Chapter 11
Probability

Abstract Probability is the branch of mathematics that deals with events whose individual outcomes are unpredictable, but whose outcomes on average are predictable. In this chapter we shall describe the rules of probability. We shall apply these rules to specific situations. As you will see, the notions and methods of calculus play an extremely important part in these applications. In particular, the logarithmic and exponential functions are ubiquitous. For these reasons, this chapter has been included in this book.

The origins of calculus lie in Newtonian mechanics, of which a brief preview was given in Sect. 10.1, where we considered the motion of a particle under the combination of a restoring force and friction. We saw that once the force acting on a particle is ascertained and the initial position and velocity of the particle specified, the whole future course of the particle is predictable. Such a predictable motion is called *deterministic*. In fact, every system of particles moving according to Newton's laws by ascertainable forces describes a predictable path. On the other hand, when the forces acting on a particle cannot be ascertained exactly, or even approximately, or when its initial position and velocity are not under our control or even our power to observe, then the path of the object is far from being predictable. Many—one is tempted to say almost all—motions observed in everyday life are of this kind. Typical examples are the wafting of smoke, drifting clouds in the sky, dice thrown, cards shuffled and dealt. Such unpredictable motion is called *nondeterministic* or *random*.

Even though the outcome of a single throw of a die is unpredictable, the average outcome in the long run is quite predictable, at least if the die is the standard kind: each number will appear in about one-sixth of a large number of throws. Similarly, if we repeatedly shuffle and deal out the top card of a deck of 52 cards, each card will appear about $1/52$ times the number of deals. With certain types of cloud formations, experience may indicate rain in three out of five cases on average.

P.D. Lax and M.S. Terrell, *Calculus With Applications*, Undergraduate Texts in Mathematics, DOI 10.1007/978-1-4614-7946-8_11,

11.1 Discrete Probability

We shall consider some simple, almost simplistic, experiments such as the tossing of a die, the shuffling of a deck of cards and dealing the top card, and the tossing of a coin. A more realistic example is the performance of a physical experiment. The two stages of an experiment are setting it up and observing its outcome. In many cases, such as meteorology, geology, oceanography, the setting up of the experiment is beyond our power; we can merely observe what has been set up by nature.

We shall deal with experiments that are *repeatable* and *nondeterministic*. Repeatable means that it can be set up repeatedly any number of times. Nondeterministic means that any single performance of the experiment may result in a variety of *outcomes*. In the simple examples mentioned at the beginning of this section, the possible outcomes are respectively a whole number between 1 and 6, any one of 52 cards, heads or tails. In this section, we shall deal with experiments that like the examples above, have *a finite number of possible outcomes*. We denote the number of possible outcomes by n, and shall number them from 1 to n.

Finally, we assume that the outcome of the experiment, unpredictable in any individual instance, is *predictable on average*. By this we mean the following: Suppose we could repeat the experiment as many times as we wished. Denote by S_j the number of instances among the first N experiments in which the jth outcome was observed to take place. Then the frequency $\dfrac{S_j}{N}$ with which the jth outcome has been observed to occur tends to a limit as N tends to infinity. We call this limit the *probability of the jth outcome* and denote it by p_j:

$$p_j = \lim_{N\to\infty} \frac{S_j}{N}. \tag{11.1}$$

These probabilities have the following properties:

(a) Each probability p_j is a real number between 0 and 1:

$$0 \le p_j \le 1.$$

(b) The sum of all probabilities equals 1:

$$p_1 + p_2 + \cdots + p_n = 1.$$

Both these properties follow from Eq. (11.1), for $\dfrac{S_j}{N}$ lies between 0 and 1, and therefore so does its limit p_j. This proves the first assertion. On the other hand, there are altogether n possible outcomes, so that each of the first N outcomes of the sequence of experiments performed falls into one of these n cases. Since S_j is the number of instances among the first N when the jth outcome was observed, it follows that

$$S_1 + S_2 + \cdots + S_n = N.$$

Dividing by N, we get

$$\frac{S_1}{N} + \frac{S_2}{N} + \cdots + \frac{S_n}{N} = 1.$$

Now let N tend to infinity. The limit of $\frac{S_1}{N}$ is p_1, that of $\frac{S_2}{N}$ is p_2, etc., so in the limit, we see that $p_1 + p_2 + \cdots + p_n = 1$, as asserted.

Sometimes, in fact very often, we are not interested in all the details of the outcome of an experiment, but merely in a particular aspect of it. For example, in drawing a card we may be interested only in the suit to which it belongs, and in throwing a die, we may be interested only in whether the outcome is even or odd. An occurrence such as drawing a spade or throwing an even number is called an *event*. In general, *we define an event E as any collection of possible outcomes*. Thus drawing a spade is the collective name for the outcomes of drawing the deuce of spades, the three of spades, etc., all the way up to drawing the ace of spades. Similarly, an even throw of a die is the collective name for throwing a two, a four or a six.

We define the probability $p(E)$ of an event E similarly to the way we defined the probability of an outcome:

$$p(E) = \lim_{N\to\infty} \frac{S(E)}{N},$$

where $S(E)$ is the number of instances among the first N performances of the experiment when the event E took place. It is easy to show that this limit exists. In fact, it is easy to give a formula for $p(E)$. For by definition, the event E takes place whenever the outcome belongs to the collection of the possible outcomes that make up the event E. Therefore, $S(E)$, the number of instances in which E has occurred, is the sum of all S_j for those j that make up E:

$$S(E) = \sum_{j \text{ in } E} S_j.$$

Divide by N:

$$\frac{S(E)}{N} = \sum_{j \text{ in } E} \frac{S_j}{N}.$$

This relation says that $\frac{S(E)}{N}$ is the sum of the frequencies $\frac{S_j}{N}$, where j is in E. We deduce that in the limit as N tends to infinity,

$$p(E) = \sum_{j \text{ in } E} p_j. \tag{11.2}$$

Two events E_1 and E_2 are called *disjoint* if both cannot take place simultaneously. That is, the set of outcomes that constitute the event E_1 and the set of outcomes that constitute the event E_2 have nothing in common. Here are some examples of disjoint events:

Example 11.1. If the experiment consists in drawing one card, let E_1 be the event of drawing a spade, and E_2 the event of drawing a heart:

$$E_1 = \{2\spadesuit, 3\spadesuit, \ldots, 10\spadesuit, J\spadesuit, Q\spadesuit, K\spadesuit, A\spadesuit\},\;\; E_2 = \{2\heartsuit, 3\heartsuit, \ldots, A\heartsuit\}.$$

Each event contains 13 outcomes, and they have no outcome in common; E_1 and E_2 are disjoint events.

Example 11.2. Suppose the experiment is to roll one die. Let E_1 be the event of throwing an even number, and E_2 the event of throwing a 3. Then E_1 consists of the outcomes 2, 4, and 6, while E_2 consists of outcome 3 only. These are disjoint.

We define the *union* of two events E_1 and E_2, denoted by $E_1 \cup E_2$, as the event of either E_1 or E_2 (or both) taking place. That is, the outcomes that constitute $E_1 \cup E_2$ are the outcomes that constitute E_1 combined with the outcomes that constitute E_2.

Example 11.3. In the card experiment, Example 11.1, $E_1 \cup E_2$ consists of half the deck: all the spades and all the hearts. In the die experiment, Example 11.2, $E_1 \cup E_2$ consists of outcomes 2, 3, 4, and 6.

The following observation is as important as it is simple: *The probability of the union of two disjoint events* is the sum of the probabilities of each event:

$$p(E_1 \cup E_2) = p(E_1) + p(E_2).$$

This is called the *addition rule for disjoint events*. This result follows from formula (11.2) for the probability of an event, for by definition of union,

$$p(E_1 \cup E_2) = \sum_{j \text{ in } E_1 \text{ or } E_2} p_j.$$

On the other hand, disjointness means that an outcome j may belong either to E_1 or to E_2 but not to both. Therefore,

$$p(E_1 \cup E_2) = \sum_{j \text{ in } E_1 \text{ or } E_2} p_j = \sum_{j \text{ in } E_1} p_j + \sum_{j \text{ in } E_2} p_j = p(E_1) + p(E_2),$$

as asserted.

Next we turn to another important idea in probability, the *independence* of two experiments. Take two experiments such as (1) throwing a die and (2) shuffling a deck and dealing the top card. Our common sense plus everything we know about the laws of nature tells us that these experiments are totally independent of each other in the sense that the outcome of one cannot possibly influence the other, nor is the outcome of both under the influence of a common cause. We state now, precisely in the language of probability theory, an important consequence of independence.

Given two experiments, we can compound them into a single *combined experiment* simply by performing them simultaneously. Let E be any event in the framework of one of the experiments, F any event in the framework of the other. The combined event of both E and F taking place will be denoted by $E \cap F$.

Example 11.4. For instance, if E is the event an even throw, and F is the event drawing a spade, then $E \cap F$ is the event of an even throw *and* drawing a spade.

We claim that *if the experiments are independent, then the probability of the combined event $E \cap F$ is the product of the separate probabilities of the events E and F*:

$$p(E \cap F) = p(E)p(F). \tag{11.3}$$

We refer to this relation as the *product rule* for independent experiments.

We now show how to deduce the product rule. Imagine the combined experiment repeated as many times as we wish. We look at the first N experiments of this sequence. Among the first N, count the number of times E has occurred, F has occurred, and $E \cap F$ has occurred. We denote these numbers by $S(E)$, $S(F)$, and $S(E \cap F)$. By definition of the probability of an event,

$$p(E) = \lim_{N\to\infty} \frac{S(E)}{N},$$

$$p(F) = \lim_{N\to\infty} \frac{S(F)}{N},$$

$$p(E \cap F) = \lim_{N\to\infty} \frac{S(E \cap F)}{N}.$$

Suppose that we single out from the sequence of combined experiments the *subsequence* of those in which E occurred. The frequency of occurrence of F in this subsequence is $\dfrac{S(E \cap F)}{S(E)}$. If the two events E and F are truly independent, the frequency with which F occurs in this subsequence should be the same as the frequency with which F occurs in the original sequence. Therefore,

$$\lim_{N\to\infty} \frac{S(E \cap F)}{S(E)} = \lim_{N\to\infty} \frac{S(F)}{N} = p(F).$$

Now we write the frequency $S(E \cap F)/N$ as the product

$$\frac{S(E \cap F)}{N} = \frac{S(E \cap F)}{S(E)} \frac{S(E)}{N}.$$

Then

$$\lim_{N\to\infty} \frac{S(E \cap F)}{N} = \lim_{N\to\infty} \frac{S(E \cap F)}{S(E)} \cdot \lim_{N\to\infty} \frac{S(E)}{N}.$$

Therefore, $p(E \cap F) = p(E)p(F)$.

