1

Preliminaries

1.1 Probability and Bayes' Theorem

1.1.1 Notation

The notation will be kept simple as possible, but it is useful to express statements about probability in the language of set theory. You probably know most of the symbols undermentioned, but if you do not you will find it easy enough to get the hang of this useful shorthand. We consider sets $A, B, C, \ldots$ of elements $x, y, z, \ldots$ and we use the word 'iff' to mean 'if and only if'. Then we write

$x \in A$ iff x is a member of A;

$x \notin A$ iff x is *not* a member of A;

$A = \{x, y, z\}$ iff A is the set whose only members are x, y and z (and similarly for larger or smaller sets);

$A = \{x;\ S(x)\}$ iff A is the set of elements for which the statement $S(x)$ is true;

$\emptyset = \{x;\ x \neq x\}$ for the null set, that is the set with no elements;

$x \notin \emptyset$ for all x;

$A \subset B$ (i.e. A is a subset of B) iff $x \in A$ implies $x \in B$;

$A \supset B$ (i.e. A is a superset of B) iff $x \in A$ is implied by $x \in B$;

$\emptyset \subset A$, $A \subset A$ and $A \supset A$ for all A;

$A \cup B = \{x;\ x \in A$ or $x \in B\}$ (where 'P or Q' means 'P or Q or both') (referred to as the union of A and B or as A union B);

$AB = A \cap B = \{x;\ x \in A$ and $x \in B\}$ (referred to as the intersection of A and B or as A intersect B);

A and B are disjoint iff $AB = \emptyset$;

$A \backslash B = \{x;\ x \in A,$ but $x \notin B\}$ (referred to as the difference set A less B).

Bayesian Statistics: An Introduction, Fourth Edition. Peter M. Lee.

Let (A_n) be a sequence $A_1, A_2, A_3, \ldots$ of sets. Then

$\bigcup_{n=1}^{\infty} A_n = \{x;\ x \in A_n \text{ for one or more } n\}$;
$\bigcap_{n=1}^{\infty} A_n = \{x;\ x \in A_n \text{ for all } n\}$;
(A_n) exhausts B if $\bigcup_{i=1}^{\infty} A_n \supset B$;
(A_n) consists of exclusive sets if $A_m A_n = \emptyset$ for $m \neq n$;
(A_n) consists of exclusive sets given B if $A_m A_n B = \emptyset$ for $m \neq n$;
(A_n) is non-decreasing if $A_1 \subset A_2 \subset \ldots$, that is $A_n \subset A_{n+1}$ for all n;
(A_n) is non-increasing if $A_1 \supset A_2 \supset \ldots$, that is $A_n \supset A_{n+1}$ for all n.

We sometimes need a notation for intervals on the real line, namely

$[a, b] = \{x;\ a \leqslant x \leqslant b\}$;
$(a, b) = \{x;\ a < x < b\}$;
$[a, b) = \{x;\ a \leqslant x < b\}$;
$(a, b] = \{x;\ a < x \leqslant b\}$

where a and b are real numbers or $+\infty$ or $-\infty$.

1.1.2 Axioms for probability

In the study of probability and statistics, we refer to as complete a description of the situation as we need in a particular context as an *elementary event*.

Thus, if we are concerned with the tossing of a red die and a blue die, then a typical elementary event is 'red three, blue five', or if we are concerned with the numbers of Labour and Conservative MPs in the next parliament, a typical elementary event is 'Labour 350, Conservative 250'. Often, however, we want to talk about one aspect of the situation. Thus, in the case of the first example, we might be interested in whether or not we get a red three, which possibility includes 'red three, blue one', 'red three, blue two', etc. Similarly, in the other example, we could be interested in whether there is a Labour majority of at least 100, which can also be analyzed into elementary events. With this in mind, an *event* is defined as a set of elementary events (this has the slightly curious consequence that, if you are *very* pedantic, an elementary event is not an event since it is an element rather than a set). We find it useful to say that one event E implies another event F if E is contained in F. Sometimes it is useful to generalize this by saying that, given H, E implies F if EH is contained in F. For example, given a red three has been thrown, throwing a blue three implies throwing an even total.

Note that the definition of an elementary event depends on the context. If we were never going to consider the blue die, then we could perfectly well treat events such as 'red three' as elementary events. In a particular context, the elementary events in terms of which it is sensible to work are usually clear enough.