Suppose that one experiment has m outcomes numbered $1, 2, \ldots, j, \ldots, m$, and the other has n outcomes numbered $1, 2, \ldots, k, \ldots, n$. Denote their respective probabilities by $p_1, \ldots, p_m$ and $q_1, \ldots, q_n$. The combined experiment then has mn possible outcomes, namely all pairs of outcomes (j, k). If the experiments are independent, then the product rule tells us that the outcome (j,k) of the combined experiment has probability

$$p_j q_k.$$

This formula plays a very important role in probability theory. We now give an illustration of its use.

Suppose that both of the two experiments we have been discussing are the tossing of a die. Then the combined experiment is the tossing of a pair of dice. Each experiment has six possible outcomes, with probability $\frac{1}{6}$. There are 36 outcomes for the combined experiment, which we can list from $(1,1)$ to $(6,6)$. According to the product rule for independent events, $p(E \cap F) = p(E)p(F)$, so each combined outcome has probability $\frac{1}{36}$. We now ask the following question: What is the probability of the event of tossing a 7? There are six ways of tossing a 7:

$$(1,6),\ (2,5),\ (3,4),\ (4,3),\ (5,2),\ (6,1).$$

The probability of tossing a 7 is the sum of the probabilities of these six outcomes that constitute the event. That sum is

$$\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}+\frac{1}{36}=\frac{1}{6}.$$

Similarly, we can calculate the probability of tossing any number between 2 and 12. We ask you in Problem 11.6 to go through the calculations of determining the probabilities that the numbers $2,3,\ldots,12$ will be thrown. The results are in Table 11.1.

Table 11.1 Probabilities for the sum of two independent dice

Throw	2	3	4	5	6	7	8	9	10	11	12
Probability	$\frac{1}{36}$	$\frac{1}{18}$	$\frac{1}{12}$	$\frac{1}{9}$	$\frac{5}{36}$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{9}$	$\frac{1}{12}$	$\frac{1}{18}$	$\frac{1}{36}$

Numerical Outcome. We now turn to another important concept of probability, the *numerical outcome* of an experiment. In physical experiments designed to measure the value of a single physical quantity, the numerical outcome is simply the measured value of the quantity in question. For the simple example of throwing a pair of dice, the numerical outcome might be the sum of the face values of each die. For the experiment of dealing a bridge hand, the numerical outcome might be the point count of the bridge hand. In general, the *numerical outcome* of an experiment *means the assignment of a real number x_j to each of the possible outcomes*, $j = 1,2,\ldots,\ n$.

Note that different outcomes may be assigned the same number, as in the case of the dice; the numerical outcome 7 is assigned to the six different outcomes $(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$.

Expectation. We show now that in a random experiment with n possible outcomes of probability p_j and numerical outcome x_j $(j = 1,2\ldots n)$, the average numerical outcome, called the *mean* of x, or *expectation* of x, denoted by $\overline{x}$ or $E(x)$, is given by the formula

$$\overline{x} = E(x) = p_1x_1 + \cdots + p_nx_n. \tag{11.4}$$

To prove this, denote as before by S_j the number of instances among the first N in which the jth outcome was observed. The average numerical outcome among the first N is therefore

$$\frac{S_1x_1 + S_2x_2 + \cdots + S_nx_n}{N}.$$

We rewrite this as

$$\frac{S_1}{N}x_1 + \frac{S_2}{N} + \cdots + \frac{S_n}{N}x_n.$$

By hypothesis, each of the ratios $\frac{S_j}{N}$ tends to the limit p_j. It follows that the average numerical outcome tends to $\overline{x}$, as asserted.

We now give an example of formula (11.4) for the average numerical outcome. Take the experiment of throwing a pair of dice. We classify the outcomes as throwing a 2, 3, ..., up to 12. We take these numbers to be the numerical outcomes of the experiment. The probability of each outcome is given in Table 11.1. We get the following value for the average numerical outcome of a throw of a pair of dice:

$$\overline{x} = \frac{1}{36}2 + \frac{1}{18}3 + \frac{1}{12}4 + \frac{1}{9}5 + \frac{5}{36}6 + \frac{1}{6}7 + \frac{5}{36}8 + \frac{1}{9}9 + \frac{1}{12}10 + \frac{1}{18}11 + \frac{1}{36}12 = 7.$$

Variance. We have shown that if we perform a random experiment with numerical outcomes many times, the average of the numerical outcomes will be very close to the mean, given by Eq. (11.4). A natural question is this: by how much do the numerical outcomes differ on average from the mean? The average difference is

$$\sum_{i=1}^{n}(x_i - \overline{x})p_i = \sum_{i=1}^{n} p_ix_i - \Big(\sum_{i=1}^{n} p_i\Big)\overline{x} = \overline{x} - \overline{x} = 0,$$

not very informative. It turns out that a related concept, the *variance*, has much better mathematical properties.

Definition 11.1. The *variance*, denoted by V, is the expected value of the square of the difference of the outcome and its expected value:

$$V = \overline{(x-\overline{x})^2} = E\big((x - E(x))^2\big).$$

We show how to express the variance in terms of the numerical outcomes and their probabilities. The numerical outcome x_j differs from the mean $\overline{x}$ by $x_j - \overline{x}$. Its square is $(x_j - \overline{x})^2$, which is equal to

$$x_j^2 - 2x_j\overline{x} + (\overline{x})^2. \tag{11.5}$$

Denote as before by S_j the number of times the jth outcome occurred among the first N events. The expected value of the quantity in Eq. (11.5) is

$$\frac{S_1x_1^2+\cdots+S_nx_n^2}{N}-2\frac{S_1x_1+\cdots+S_nx_n}{N}\overline{x}+(\overline{x})^2.$$

As N tends to infinity, $\frac{S_j}{N}$ tends to p_j. Therefore, the expected value above tends to

$$V=p_1x_1^2+\cdots+p_nx_n^2-2(p_1x_1+\cdots+p_nx_n)\overline{x}+(\overline{x})^2=\overline{x^2}-(\overline{x})^2.$$

We denote the expected value of the square of the outcome by $\overline{x^2}=E(x^2)$. This leads to an alternative way to calculate the variance:

$$V=E\big((x-E(x))^2\big)=E(x^2)-\big(E(x)\big)^2. \tag{11.6}$$

Definition 11.2. The square root of the variance is called the *standard deviation*.

The Binomial Distribution. Suppose a random experiment has two possible outcomes A and B, with probabilities p and q respectively, where $p+q=1$. For example, think of a coin toss, or an experiment with two outcomes, A success and B failure. Choose any positive integer N and repeat the experiment N times. Assume that the repeated experiments are independent of each other. This new combined experiment has outcomes that are a string of successes and failures. For example, if $N=5$, then *ABAAB* is one possible outcome, and another is *BABBB*. If we let x be the numerical outcome that this repeated experiment results in A occurring exactly x times, we see that x has possible values

$$x=0,\ 1,\ ,\ldots,N.$$

The probability that the outcome A occurs exactly k times and the outcome B exactly $N-k$ times is given by the expression

$$b_k(N)=\binom{N}{k}p^kq^{N-k}.$$

For since the outcomes of the experiments are independent of each other, the probability of a particular sequence of (k) A's and $(N-k)$ B's is p^kq^{N-k}. Since there are exactly $\binom{N}{k}$ arrangements of (k) A's and $(N-k)$ B's, this proves the result. We make the following definition.

Definition 11.3. The probabilities

$$b_k(N) = \binom{N}{k} p^k q^{N-k} \tag{11.7}$$

are called the *binomial* distribution. We call $b_k(N)$ the probability of k successes in N independent trials, where p is the probability of success in one trial, and $q = 1 - p$ it the probability of failure.

The sum of the probabilities of all possible outcomes is

$$\sum_{k=0}^{N} \binom{N}{k} p^k q^{N-k} = \sum_{k=0}^{N} \binom{N}{k} p^k (1-p)^{N-k}.$$

According to the binomial theorem, this sum is equal to $(p+1-p)^N = 1^N = 1$.

Example 11.5. Suppose a fair coin is tossed 10 times, and x is the number of heads. What is the probability of getting 7 heads and 3 tails? "Fair" means that the probabilities of heads or tails on one toss are each $\frac{1}{2}$. We take $N = 10$. There are $\binom{10}{7} = 120$ ways for 7 heads to turn up out of 10 tosses. The probability of each is $\left(\frac{1}{2}\right)^{10}$. Therefore,

$$\begin{aligned} p(x=7) &= \binom{10}{7} \left(\frac{1}{2}\right)^7 \left(\frac{1}{2}\right)^3 \\ &= \frac{(10)(9)(8)}{3!} \left(\frac{1}{2}\right)^{10} = (120)\frac{1}{1024} = 0.1171875 \end{aligned}$$

We calculate now the expected value of the number occurrences of outcome A when there are N independent trials. Let $(x = k)$ be the numerical outcome that there are exactly k occurrences of A, and let $p(x = k)$ be its probability. We saw above that $p(x=k) = \binom{N}{k} p^k (1-p)^{N-k}$ for each possible value of $x = 0, 1, 2, \ldots, k, \ldots, N$. By definition of expectation,

$$E(x) = \sum_{k=0}^{N} kp(x=k) = \sum_{k=0}^{N} k \binom{N}{k} p^k q^{N-k} = \sum_{k=1}^{N} k \binom{N}{k} p^k q^{N-k}.$$

Using the formula for the binomial coefficients, we can write the formula for the expected value as

$$E(x) = \sum_{k=1}^{N} \frac{kN!}{k!(N-k)!} p^k q^{N-k} = Np \sum_{k=1}^{N} \frac{(N-1)!}{(k-1)!(N-k)!} p^{k-1} q^{N-k}.$$

Using the binomial theorem, we can rewrite the last sum as

$$Np(p+q)^{N-1} = Np.$$

Thus we have proved that for the binomial distribution, the expected number of successes, $E(x)$, is Np.

Note that since the probability of the outcome A in a single trial is p, it is reasonable to expect the outcome A to occur Np times after the experiment has been performed N times.

The Poisson Distribution. Suppose you know that each week, a large number of vehicles pass through a busy intersection and there are on average u accidents. Let us assume that the probability of a vehicle having an accident is independent of the occurrence of previous accidents. We use a binomial distribution to determine the probability of k accidents in a week:

$$b_k(N) = \binom{N}{k} p^k (1-p)^{N-k} = \frac{N(N-1)\dots(N-k+1)}{k!} p^k (1-p)^{N-k}$$
$$= \left(1-\frac{1}{N}\right)\dots\left(1-\frac{k-1}{N}\right)\frac{N^k p^k (1-p)^{N-k}}{k!}.$$

Setting $p = \dfrac{u}{N}$, we can rewrite this as

$$= \left[\frac{\left(1-\frac{1}{N}\right)\cdots\left(1-\frac{k-1}{N}\right)}{(1-p)^k}\right]\frac{u^k}{k!}\left(1-\frac{u}{N}\right)^N.$$

As N tends to infinity, p tends to 0, and the factor in brackets tends to 1, because both the numerator and the denominator tend to 1. As shown in Sect. 1.4, the third factor tends to e^{-u}. Therefore,

$$\lim_{N\to\infty,\ u=Np} b_k(N) = \frac{u^k}{k!}\mathrm{e}^{-u}.$$

This gives us an estimate for $b_k(N)$ when N is large, p is small, and $Np = u$.

Definition 11.4. The *Poisson* distribution is the set of probabilities

$$p_k(u) = \frac{u^k}{k!}\mathrm{e}^{-u}, \tag{11.8}$$

where u is a parameter. The number p_k is the probability of k favorable outcomes, $k = 0, 1, 2, \dots$.

Note that the sum of the $p_k(u)$ equals 1:

$$\sum_{k=0}^{\infty} p_k(u) = \mathrm{e}^{-u} \sum_{k=0}^{\infty} \frac{u^k}{k!} = \mathrm{e}^{-u}\mathrm{e}^{u} = 1.$$

Here we have used the expression of the exponential function given by its Taylor series. The Poisson process is an example of discrete probability with infinitely many possible outcomes.

Next, we show that the combination of two Poisson processes is again a Poisson process. Denote by $p_k(u)$ and $p_k(v)$ the probability of k favorable outcomes in these processes, where p_k is given by formula (11.8). We claim that the probability of k favorable outcomes when both experiments are performed is $p_k(u+v)$, assuming that the experiments are independent.

Proof. There will be k favorable outcomes for the combined experiment if the first experiment has j favorable outcomes and the second experiment has $(k-j)$. If the experiments are independent, the probability of such a combined outcome is the product of the probabilities,

$$p_j(u)p_{k-j}(v),$$

so the probability of the combined experiment to have k favorable outcomes is the sum

$$\sum_j p_j(u)p_{k-j}(v) = \sum_j \frac{u^j}{j!}\mathrm{e}^{-u}\frac{v^{k-j}}{(k-j)!}\mathrm{e}^{-v}.$$

We rewrite this sum as

$$\frac{1}{k!}\mathrm{e}^{-(u+v)}\sum_j \frac{k!}{j!(k-j)!}u^j v^{k-j}.$$

The sum in this formula is the binomial expression for $(u+v)^k$. Therefore, the probability of k favorable outcomes for the combined experiment is

$$\frac{1}{k!}(u+v)^k e^{-(u+v)},$$

which is the Poisson distribution $p_k(u+v)$. □

Problems

11.1. Calculate the variance of the outcome when one die is rolled.

11.2. Find the probability of getting exactly three heads in six independent tosses of a fair coin.

11.3. Let E be an event consisting of a certain collection of outcomes of an experiment. We may call these outcomes *favorable* from the point of view of the event that interests us. The collection of all unfavorable outcomes, i.e., those that do not belong to E, is called the event *complementary* to E. Denote by E' the complementary event. Prove that

$$p(E) + p(E') = 1.$$

11.4. We have said that the probability of the outcome (j,k) of the combination of two independent experiments is $p_j q_k$ when the outcomes of one experiment have probabilities p_j and the other q_k, where $j = 1, 2, \ldots, n$ and $k = 1, 2, \ldots, m$. Show that the sum of all these probabilities is 1.

11.5. An event E is *included* in the event F if whenever E takes place, F also takes place. Another way of expressing this relationship is to say that the outcomes that constitute E form a *subset* of the outcomes that make up F. The assertion "the event E is included in the event F" is expressed in symbols by $E \subset F$. For example, the event E of drawing a spade is included in the event F of drawing a black card.

Show that if $E \subset F$, then

$$p(E) \leq p(F).$$

11.6. Verify the probabilities shown for two independent dice in Table 11.1.

11.7. Let $E_1, E_2, \ldots, E_m$ be a collection of m events that are *disjoint* in the sense that no outcome can belong to more than one event. Denote the union of the events E_j by

$$E = E_1 \cup E_2 \cup \cdots \cup E_m.$$

Show that the *additive* rule holds:

$$p(E) = p(E_1) + \cdots + p(E_m).$$

11.8. Show that the variance of the number of successes in the binomial distribution is $Np(1-p)$.

11.9. Let x represent the number of successes in a Poisson distribution (11.8). Show that the expected value

$$E(x) = \sum_{k=0}^{\infty} k p_k(u)$$

is equal to u.

11.2 Information Theory: How Interesting Is Interesting?

It is a universal human experience that some information is dull, some interesting. Man bites dog is news, dog bites man is not. In this section, we describe a way of assigning a quantitative measure to the value of a piece of information.

"Interesting," in this discussion, shall mean the degree of surprise at being informed that a certain event E, whose occurrence is subject to chance, has occurred.

An event is a collection of possible outcomes of an experiment. The frequency with which the event E occurs in a large number of performances of the experiment is its probability $p(E)$. We assume in this theory that the information gained on learning that an event has occurred depends only on the probability p of the event. We denote by $f(p)$ the information thus gained. In other words, we could think of $f(p)$ as a measure of the element of surprise generated by the event that has occurred.

What properties does this function f have? We claim that the following four are mandatory:

(a) $f(p)$ increases as p decreases.
(b) $f(1)=0$.
(c) $f(p)$ tends to infinity as p tends to 0.
(d) $f(pq)=f(p)+f(q)$.

Property (a) expresses the fact that the occurrence of a less probable event is more surprising than the occurrence of a more probable one and therefore carries more information. Property (b) says that the occurrence of an event that is a near certainty imparts almost no new information, while property (c) says that the occurrence of a rare event is of great interest and furnishes a great deal of new information.

Property (d) expresses a property of independent events. Suppose that two events E and F are *independent*. Since such events are totally unrelated, being informed that both of them have occurred conveys no more information than learning that each has occurred separately, i.e., the information gained on learning that both have occurred is the *sum* of the information gained by learning of each occurrence separately. Denote by p and q the probabilities of events E and F, respectively. According to the product rule (11.3), the probability of the combined event $E \cap F$ is the product pq. It is not hard to show that the only continuous function that satisfies property (d) is a constant multiple of $\log p$. So we conclude that

$$f(p)=k\log p. \tag{11.9}$$

What is the value of this constant? According to property (a), $f(p)$ increases with decreasing p. Since $\log p$ increases with increasing p, we conclude that the constant must be *negative*. What about its magnitude? There is no way of deciding that without first adopting an arbitrary unit of information. For convenience, we choose the constant to be -1, and so define

$$f(p)=-\log p.$$

We ask you to verify properties (b) and (c) in Problem 11.10.

Now consider an experiment with n possible outcomes having probabilities $p_1, p_2, \ldots, p_n$. If in a single performance of the experiment, the jth outcome occurs, we have gained information in the amount $-\log p_j$. We now ask the following question: If we perform the experiment repeatedly many times, what is the *average information gain*? The answer to this question is contained in formula (11.4) concerning the average numerical outcome of a series of experiments. According to that formula, if the jth numerical outcome is x_j, and the average numerical outcome is

$p_1x_1 + \cdots + p_nx_n$. In our case, the numerical outcome, the information gained in the jth outcome, is

$$x_j = -\log p_j.$$

So the average information gain I is $I = -(p_1 \log p_1 + p_2 \log p_2 + \cdots + p_n \log p_n)$. To indicate the dependence of I on the probabilities, we write

$$I = I(p_1, \ldots, p_n) = -p_1 \log p_1 - p_2 \log p_2 - \cdots - p_n \log p_n. \tag{11.10}$$

This definition of information is due to the physicist Léo Szilárd. It was introduced in the mathematical literature by Claude Shannon.

Let us look at the simplest case that there are only two possible outcomes, with probabilities p and $1-p$. We can write the formula for information gain as follows:

$$I = -p \log p + (p-1)\log(1-p).$$

How does I depend on p? To study how I changes with p, we use the methods of calculus: we differentiate I with respect to p and get

$$\frac{\mathrm{d}I}{\mathrm{d}p} = -\log p - 1 + \log(1-p) + 1 = -\log p + \log(1-p).$$

Using the functional equation of the logarithm function, we can rewrite this as

$$\frac{\mathrm{d}I}{\mathrm{d}p} = \log\left(\frac{1-p}{p}\right).$$

We know that $\log x$ is positive for $x > 1$ and negative for $x < 1$. Also,

$$\frac{(1-p)}{p} \begin{cases} > 1 \text{ for } 0 < p < \frac{1}{2}, \\ < 1 \text{ for } \frac{1}{2} < p < 1. \end{cases}$$

Therefore,

$$\frac{\mathrm{d}I}{\mathrm{d}p} \begin{cases} > 0 \text{ for } 0 < p < \frac{1}{2}, \\ < 0 \text{ for } \frac{1}{2} < p < 1. \end{cases}$$

It follows that $I(p)$ is an increasing function of p from 0 to $\frac{1}{2}$, and a decreasing function as p goes from $\frac{1}{2}$ to 1. Therefore, *the largest value of I occurs when $p = \frac{1}{2}$*. In words: the most information that can be gained on average from an experiment with two possible outcomes occurs when the probabilities of the two outcomes are equal.

We now extend this result to experiments with n possible outcomes.

Theorem 11.1. *The function*

$$I(p_1,\dots,p_n) = -p_1\log p_1 - p_2\log p_2 - \cdots - p_n\log p_n,$$

defined for positive numbers with $p_1+p_2+\cdots+p_n = 1$*, is largest when*

$$p_1 = p_2 = \cdots = p_n = \frac{1}{n}.$$

Proof. We have to show that

$$I(p_1,\dots,p_n) < I\Big(\frac{1}{n},\dots,\frac{1}{n}\Big)$$

unless all the p_j are equal to $\frac{1}{n}$. In order to apply the methods of calculus to proving this inequality, we consider the following functions $r_j(s)$:

$$r_j(s) = sp_j + (1-s)\frac{1}{n}, \qquad\qquad j = 1,\dots,n.$$

These functions are designed so that at $s = 0$, the value of each r_j is $\frac{1}{n}$, and at $s = 1$, the value of r_j is p_j:

$$r_j(0) = \frac{1}{n}, \qquad r_j(1) = p_j, \qquad j = 1,\dots,n.$$

So, if we define the function $J(s) = I(r_1(s),\dots,r_n(s))$, then

$$J(0) = I\Big(\frac{1}{n},\dots,\frac{1}{n}\Big), \qquad J(1) = I(p_1,\dots,p_n).$$