Events are referred to above as possible future occurrences, but they can also describe present circumstances, known or unknown. Indeed, the relationship which probability attempts to describe is one between what you currently know and

something else about which you are uncertain, both of them being referred to as events. In other words, for at least some pairs of events E and H there is a number $\mathsf{P}(E|H)$ defined which is called the probability of the event E given the hypothesis H. I might, for example, talk of the probability of the event E that I throw a red three given the hypothesis H that I have rolled two fair dice once, or the probability of the event E of a Labour majority of at least 100 given the hypothesis H which consists of my knowledge of the political situation to date. Note that in this context, the term 'hypothesis' can be applied to a large class of events, although later on we will find that in statistical arguments, we are usually concerned with hypotheses which are more like the hypotheses in the ordinary meaning of the word.

Various attempts have been made to define the notion of probability. Many early writers claimed that $\mathsf{P}(E|H)$ was m/n where there were n symmetrical and so equally likely possibilities given H of which m resulted in the occurrence of E. Others have argued that $\mathsf{P}(E|H)$ should be taken as the long run frequency with which E happens when H holds. These notions can help your intuition in some cases, but I think they are impossible to turn into precise, rigourous definitions. The difficulty with the first lies in finding genuinely 'symmetrical' possibilities – for example, real dice are only approximately symmetrical. In any case, there is a danger of circularity in the definitions of symmetry and probability. The difficulty with the second is that we never know how long we have to go on trying before we are within, say, 1% of the true value of the probability. Of course, we may be able to give a value for the number of trials we need to be within 1% of the true value with, say, probability 0.99, but this is leading to another vicious circle of definitions. Another difficulty is that sometimes we talk of the probability of events (e.g. nuclear war in the next 5 years) about which it is hard to believe in a large numbers of trials, some resulting in 'success' and some in 'failure'. A good, brief discussion is to be found in Nagel (1939) and a fuller, more up-to-date one in Chatterjee (2003).

It seems to me, and to an increasing number of statisticians, that the only satisfactory way of thinking of $\mathsf{P}(E|H)$ is as a measure of my degree of belief in E given that I know that H is true. It seems reasonable that this measure should abide by the following axioms:

P1	$\mathsf{P}(E\|H)$	$\geqslant$	0 for all E, H.
P2	$\mathsf{P}(H\|H)$	$=$	1 for all H.
P3	$\mathsf{P}(E \cup F\|H)$	$=$	$\mathsf{P}(E\|H) + \mathsf{P}(F\|H)$ when $EFH = \emptyset$.
P4	$\mathsf{P}(E\|FH)\,\mathsf{P}(F\|H)$	$=$	$\mathsf{P}(EF\|H)$.

By taking $F = H \backslash E$ in P3 and using P1 and P2, it easily follows that

$$\mathsf{P}(E|H) \leqslant 1 \text{ for all } E, H,$$

so that $\mathsf{P}(E|H)$ is always between 0 and 1. Also by taking $F = \emptyset$ in P3 it follows that

$$\mathsf{P}(\emptyset|H) = 0.$$

Now intuitive notions about probability always seem to agree that it should be a quantity between 0 and 1 which falls to 0 when we talk of the probability of something we are certain will not happen and rises to 1 when we are certain it will happen (and we are certain that H is true given H is true). Further, the additive property in P3 seems highly reasonable – we would, for example, expect the probability that the red die lands three or four should be the sum of the probability that it lands three and the probability that it lands four.

Axiom P4 may seem less familiar. It is sometimes written as

$$\mathsf{P}(E|FH) = \frac{\mathsf{P}(EF|H)}{\mathsf{P}(F|H)}$$

although, of course, this form cannot be used if the denominator (and hence the numerator) on the right-hand side vanishes. To see that it is a reasonable thing to assume, consider the following data on criminality among the twin brothers or sisters of criminals [quoted in his famous book by Fisher (1925b)]. The twins were classified according as they had a criminal conviction (C) or not (N) and according as they were monozygotic (M) (which is more or less the same as identical – we will return to this in Section 1.2) or dizygotic (D), resulting in the following table:

	C	N	Total
M	10	3	13
D	2	15	17
Total	12	18	30

If we denote by H the knowledge that an individual has been picked at random from this population, then it seems reasonable to say that

$$\mathsf{P}(C|H) = 12/30,$$
$$\mathsf{P}(MC|H) = 10/30.$$

If on the other hand, we consider an individual picked at random from among the twins with a criminal conviction in the population, we see that

$$\mathsf{P}(M|CH) = 10/12$$

and hence

$$\mathsf{P}(M|CH)\mathsf{P}(C|H) = \mathsf{P}(MC|H),$$

so that P4 holds in this case. It is easy to see that this relationship does not depend on the particular numbers that happen to appear in the data.

In many ways, the argument in the preceding paragraph is related to derivations of probabilities from symmetry considerations, so perhaps it should be stressed that

while in certain circumstances we may believe in symmetries or in equally probable cases, we cannot base a general definition of probability on such arguments.