Therefore, the inequality to be proved can be expressed simply as $J(1) < J(0)$. We shall prove this by showing that $J(s)$ is a decreasing function of s. We use the monotonicity criterion to demonstrate the decreasing character of $J(s)$, by verifying that its derivative is negative. To calculate the derivative of $J(s)$, we need to know the derivative of each r_j with respect to s. This is easily calculated:

$$\frac{\mathrm{d}r_j(s)}{\mathrm{d}s} = p_j - \frac{1}{n}. \tag{11.11}$$

Note that the derivative of each r_j is constant, since each r_j is a linear function of s. Using the definition of I, we have

$$J(s) = -r_1\log r_1 - \cdots - r_n\log r_n.$$

We calculate the derivative of J using the chain rule and Eq. (11.11):

$$\begin{aligned}\frac{\mathrm{d}J}{\mathrm{d}s} &= -(1+\log r_1)\frac{\mathrm{d}r_1}{\mathrm{d}s} - \cdots - (1+\log r_n)\frac{\mathrm{d}r_n}{\mathrm{d}s} \\ &= -(1+\log r_1)\left(p_1 - \frac{1}{n}\right) - \cdots - (1+\log r_n)\left(p_n - \frac{1}{n}\right). \end{aligned} \tag{11.12}$$

Since $r_j(0) = \dfrac{1}{n}$, we get with $s = 0$ that

$$\frac{\mathrm{d}J}{\mathrm{d}s}(0) = -\left(1+\log\left(\frac{1}{n}\right)\right)\left(p_1 - \frac{1}{n}\right) - \cdots - \left(1+\log\left(\frac{1}{n}\right)\right)\left(p_n - \frac{1}{n}\right).$$

Since the sum of the p_j is 1, this gives

$$\frac{\mathrm{d}J}{\mathrm{d}s}(0) = -\left(1+\log\left(\frac{1}{n}\right)\right)\left(1 - n\frac{1}{n}\right) = 0.$$

Switching to the J' notation, we have $J'(0) = 0$. We claim that for all positive values of s,

$$J'(s) < 0. \tag{11.13}$$

If we can show this, our proof that J is a decreasing function is complete. To verify Eq. (11.13), we shall show that $J'(s)$ itself is a decreasing function of s. Since $J'(0) = 0$, then J' will be negative for all positive s.

To show that J' is decreasing, we apply the monotonicity criterion once more, this time to J', and show that J'' is negative. We compute J'' by differentiating Eq. (11.12) and using Eq. (11.11):

$$\begin{aligned} J'' &= -\frac{1}{r_1}r_1'\left(p_1 - \frac{1}{n}\right) - \cdots - \frac{1}{r_n}r_n'\left(p_n - \frac{1}{n}\right) \\ &= -\frac{1}{r_1}\left(p_1 - \frac{1}{n}\right)^2 - \cdots - \frac{1}{r_n}\left(p_n - \frac{1}{n}\right)^2. \end{aligned}$$

Each term in this sum is negative or zero. Since not all p_j are equal to $\dfrac{1}{n}$, at least some terms are negative. This proves $J'' < 0$, and completes the proof of our theorem. □

Problems

11.10. Verify that the function $f(p) = -k\log p$ has properties (b) and (c) that we listed at the outset of Sect. 11.2.

11.11. Suppose that an experiment has three possible outcomes, with probabilities p, q, and r, with

$$p+q+r=1.$$

Suppose that we simplify the description of our experiment by lumping the last two cases together, i.e., we look on the experiment as having two possible outcomes, one with probability p, the other with probability $1-p$. The average information gain in looking at the full description of the experiment is

$$-p\log p - q\log q - r\log r.$$

In looking at the simplified description, the average information gain is

$$-p\log p - (1-p)\log(1-p).$$

Prove that the average information gain from the full experiment is greater than that obtained from its simplified description. The result is to be expected: if we lump data together, we lose information.

11.12. Let $p_1, \ldots, p_n$ be the probabilities of the n possible outcomes of an experiment, and $q_1, \ldots, q_m$ the probabilities of the outcomes of another experiment. Suppose that the experiments are *independent*, i.e., if we combine the two experiments, the probability of the first experiment having the jth outcome and the second experiment having the kth outcome is the product

$$r_{jk} = p_j q_k.$$

Show that in this case, the average information gain from the combined experiment is the *sum* of the average information gains in the performance of each experiment separately:

$$I(r_{11},\ldots,r_{mn}) = I(p_1,\ldots,p_n) + I(q_1,\ldots,q_m).$$

11.13. Suppose an experiment can have n possible outcomes, the jth having probability p_j, $j=1,\ldots,n$. The information gained from this experiment is on average

$$-p_1\log p_1 - \cdots - p_n\log p_n.$$

Suppose we simplify the description of the experiment by lumping the last $n-1$ outcomes together as failures of the first case. The average information gain from this description is

$$-p_1\log p_1 - (1-p_1)\log(1-p_1).$$

Prove that we gain on average more information from the full description than from the simplified description.

11.3 Continuous Probability

The probability theory developed in Sect. 11.1 deals with experiments that have finitely many possible numerical outcomes. This is a good model for experiments such as tossing a coin (labeling the numerical outcome 0 or 1) or throwing a die, but it is artificial for experiments such as making a physical measurement with an apparatus subject to random disturbances that can be reduced but not totally eliminated. *Every real number is a possible numerical outcome* of such an experiment. This section is devoted to developing a probability theory for such situations. The experiments we study are, just like the previous ones, repeatable and nondeterministic but predictable on average.

By "predictable on average" we mean this: Repeat the experiment as many times as we wish and denote by $S(x)$ the number of instances among the first N performances for which the numerical outcome was less than x. Then the frequency $\frac{S(x)}{N}$ with which this event occurs tends to a limit as N tends to infinity. This limit is the *probability that the outcome is less than x*, and is denoted by $P(x)$:

$$P(x) = \lim_{N\to\infty} \frac{S(x)}{N}.$$

The probability $P(x)$ has the following properties:

(i) Each probability lies between 0 and 1:

$$0 \le P(x) \le 1.$$

(ii) $P(x)$ is a nondecreasing function of x.

Properties (i) and (ii) are consequences of the definition, for the number $S(x)$ lies between 0 and N, so that the ratio $\frac{S(x)}{N}$ lies between 0 and 1; but then so does the limit $P(x)$. Secondly, $S(x)$ is a nondecreasing function of x, so that the ratio $\frac{S(x)}{N}$ is a nondecreasing function of x; then so is the limit $P(x)$. We shall assume two further properties of $P(x)$:

(iii) $P(x)$ tends to 0 as x tends to minus infinity.
(iv) $P(x)$ tends to 1 as x tends to infinity.

Property (iii) says that the probability of a very large negative outcome is very small. Property (iv) implies that very large positive outcomes are very improbable, as we ask you to explain in Problem 11.14. As in Sect. 11.1, we shall be interested in collections of outcomes, which we call *events*.

Example 11.6. Examples of events are:

(a) The outcome is less than x.

(b) The outcome lies in the interval I.
(c) The outcome lies in a given collection of intervals.

The probability of an event E, which we denote by $P(E)$, is defined as in Sect. 11.1, as the limit of the frequencies:

$$\lim_{N\to\infty} \frac{S(E)}{N} = P(E),$$

where $S(E)$ the number of times the event E took place among the first N of an infinite sequence of performances of an experiment. The argument presented in Sect. 11.1 can be used in the present context to show the additive rules for disjoint events: Suppose E and F are two events that have probabilities $P(E)$ and $P(F)$, and suppose that they are *disjoint* in the sense that one event precludes the other. That is, no outcome can belong to both E and F. In this case, the union $E \cup F$ of the events, consisting of all outcomes either in E or in F, also has a probability that is the sum of the probabilities of E and F:

$$P(E \cup F) = P(E) + P(F).$$

We apply this to the events

$$E: \text{ the outcome } x < a$$

and

$$F: \text{ the outcome } a \le x < b.$$

The union of these two is

$$E \cup F: \text{ the outcome } x < b.$$

Then

$$P(E) = P(a), \quad P(E \cup F) = P(b).$$

We conclude that

$$P(F) = P(b) - P(a)$$

is the probability of an outcome less than b but greater than or equal to a.

We now make the following assumption:

(v) $P(x)$ is a continuously differentiable function.

This assumption holds in many important cases and allows us to use the methods of calculus. We denote the derivative of P by p:

$$\frac{\mathrm{d}P(x)}{\mathrm{d}x} = p(x)$$

The function $p(x)$ is called the *probability density*. According to the mean value theorem, for every a and b, there is a number c lying between a and b such that

$$P(b) - P(a) = p(c)(b-a). \tag{11.14}$$

According to the fundamental theorem of calculus,

$$P(b) - P(a) = \int_a^b p(x)\,\mathrm{d}x. \tag{11.15}$$

Since by assumption (iii), $P(a)$ tends to 0 as a tends to minus infinity, we conclude that

$$P(b) = \int_{-\infty}^{b} p(x)\,\mathrm{d}x.$$

Since by assumption (iv), $P(b)$ tends to 1 as b tends to infinity, we conclude that

$$1 = \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x.$$

This is the continuous analogue of the basic fact that $p_1 + p_2 + \cdots + p_n = 1$ in discrete probability. According to property (ii), $P(x)$ is a nondecreasing function of x. Since the derivative of a nondecreasing function is nowhere negative, we conclude that $p(x)$ is nonnegative for all x:

$$0 \le p(x).$$

We now define the *expectation*, or mean, $\overline{x}$ of an experiment analogously to the discrete case. Imagine the experiment performed as many times as we wish, and denote the sequence of outcomes by

$$a_1, a_2, \ldots, a_N, \ldots.$$

Theorem 11.2. *If an experiment is predictable on average, and if the outcomes are restricted to lie in a finite interval, then*

$$\overline{x} = \lim_{N\to\infty} \frac{a_1 + \cdots + a_N}{N}$$

exists and is equal to

$$\overline{x} = \int_{-\infty}^{\infty} x p(x)\,\mathrm{d}x. \tag{11.16}$$

The assumption that the outcomes lie in a finite interval is a realistic one if one thinks of the experiment as a measurement. After all, every measuring apparatus has a finite range. However, there are probability densities of great theoretical interest, such as the ones we shall discuss in Sect. 11.4, that are positive for all real x. Theorem 11.2 remains true for these experiments, too, under the additional assumption that the improper integral defining $\overline{x}$ exists.

Proof. Divide the interval I in which all outcomes lie into n subintervals $I_1,\dots,I_n$. Denote the endpoints by

$$e_0 < e_1 < \cdots < e_n.$$

The probability P_j of an outcome lying in interval I_j is the difference of the values of P at the endpoints of I_j. According to formula (11.14), this difference is equal to

$$P_j = P(e_j) - P(e_{j-1}) = p(x_j)(e_j - e_{j-1}), \tag{11.17}$$

where x_j is a point in I_j guaranteed by the mean value theorem, and $(e_j - e_{j-1})$ denotes the length of I_j. We now simplify the original experiment by recording merely the intervals I_j in which the outcome falls, and calling the numerical outcome in this case x_j, the point in I_j that appears in formula (11.17). The actual outcome of the full experiment and the numerical outcome of the simplified experiment always belong to the same subinterval of the subdivision we have taken. Therefore, *these two outcomes differ by at most* w, *the length of largest of the subintervals* I_j.

Now consider the sequence of outcomes $a_1,\ a_2,\dots$ of the original experiment. Denote the corresponding outcomes of the simplified experiment by $b_1,\ b_2,\dots$. The simplified experiment has a finite number of outcomes. For such discrete experiments, we have shown in Sect. 11.1 that the average of the numerical outcomes tends to a limit, called the expectation. We denote it by $\overline{x}_n$:

$$\overline{x}_n = \lim_{N\to\infty} \frac{b_1 + \cdots + b_N}{N}, \tag{11.18}$$

where n is the number of subintervals of I. The expectation $\overline{x}_n$ of the simplified experiment can be calculated by formula (11.4):

$$\overline{x}_n = P_1 x_1 + \cdots + P_n x_n. \tag{11.19}$$

By Eq. (11.17), this is

$$= p(x_1)x_1(e_1 - e_0) + \cdots + p(x_n)x_n(e_n - e_{n-1}).$$

We recognize this as an approximating sum for the integral of $xp(x)$ over I. If the subdivision is fine enough, *the approximating sum* $\overline{x}_n$ *differs very little from the value of the integral*

$$\int_{e_0}^{e_n} xp(x)\,\mathrm{d}x. \tag{11.20}$$

We recall that the outcomes of the simplified experiment and the full experiment differ by less than w, the length of the largest subinterval I_j. Therefore, the expectation of the simplified experiment tends to the expectation of the full experiment as the lengths of the subintervals tend to zero. This proves that the expectation of the full experiment is given by the integral (11.20). Since $p(x)$ is zero outside the interval I, the integrals (11.20) and (11.16) are equal. This concludes the proof of Theorem 11.2. □

We now give some examples of expectation.

Example 11.7. Let A be a positive number, and define $p(x)$ by

$$p(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1/A & \text{for } 0 \le x < A, \\ 0 & \text{for } A \le x. \end{cases}$$

This is intended to mean that the numerical outcome x is equally likely to occur anywhere in $[0,A]$. This choice of p satisfies $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = \int_0^A \frac{\mathrm{d}x}{A} = 1$. We now compute the expected value

$$\overline{x} = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x = \int_0^A \frac{x}{A}\,\mathrm{d}x = \left[\frac{x^2}{2A}\right]_0^A = \frac{A}{2}.$$

Example 11.8. Let A be a positive number and set

$$p(x) = \begin{cases} 0 & \text{for } x < 0, \\ A\mathrm{e}^{-Ax} & \text{for } 0 \le x. \end{cases}$$

Let us check that p satisfies $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$. Using the fundamental theorem of calculus, we have

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = \int_0^{\infty} A\mathrm{e}^{-Ax}\,\mathrm{d}x = -\mathrm{e}^{-Ax}\big|_0^{\infty} = 1.$$

We now compute $\overline{x}$. Using integration by parts and then the fundamental theorem, we have

$$\overline{x} = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x = \int_0^{\infty} xA\mathrm{e}^{-Ax}\,\mathrm{d}x = \int_0^{\infty} \mathrm{e}^{-Ax}\,\mathrm{d}x = \left[\frac{-\mathrm{e}^{-Ax}}{A}\right]_0^{\infty} = \frac{1}{A}.$$

Example 11.9. Assume that $p(x)$ is an even function. Then $xp(x)$ is an odd function, and so

$$\overline{x} = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x = 0.$$

Let $f(x)$ be any function of x. We define the *expected value of* f with respect to the probability density $p(x)$ as

$$\overline{f} = \int_{-\infty}^{\infty} f(x)p(x)\,\mathrm{d}x.$$

One can show, analogously to the foregoing discussion, that if $a_1, \dots, a_N, \dots$ is a sequence of outcomes, then

$$\lim_{N\to\infty} \frac{f(a_1) + \cdots + f(a_N)}{N} = \overline{f}.$$

Independence. We now turn to the important concept of *independence*. The intuitive notion is the same as in the discrete models discussed in Sect. 11.1: two experiments are independent if the outcome of either has no influence on the other, nor are they both influenced by a common cause. We analyze the consequences of independence the same way we did previously, by constructing a *combined experiment* consisting of performing both experiments.

We first analyze the case that the outcome of the first experiment may be any real number, but the second experiment can have only a finite number of outcomes. As before, we denote by $P(a)$ the probability that the numerical outcome of the first experiment is less than a. The second experiment has n possible numerical outcomes $a_1, \dots, a_n$, which occur with probabilities $Q_1,\ Q_2,\ \dots,\ Q_n$. We define the *numerical outcome* of the combined experiment to be the *sum* of the separate numerical outcomes of the two experiments that constitute it.

We now derive a useful and important formula for the probability that the numerical outcome of the combined experiment is less than x. We denote this event by $E(x)$, and denote its probability by $U(x)$. We shall show that

$$U(x) = Q_1 P(x - a_1) + \cdots + Q_n P(x - a_n). \tag{11.21}$$

Proof. The numerical outcome of the second experiment is one of the n numbers a_j. The numerical outcome of the combined experiment is then less than x if and only if the outcome of the first experiment is less than $x - a_j$. We denote this event by $E_j(x)$. Thus the event $E(x)$ is the union

$$E(x) = E_1(x) \cup \cdots \cup E_n(x).$$

The events $E_j(x)$ are disjoint, that is, an outcome cannot belong to two distinct events $E_j(x)$ and $E_k(x)$. It follows then from the addition rule for disjoint events that the probability of their union $E(x)$ is the sum of the probabilities of the events $E_j(x)$.

Since the two experiments are independent, the probability of $E_j(x)$ is given by the product of the probabilities of the two experiments,

$$Q_j P(x - a_j).$$

The sum of the probabilities of the $E_j(x)$ is $U(x)$, the probability of $E(x)$. This completes the proof of Eq. (11.21). □

We now turn to the situation in which both experiments can have any real number as outcome. We denote by $P(a)$ and $Q(a)$ the probabilities that the outcome is less than a in each of the two experiments, respectively.

We shall prove the following analogue of formula (11.21): Suppose that $Q(x)$ is continuously differentiable, and denote its derivative by $q(x)$. Then $U(x)$, the probability that the outcome of the combined experiment is less than x, is given by

$$U(x) = \int_{-\infty}^{\infty} q(a)P(x-a)\,\mathrm{d}a. \tag{11.22}$$

The proof deduces Eq. (11.22) from Eq. (11.21). We assume that the outcome of the second experiment always lies in some finite interval I. We subdivide I into a finite number n of subintervals $I_j = [e_{j-1}, e_j]$. Let us denote by Q_j the probability that the outcome of the experiment Q lies in I_j. According to the mean value theorem,

$$Q_j = Q(e_j) - Q(e_{j-1}) = q(a_j)(e_j - e_{j-1}), \tag{11.23}$$

where a_j is some point in I_j.

We *discretize* the second experiment by lumping together all outcomes that lie in the interval I_j and *redefine* the numerical outcome in that case to be a_j, the number guaranteed to exist by the mean value theorem in Eq. (11.23). The probability of the outcome a_j is then the probability that the outcome lies in the interval I_j, i.e., it is Q_j.

Substitute for each Q_j the expressions given in Eq. (11.23). According to formula (11.21), the probability that the outcome of the discretized experiment is less than x is

$$U_n(x) = q(a_1)P(x-a_1)(e_1-e_0) + \cdots + q(a_n)P(x-a_n)(e_n - e_{n-1}).$$

The sum on the right is an approximating sum for the integral

$$\int_{-\infty}^{\infty} q(a)P(x-a)\,\mathrm{d}a.$$

This function was denoted by $U(x)$ in formula (11.22). Since approximating sums tend to the integral as the subdivision is made finer and finer, we conclude that for every x, $U_n(x)$ tends to $U(x)$. This proves our contention.

Now suppose that $P(x)$ is continuously differentiable, and denote its derivative by $p(x)$. It follows from Theorem 7.8 that $U(x)$ as defined by Eq. (11.22) is differentiable, and its derivative, which we denote by $u(x)$, can be obtained by differentiating the integrand with respect to x:

$$u(x) = \int_{-\infty}^{\infty} q(a)p(x-a)\,\mathrm{d}a. \tag{11.24}$$

We summarize what we have proved:

Theorem 11.3. *Consider two experiments whose outcomes lie in some finite interval and have probability densities p and q respectively. Suppose the experiments are independent. In the combined experiment consisting of performing both experiments, define the outcome of the combined experiment to be the sum of the outcomes of the individual experiments. Then the combined experiment has probability density $u(x)$ given by $u(x) = \int_{-\infty}^{\infty} q(a)p(x-a)\,\mathrm{d}a$.*

The restriction of the outcomes of the experiments to a finite interval is too confining for many important applications. Fortunately, the theorem, although not our proof, holds under more general conditions.

Definition 11.5. The function u defined by $u(x) = \int_{-\infty}^{\infty} q(a)p(x-a)\,\mathrm{d}a$ is called the *convolution* of the functions q and p. This relation is denoted by

$$u = q * p. \tag{11.25}$$

Example 11.10. Consider the following example of evaluating the convolution of two functions, where A and B are positive numbers.

$$p(a) = \begin{cases} 0 & \text{for } a < 0, \\ \mathrm{e}^{-Aa} & \text{for } 0 \le a, \end{cases} \qquad q(a) = \begin{cases} 0 & \text{for } a < 0, \\ \mathrm{e}^{-Ba} & \text{for } 0 \le a. \end{cases}$$

Substitute these definitions of p and q into the definition of the convolution:

$$u(x) = (p*q)(x) = \int_{-\infty}^{\infty} p(a)q(x-a)\,\mathrm{d}a.$$

Both $p(t)$ and $q(t)$ were defined to be zero for $t < 0$. It follows from this that the first factor in the integrand, $p(a)$, is zero for a negative. If $x < 0$, the second factor, $q(x-a)$, is zero for a positive. So for x negative, the integrand is zero for all values of a, and therefore so is the integral. This shows that $u(x) = 0$ for $x < 0$. For $x > 0$, the same analysis shows that the integrand is nonzero only in the range $0 \le a \le x$. So for $x > 0$,

$$u(x) = \int_0^x \mathrm{e}^{-Aa-B(x-a)}\,\mathrm{d}a$$

$$= \mathrm{e}^{-Bx}\int_0^x \mathrm{e}^{(B-A)a}\,\mathrm{d}a = \left[\mathrm{e}^{-Bx}\frac{\mathrm{e}^{(B-A)a}}{B-A}\right]_{a=0}^{x} = \frac{1}{B-A}(\mathrm{e}^{-Ax} - \mathrm{e}^{-Bx}).$$

Convolution is an important operation among functions, with many uses. We now state and prove some of its basic properties without any reference to probability.

Theorem 11.4. *Let $q_1(x)$, $q_2(x)$, and $p(x)$ be continuous functions defined for all real numbers x, and assume that the functions are zero outside a finite interval.*

(a) Convolution is distributive: $(q_1+q_2)*p = q_1*p+q_2*p$.
(b) Let k be any constant. Then $(kq)*p = k(q*p)$.
(c) Convolution is commutative: $q*p = p*q$.

Proof. The first result follows from the additivity of integrals:

$$(q_1+q_2)*p(x) = \int_{-\infty}^{\infty} \big(q_1(a)+q_2(a)\big)p(x-a)\,\mathrm{d}a$$

$$= \int_{-\infty}^{\infty} q_1(a)p(x-a)\,\mathrm{d}a + \int_{-\infty}^{\infty} q_2(a)p(x-a)\,\mathrm{d}a = q_1*p(x)+q_2*p(x).$$

The second result follows from

$$(kq)*p(x) = \int_{-\infty}^{\infty} kq(a)p(x-a)\,\mathrm{d}a = k\int_{-\infty}^{\infty} q(a)p(x-a)\,\mathrm{d}a = k(q*p)(x).$$

The third result follows if we make the change of variable $b = x-a$:

$$q*p(x) = \int_{-\infty}^{\infty} q(a)p(x-a)\,\mathrm{d}a = \int_{-\infty}^{\infty} q(x-b)p(b)\,\mathrm{d}b = p*q(x).$$

□

The following result is another basic property of convolution.

Theorem 11.5. *Suppose p and q are continuous functions, both zero outside some finite interval. Denote their convolution by u:*

$$u = p*q.$$

Then the integral of the convolution is the product of the integrals of the factors:

$$\int_{-\infty}^{\infty} u(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} p(x)\,\mathrm{d}x \int_{-\infty}^{\infty} q(a)\,\mathrm{d}a. \tag{11.26}$$

Proof. By definition of the convolution $u = p*q$,

$$u(x) = \int_{-\infty}^{\infty} p(x-a)q(a)\,\mathrm{d}a. \tag{11.27}$$

Suppose that the function $p(a)$ is zero outside the interval $I = [-b,b]$, so that $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = \int_{-b}^{b} p(x)\,\mathrm{d}x$, and $q(a)$ is zero outside the interval J. It follows that $u(x)$ is zero when x lies outside the interval $I \cup J$.

Approximate the integral (11.27) by the sum

$$u_n(x) = \sum_{j=1}^{n} p(x-a_j)q(a_j)(a_{j+1}-a_j), \tag{11.28}$$

where the numbers $a_1, \dots, a_n$ are n equally spaced points in the interval J of integration. It follows from the definition of integral as the limit of approximate sums that $u_n(x)$ tends to $u(x)$, uniformly for all x in the interval $I \cup J$. It follows that the integral of $u_n(x)$ with respect to x over $I \cup J$ tends to the integral of $u(x)$. It follows from formula (11.28) that the integral of $u_n(x)$ over $I \cup J$ is

$$\int_{-b}^{b} p(x)\,\mathrm{d}x \sum_{j=1}^{n} q(a_j)(a_{j+1}-a_j).$$

The limit of this sum as n tends to infinity is the integral of q. This concludes the proof of Eq. (11.26) in Theorem 11.5. □

The numerical outcome of the combination of two experiments was defined as the *sum* of the numerical outcomes of its two constituents. We now give some realistic examples to illustrate why this definition is of interest.

Suppose the outcomes of the two experiments represent *income* from two entirely different sources. Their sum is then the total income; its probability distribution is of considerable interest.

Here is another example: Suppose the two outcomes represent amounts of water entering a reservoir in a given period from two different sources. Their sum represents the total inflow, again an object of considerable interest.

Problems

11.14. We have said that the assumption $P(x)$ tends to 1 as x tends to infinity means that very large positive outcomes x are improbable. Justify that statement.

11.15. Define p by

$$p(x) = \begin{cases} 0 & \text{for } x < 0, \\ \frac{2}{A}\left(1 - \frac{x}{A}\right) & \text{for } 0 \le x \le A, \\ 0 & \text{for } A < x. \end{cases}$$

(a) Show that $\displaystyle\int_{-\infty}^{\infty} p(x)\,dx = 1$.

(b) Calculate the expected value of x, i.e., find $\overline{x} = \displaystyle\int_{-\infty}^{\infty} xp(x)\,dx$.

(c) Calculate the expected value $\overline{x^2} = \int_{-\infty}^{\infty} x^2 p(x)\,dx$.
(d) Give a definition of standard deviation and calculate it for this case.

11.16. Define p by

$$p(x) = k|x|\mathrm{e}^{-kx^2}, \quad k > 0.$$

Show that p is a probability density, i.e.,

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1.$$

11.17. Let A and B be two positive numbers. Define p and q by

$$p(t) = \begin{cases} 0 & \text{for } t < 0, \\ \frac{1}{A} & \text{for } 0 \le t \le A, \\ 0 & \text{for } A < t, \end{cases} \quad q(t) = \begin{cases} 0 & \text{for } t < 0, \\ \frac{1}{B} & \text{for } 0 \le t \le B, \\ 0 & \text{for } B < t. \end{cases}$$

(a) Show that p and q are probability densities, i.e., that they satisfy

$$\int_{-\infty}^{\infty} p(t)\,\mathrm{d}t = 1, \qquad \int_{-\infty}^{\infty} q(t)\,\mathrm{d}t = 1.$$

(b) Let u denote the convolution of p and q. Show that $u(x) = 0$ for $x < 0$ and for $x > A + B$.
(c) Verify that $u(x)$ is constant if $B < x < A$.
(d) Determine all values of $u(x)$ for the case $B < A$.

11.18. The purpose of this problem is to give an alternative proof of Theorem 11.5. Let p and q be a pair of functions, both zero outside some finite interval J. Let u be the convolution of p and q.

(a) Let h be a small number. Show that the sum

$$\sum_i p(ih)q(x - ih)h \tag{11.29}$$

is an approximating sum to the integral defining $u(x)$.
(b) Show that

$$\sum_j u(jh)h \tag{11.30}$$

is an approximating sum to the integral $\int_{-\infty}^{\infty} u(x)\,\mathrm{d}x$.
(c) Substitute the approximations (11.29) for $u(x)$ into Eq. (11.30) with $x = jh$. Show that the result is the *double sum* $\sum_{i,j} p(ih)q((j-i)h)h^2$.
(d) Denote $j - i$ by ℓ and rewrite the above double sum as $\sum_{i,\ell} p(ih)q(\ell h)h^2$.

(e) Show that this double sum can be written as the product of two single sums:

$$\left(\sum_i p(ih)h\right)\left(\sum_\ell q(\ell h)h\right).$$

(f) Show that the single sums are approximations to the integrals

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x \quad \text{and} \quad \int_{-\infty}^{\infty} q(x)\,\mathrm{d}x.$$

(g) Show that as h tends to zero, you obtain the identity in Theorem 11.5.

11.19. Define

$$|u|_1 = \int_{-\infty}^{\infty} |u(x)|\,\mathrm{d}x$$

as a quantity that measures the size of functions $u(x)$ that are defined for all x and are zero outside a finite interval.

(a) Evaluate $|u|_1$ if $u(x) = 5$ on $[a,b]$, and zero outside of $[a,b]$.
(b) Verify the properties $|cu|_1 = |c||u|_1$ when c is constant, and $|u+v|_1 \le |u|_1 + |v|_1$.
(c) Prove for convolution that $|u*v|_1 \le |u|_1|v|_1$.

11.4 The Law of Errors

In this section, we shall analyze a particular experiment. The experiment consists in dropping pellets from a fixed point at a certain height onto a horizontal plane. If the hand that releases the pellet were perfectly still and if there were no air currents diverting the pellet on its downward path, then we could predict with certainty that the pellet would end up directly below the point where it was released. But even the steadiest hand trembles a little, and even on the stillest day, minute air currents buffet the pellet in its downward flight, in a random fashion. These effects become magnified and very noticeable if the pellets are dropped from a great height, say the tenth floor of a building. Under such circumstances, the experiment appears to be nondeterministic, i.e., it is impossible to predict where each pellet is going to land.[1]

Although it is impossible to predict where any particular pellet would fall, the outcome can be predicted very well on average. That is, let G be any region such as a square, rectangle, triangle, or circle. Denote by $S(G)$ the number of instances

[1] G.I. Taylor (1886–1975), a famous British applied mathematician, described the following experience during the First World War: Taylor was working on a project to develop aerial darts; his task was to record the patterns created when a large number of darts were dropped from an airplane. This he did by putting a piece of paper under each dart where it had fallen in the field. These papers were to be photographed from the air. He had just finished this tedious task when a cavalry officer came by on horseback and demanded to know what Taylor was doing. Taylor explained the dart project, whereupon the officer exclaimed, “And you chaps managed to hit all those bits of paper? Good show!”

among the first N in a sequence of experiments in which the pellet landed in G. Then the frequencies $\frac{S(G)}{N}$ tend to a limit, called the probability of landing in G and denoted by $C(G)$:

$$\lim_{N\to\infty} \frac{S(G)}{N} = C(G).$$

In this section, we shall investigate the nature of this probability.

Suppose that the region G is a very small one. Then we expect the probability of landing in G to be nearly proportional to the area $A(G)$ of G. We can express this surmise more precisely as follows: Let g be any point in the plane. Then there is a number $c = c(g)$, called the *probability density* at g, such that for any region G containing g

$$C(G) = \big(c(g) + \text{small}\big)A(G),$$

where "small" means a quantity that tends to zero as G shrinks to the point g.

What can we say about the probability density $c(g)$? It depends on how close g is to the bullseye, i.e., the point directly underneath where the pellet is released. The closer g is, the greater the probability of a hit near g. In particular, the maximum value of c is achieved when g is the bullseye. We now adopt the following two hypotheses about the way in which the uncontrolled tremors of the hand and the unpredictable gusts of wind influence the distribution of hits and misses:

(i) $c(g)$ depends only on the distance of g from the bullseye, and not on the direction in which g lies.
(ii) Let x and y be perpendicular directions. Displacement of pellets in the x-direction is independent of their displacement in the y-direction.

Example 11.11. A special case illustrating hypothesis (ii) consists of two half-spaces bounded by lines through the origin in perpendicular directions. The probability of the pellet falling in either half-plane is $\frac{1}{2}$. The probability that the pellet falls in the quarter-plane that is the intersection of the two half-planes is $\frac{1}{4}$, and this is equal to $(\frac{1}{2})(\frac{1}{2})$.

To express these hypotheses in a mathematical form, we introduce a Cartesian coordinate system with the origin, naturally, at the bullseye. We denote by (a,d) the coordinates of the point g, as in Fig. 11.1. We denote by $P(a)$ the probability that the pellet falls in the half-plane

$$x < a.$$

The probability that the pellet falls in the strip $a \le x < b$ is

$$P(b) - P(a).$$

Assume that $P(a)$ has a continuous derivative for all a. We denote it by $p(a)$. According to the mean value theorem, the difference

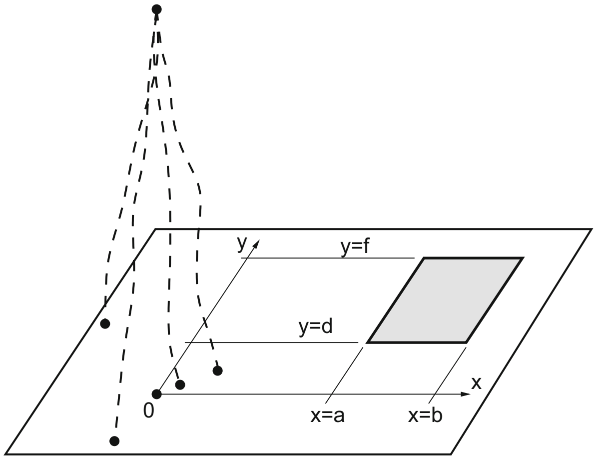

Fig. 11.1 Pellets dropped directly over the origin might land in the *shaded rectangle*

$$P(b) - P(a) \quad \text{is equal to} \quad p(a_1)(b-a),$$

for some number a_1 between a and b.

What is the probability that the pellet falls in the rectangle

$$a \le x < b, \quad d \le y < f?$$

This event occurs when the pellet falls in the strip $a \le x < b$ and the strip $d \le y < f$. According to hypothesis (ii), these two events are independent, and therefore, according to the *product rule*, the probability of the combined event is the product of the probabilities of the two separate events whose simultaneous occurrence constitutes the combined event. Thus the probability of a pellet falling in the rectangle is the product

$$p(a_1)(b-a)p(d_1)(f-d).$$

Since the product $(b-a)(f-d)$ is the area A of the rectangle, we can rewrite this as

$$p(a_1)p(d_1)A.$$

Now consider a sequence of rectangles that tend to the point $g = (a,d)$ by letting b tend to a and f tend to d. Since a_1 lies between a and b and d_1 lies between d and f, and since p is a continuous function, it follows that $p(a_1)$ tends to $p(a)$ and $p(d_1)$ tends to $p(d)$. Thus, in this case, we can express the probability that the pellet lands in the rectangle as

$$\big(p(a)p(d) + \text{small}\big)A.$$

We conclude that the probability density c at the point $g = (a, d)$ is

$$c(g) = p(a)p(d). \tag{11.31}$$

Next, we exploit the symmetry of the experimental setup around the bullseye by introducing another coordinate system, as in Fig. 11.2, whose origin is still the bullseye but where one of the coordinate axes is chosen to go through the point g whose coordinates in the old system were (a, d). The coordinates of g in the new system are

$$\left(0, \sqrt{a^2 + d^2}\right).$$

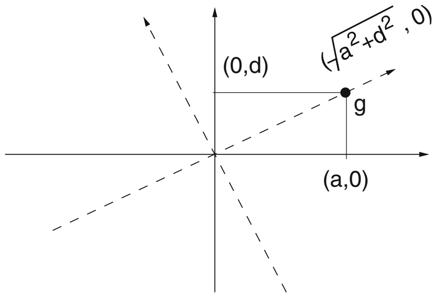

Fig. 11.2 A rotated coordinate system

According to hypothesis (i), we can apply relation (11.31) in any coordinate system. Then c in the new coordinate system is

$$c(g) = p(0)p\left(\sqrt{a^2 + d^2}\right). \tag{11.32}$$

Since the value of $c(g)$ expressed in two different coordinate systems as in Eqs. (11.31) and (11.32) are equal, we conclude that

$$p(a)p(d) = p(0)p\left(\sqrt{a^2 + d^2}\right). \tag{11.33}$$

This is a functional equation for $p(x)$. It can be solved by the trick of writing the function $p(x)$ in terms of another function $f(x) = \dfrac{p\left(\sqrt{x}\right)}{p(0)}$. Set $x = a^2$ and $y = d^2$ in Eq. (11.33), which gives

$$f(x)f(y) = f(x+y).$$

This, at last, is the familiar functional equation satisfied by exponential functions and only by them, as explained in Sect. 2.5c. So we conclude that $f(x) = \mathrm{e}^{Kx}$. Using the relation $f(x) = \dfrac{p\left(\sqrt{x}\right)}{p(0)}$, we deduce that $p(a) = p(0)\mathrm{e}^{Ka^2}$. We claim that the constant K is negative. For as a tends to infinity, the probability density $p(a)$ tends to zero, and this is the case only if K is negative. To put this into evidence, we rename K as $-k$, and rewrite

$$p(x) = p(0)\mathrm{e}^{-kx^2}. \tag{11.34}$$

Since p is a probability density, it satisfies $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$. Substituting Eq. (11.34) into this relation gives

$$p(0)\int_{-\infty}^{\infty} \mathrm{e}^{-kx^2}\,\mathrm{d}x = 1. \tag{11.35}$$

Introduce $y = \sqrt{2k}x$ as a new variable of integration. We get

$$\int_{-\infty}^{\infty} \mathrm{e}^{-kx^2}\,\mathrm{d}x = \frac{1}{\sqrt{2k}}\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{y^2}{2}}\,\mathrm{d}y. \tag{11.36}$$

It follows from Eq. (7.10) that $\int_{-\infty}^{\infty} \mathrm{e}^{-\frac{y^2}{2}}\,\mathrm{d}y = \sqrt{2\pi}$. Therefore,

$$\int_{-\infty}^{\infty} \mathrm{e}^{-kx^2}\,\mathrm{d}x = \sqrt{\frac{\pi}{k}}.$$

Setting this in Eq. (11.35) gives $p(0) = \sqrt{\frac{k}{\pi}}$. Therefore, using Eq. (11.34), we get

$$p(x) = \sqrt{\frac{k}{\pi}}\mathrm{e}^{-kx^2}. \tag{11.37}$$

Substituting this into $c(g) = p(a)p(d)$, we deduce

$$c(x,y) = \frac{k}{\pi}\mathrm{e}^{-k(x^2+y^2)}. \tag{11.38}$$

The derivation of the law of errors presented above is due to the physicist James Clerk Maxwell (1831–1879), who made profound investigations of the significance of probability densities of the form (11.37) and (11.38) in physics. For this reason, such densities in physics are called *Maxwellian*. Even before Maxwell, Carl Friedrich Gauss (1777–1855) investigated probabilities of the form (11.37). Mathematicians call such densities Gaussian. Another name for probabilities of this form is *normal*.

In Fig. 11.3, we see the shape of the normal distributions $p(x)$ for three different values: $k = 0.5,\ k = 1,\ k = 2$. These graphs indicate that the larger the value of k, the greater the concentration of the probability near the bullseye. The rest of this section is about some of the basic properties of normal distributions.

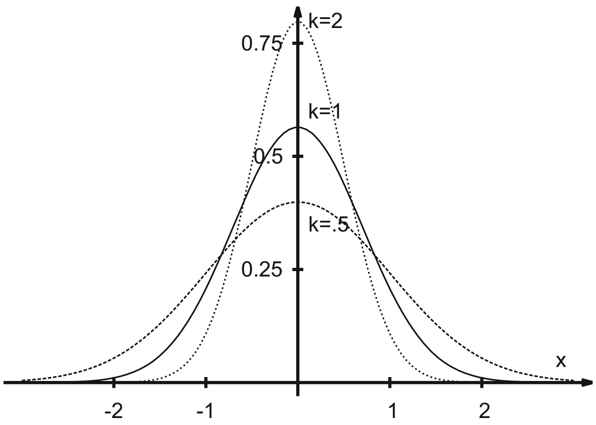

Fig. 11.3 Normal distributions with $k = 0.5$, $k = 1$, and $k = 2$

Theorem 11.6. *The convolution of two normal distributions is normal.*

Proof. We denote the two normal distributions by

$$p(x) = \sqrt{\frac{k}{\pi}}\mathrm{e}^{-kx^2}, \qquad \text{and} \quad q(x) = \sqrt{\frac{m}{\pi}}\mathrm{e}^{-mx^2}. \tag{11.39}$$

Their convolution is

$$(q * p)(x) = \int_{-\infty}^{\infty} q(a)p(x-a)\,\mathrm{d}a$$

$$= \frac{\sqrt{mk}}{\pi}\int_{-\infty}^{\infty} \mathrm{e}^{-ma^2-k(x-a)^2}\,\mathrm{d}a = \frac{\sqrt{mk}}{\pi}\mathrm{e}^{-kx^2}\int_{-\infty}^{\infty} \mathrm{e}^{-\left((m+k)a^2-2akx\right)}\,\mathrm{d}a.$$

To evaluate the integral, we complete the exponent under the integral sign to a perfect square:

$$(m+k)a^2 - 2akx = (m+k)\left(a - \frac{kx}{m+k}\right)^2 - \frac{k^2}{m+k}x^2.$$

Setting this into the integral above, we get, using $a - \dfrac{kx}{m+k} = b$ as the new variable of integration,

$$(q * p)(x) = \frac{\sqrt{mk}}{\pi}\mathrm{e}^{-k+\frac{k^2}{m+k}x^2}\int_{-\infty}^{\infty} \mathrm{e}^{-(m+k)b^2}\,\mathrm{d}b.$$

The integral is of the same form as the integral (11.36), with $(m+k)$ in place of k. Therefore, the value of the integral is $\sqrt{\dfrac{\pi}{m+k}}$. This gives

$$(q*p)(x) = \sqrt{\frac{1}{\pi}}\sqrt{\frac{mk}{m+k}}\mathrm{e}^{-(k-\frac{k^2}{m+k})x^2} = \sqrt{\frac{1}{\pi}}\sqrt{\frac{mk}{m+k}}\mathrm{e}^{-\frac{km}{m+k}x^2}.$$

We summarize: With p and q given by Eq. (11.39),

$$q*p = \sqrt{\frac{\ell}{\pi}}\mathrm{e}^{-\ell x^2}, \qquad \text{where} \quad \ell = \frac{km}{k+m}. \tag{11.40}$$

□

We turn next to the continuous analogue of Theorem 11.1 for discrete probability:

Theorem 11.7. *Among all probability densities $q(x)$ that satisfy*

$$\int_{-\infty}^{\infty} x^2 q(x)\,\mathrm{d}x = \frac{1}{2k},$$

the quantity $I(q) = -\int_{-\infty}^{\infty} q(x)\log q(x)\,\mathrm{d}x$ is largest for the Gaussian, i.e., when $q = p$ given by $p(x) = \sqrt{\dfrac{k}{\pi}}\mathrm{e}^{-kx^2}$.

Remarks.

(i) This result is a continuous analogue of Theorem 11.1, which asserts that among all probability distributions for n events, $-\sum_{1}^{n} p_j \log p_j$ is largest when all the p_j are equal. Our proof is similar to the proof given in the discrete case.
(ii) The functional $I(q)$ is the *entropy* of $q(x)$, an important quantity.
(iii) Implicit in the statement of the theorem is that

$$\int_{-\infty}^{\infty} x^2 p(x)\,\mathrm{d}x = \frac{1}{2k} \qquad \text{when} \quad p(x) = \sqrt{\frac{k}{\pi}}\mathrm{e}^{-kx^2},$$

which we ask you to derive in Problem 11.20.

Proof. We construct the following one-parameter family of probability densities $r(s)$ in the interval $0 \le s \le 1$:

$$r(s) = sq + (1-s)p. \tag{11.41}$$

This function is designed so that $r(0) = p$ and $r(1) = q$. To show that $I(p) \ge I(q)$, it suffices to verify that $I(r(s))$, which we abbreviate as $F(s)$, is a decreasing function of s. According to the monotonicity criterion, the decreasing character of $F(s)$

can be shown by verifying that the derivative of $F(s)$ is negative. To this end, we calculate the derivative of $F(s) = -\int_{-\infty}^{\infty} r(s)\log r(s)\,\mathrm{d}x$ using the differentiation theorem for integrals, Theorem 7.8. Differentiate $r(s)\log r(s)$ with respect to s; since $\frac{\mathrm{d}r}{\mathrm{d}s} = p - q$, we get

$$\frac{\mathrm{d}}{\mathrm{d}s}\left(r(s)\log r(s)\right) = \big(1+\log r(s)\big)\frac{\mathrm{d}r}{\mathrm{d}s} = \big(1+\log r(s)\big)(p-q).$$

Thus

$$\frac{\mathrm{d}F(s)}{\mathrm{d}s} = -\int_{-\infty}^{\infty}\big(1+\log r(s)\big)\left(p(x)-q(x)\right)\mathrm{d}x. \tag{11.42}$$

For $s=0$, we have $r(0) = p(x)$ and $\log p(x) = \log\sqrt{\frac{k}{\pi}} - kx^2$. So if we set $s=0$ in the derivative, we get

$$\frac{\mathrm{d}F}{\mathrm{d}s}(0) = -\int_{-\infty}^{\infty}\left(1+\log\sqrt{\frac{k}{\pi}} - kx^2\right)\left(p(x)-q(x)\right)\mathrm{d}x$$

$$= -\left(1+\log\sqrt{\frac{k}{\pi}}\right)\int_{-\infty}^{\infty}\left(p(x)-q(x)\right)\mathrm{d}x - k\int_{-\infty}^{\infty}x^2\left(p(x)-q(x)\right)\mathrm{d}x.$$

Since both p and q are probability densities, $\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} q(x)\,\mathrm{d}x = 1$. Furthermore, $\int_{-\infty}^{\infty} x^2 p(x)\,dx = \int_{-\infty}^{\infty} x^2 q(x)\,dx = \frac{1}{2k}$. Therefore, $\int_{-\infty}^{\infty}\left(p(x)-q(x)\right)\mathrm{d}x$ and $\int_{-\infty}^{\infty} x^2\left(p(x)-q(x)\right)\mathrm{d}x$ are both zero, and $\frac{\mathrm{d}F}{\mathrm{d}s}(0) = 0$. To show that $\frac{\mathrm{d}F}{\mathrm{d}s}(s) < 0$ for all s between 0 and 1, it suffices to show that $\frac{\mathrm{d}^2F}{\mathrm{d}s^2} < 0$. We now calculate the second derivative of F by again applying the differentiation theorem for integrals to Eq. (11.42). We get that

$$\frac{\mathrm{d}^2F(s)}{\mathrm{d}s^2} = \int_{-\infty}^{\infty}\left(p(x)-q(x)\right)\frac{1}{r(s)}\frac{\mathrm{d}r}{\mathrm{d}s}\,\mathrm{d}x = \int_{-\infty}^{\infty} -\frac{\left(p(x)-q(x)\right)^2}{r(s)}\,\mathrm{d}x.$$

This last integral is negative unless q and p are identical; hence the second derivative of F is negative. □

Remark. In our proof we have applied the differentiation theorem for integrals to improper integrals over the infinite interval $(-\infty,\infty)$, whereas this differentiation theorem was proved only for proper integrals. To get around this difficulty, we assume that $q(x)$ equals $p(x)$ outside a sufficiently large interval (a,b) and derive the inequality $I(q) \le I(p)$ for this subclass of q. From this, we can deduce the inequality for any q by approximating q by a sequence of q's belonging to the subclass. We omit the details of this step in the proof.

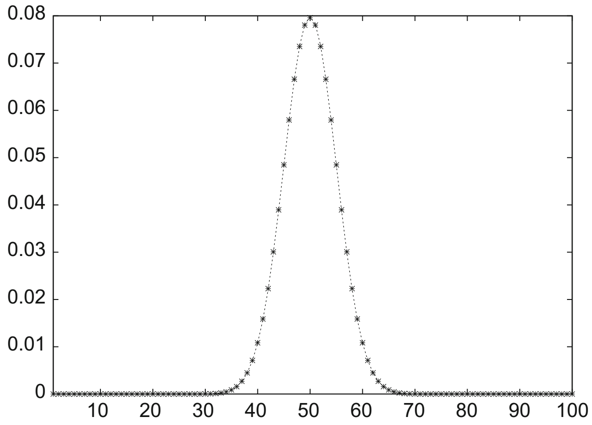

Fig. 11.4 The binomial distribution $b_k(100)$ and a normal distribution

The Limit of the Binomial Distribution. We have defined the binomial distribution as $b_k(n) = \binom{n}{k} p^k q^{n-k}$. To simplify the discussion, we take $p = q = \frac{1}{2}$. In this case,

$$b_k(n) = 2^{-n}\binom{n}{k}.$$

We have plotted these probabilities in Fig. 11.4 for $n = 100$ together with the function $\frac{1}{10}\sqrt{\frac{2}{\pi}}\mathrm{e}^{-2y^2}$, which is a multiple of the normal distribution. Note that the points b_k lie (nearly) on the graph of the normal distribution. The figures suggest that for n large, binomial distributions tend toward normal distributions. The precise statement is the following theorem.

Theorem 11.8. *The binomial distribution*

$$b_k(n) = 2^{-n}\binom{n}{k}$$

is approximately normal in this sense: set $y = \dfrac{k - \frac{1}{2}n}{\sqrt{n}}$. *Then*

$$b_k(n) \sim \frac{1}{\sqrt{n}}\sqrt{\frac{2}{\pi}}\mathrm{e}^{-2y^2},$$

where $\sim$ *means asymptotic as* n *and* k *tend to infinity with* y *fixed.*

In Problem 11.23, we guide you through a proof of this theorem. It is based on Stirling's formula (Theorem 7.5), which states that

$$m! \sim \sqrt{2\pi m}\left(\frac{m}{\mathrm{e}}\right)^m,$$

that is, the ratio of the left and right sides tends to 1 as m tends to infinity.

Problems

11.20. Integrate by parts to show that $\int_{-\infty}^{\infty} x^2 p(x)\,\mathrm{d}x = \frac{1}{2k}$ when $p(x) = \sqrt{\frac{k}{\pi}}\mathrm{e}^{-kx^2}$. Explain why this integral is called the variance of the normal distribution.

11.21. The purpose of this problem is to evaluate the integral $\int_{-\infty}^{\infty} \mathrm{e}^{-y^2}\,\mathrm{d}y$ numerically. Since the interval of integration is infinite, we truncate it by considering for large N, the approximate integral

$$I_{\text{mid}}\left(\mathrm{e}^{-y^2}, \left[-\left(N+\frac{1}{2}\right)h, \left(N+\frac{1}{2}\right)h\right]\right) = h\sum_{n=-N}^{N} \mathrm{e}^{-(nh)^2}$$

with subintervals of length h.

(a) Prove that $\sum_{n=K}^{\infty} \mathrm{e}^{-(nh)^2} \le \sum_{n=K}^{\infty} \mathrm{e}^{-Knh^2} = \frac{\mathrm{e}^{-K^2h^2}}{1-\mathrm{e}^{-Kh^2}}$. Use this to show that for $h=1$, the sum $\sum_{n=4}^{\infty} \mathrm{e}^{-(nh)^2}$ is less than 10^{-6}.

(b) Evaluate $I_{\text{mid}} \approx 1.77263\ldots$ numerically using just the sum from -3 to 3 with $h = 1$.

Remark. The value of the integral is $\sqrt{\pi} = 1.77245\ldots$. Thus we see that the midpoint rule gives an astonishingly good approximation to the value of the integral, even when we divide the interval of integration into subintervals of length $h = 1$, a rather crude subdivision.

11.22. Let

$$p(x,t) = \frac{1}{\sqrt{4\pi t}}\mathrm{e}^{-\frac{x^2}{4t}}.$$

(a) Determine the derivative of p with respect to x. Denote it by p_x.

(b) Determine the second derivative of p with respect to x. Denote it by p_{xx}.

(c) Determine the derivative of p with respect to t. Denote it by p_t.

(d) Verify that $p_t = p_{xx}$.

(e) In one application, p has an interpretation as the temperature of a metal rod, which varies with position and time. Suppose $p(x,t)$ is graphed as a function of x. In an interval of x where the graph is convex, will the temperature increase or decrease with time, according to part (d)?

11.23. Verify the following steps, which prove Theorem 11.8.

(a) Use Stirling's formula for each factorial in $2^{-n}\binom{n}{k}$ to show that

$$2^{-n}\binom{n}{k} \sim \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{k(n-k)}}\,\frac{n^n}{2^n k^k (n-k)^{n-k}}.$$

(b) Substitute $k = \frac{n}{2} + \sqrt{n}y$ and rearrange the previous expression to show that

$$2^{-n}\binom{n}{k} \sim \frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{\frac{n^2}{4} - ny^2}}\,\frac{1}{\left(1-\frac{4y^2}{n}\right)^{\frac{n}{2}}\left(\frac{n}{2}+\sqrt{n}y\right)^{\sqrt{n}y}\left(\frac{n}{2}-\sqrt{n}y\right)^{-\sqrt{n}y}}.$$

(c) We showed in Sect. 2.6 that $\left(1+\frac{x}{m}\right)^m$ tends to e^x as m tends to infinity. Use this to show that the right side in (b) is asymptotic to

$$\frac{1}{\sqrt{2\pi}}\sqrt{\frac{n}{\frac{n^2}{4} - ny^2}}\,\frac{1}{e^{-2y^2}e^{2y^2}e^{2y^2}}.$$

(d) Show that the last expression is asymptotic to $\dfrac{1}{\sqrt{n}}\sqrt{\dfrac{2}{\pi}}e^{-2y^2}$.