It is convenient to use a stronger form of axiom P3 in many contexts, namely,

$$\text{P3*} \qquad \mathsf{P}\left(\bigcup_{n=1}^{\infty} E_n \mid H\right) = \sum_{n=1}^{\infty} \mathsf{P}(E_n|H)$$

whenever the (E_n) are exclusive events given H. There is no doubt of the mathematical simplifications that result from this assumption, but we are supposed to be modelling our degrees of belief and it is questionable whether these have to obey this form of the axiom. Indeed, one of the greatest advocates of Bayesian theory, Bruno de Finetti, was strongly against the use of P3*. His views can be found in de Finetti (1972, Section 5.32) or in de Finetti (1974–1975, Section 3.11.3).

There is certainly some arbitrariness about P3*, which is sometimes referred to as an assumption of *σ-additivity*, in that it allows additivity over some but not all infinite collections of events (technically over countable but not over uncountable collections). However, it is impossible in a lot of contexts to allow additivity over any (arbitrary) collection of events. Thus, if we want a model for picking a point 'completely at random' from the unit interval

$$[0, 1] = \{x; 0 \leqslant x \leqslant 1\},$$

it seems reasonable that the probability that the point picked is in any particular sub-interval of the unit interval should equal the length of that sub-interval. However, this clearly implies that the probability of picking *any* one particular x is zero (since any such x belongs to intervals of arbitrarily small lengths). But the probability that *some* x is picked is unity, and it is impossible to get one by adding a lot of zeroes.

Mainly because of its mathematical convenience, we shall assume P3* while being aware of the problems.

1.1.3 'Unconditional' probability

Strictly speaking, there is, in my view, no such thing as an unconditional probability. However, it often happens that many probability statements are made conditional on everything that is part of an individual's knowledge at a particular time, and when many statements are to be made conditional on the same event, it makes for cumbersome notation to refer to this same conditioning event every time. There are also cases where we have so much experimental data in circumstances judged to be relevant to a particular situation that there is a fairly general agreement as to the probability of an event. Thus, in tossing a coin, you and I both have experience of tossing similar coins many times and so are likely to believe that 'heads' is approximately as likely as not, so that the probability of 'heads' is approximately $\frac{1}{2}$ given your knowledge or mine.

In these cases we write

$$\begin{aligned} \mathsf{P}(E) \quad &\text{for} \quad \mathsf{P}(E|\Omega), \\ \mathsf{P}(E|F) \quad &\text{for} \quad \mathsf{P}(E|F\Omega), \end{aligned}$$

where Ω is the set of possibilities consistent with the sum total of data available to the individual or individuals concerned. We usually consider sets F for which $F \subset \Omega$, so that $F\Omega = F$. It easily follows from the axioms that

$$\begin{gathered} 0 \leqslant \mathsf{P}(E) \leqslant 1, \\ \mathsf{P}(\Omega) = 1, \qquad \mathsf{P}(\emptyset) = 0, \\ \mathsf{P}\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mathsf{P}(E_n) \end{gathered}$$

whenever the (E_n) are exclusive events (or more properly whenever they are exclusive events given Ω), and

$$\mathsf{P}(E|F)\,\mathsf{P}(F) = \mathsf{P}(EF).$$

Many books begin by asserting that unconditional probability is an intuitive notion and use the latter formula in the form

$$\mathsf{P}(E|F) = \mathsf{P}(EF)/\mathsf{P}(F) \quad (\text{provided } \mathsf{P}(F) \neq 0)$$

to define conditional probability.

1.1.4 Odds

It is sometimes convenient to use a language more familiar to bookmakers to express probabilistic statements. We define the *odds on E against F given H* as the ratio

$$\mathsf{P}(E|H)/\mathsf{P}(F|H) \text{ to } 1$$

or equivalently

$$\mathsf{P}(E|H) \text{ to } \mathsf{P}(F|H).$$

A reference to the odds on E against F with no mention of H is to be interpreted as a reference to the odds on E against F given Ω, where Ω is some set of background knowledge as above.

Odds do not usually have properties as simple as probabilities, but sometimes, for example, in connection with Bayesian tests of hypotheses, they are more natural to consider than separate probabilities.

1.1.5 Independence

Two events E and F are said to be *independent* given H if

$$\mathsf{P}(EF|H) = \mathsf{P}(E|H)\,\mathsf{P}(F|H).$$

From axiom P4, it follows that if $\mathsf{P}(F|H) \neq 0$ this condition is equivalent to

$$\mathsf{P}(E|FH) = \mathsf{P}(E|H),$$

so that if E is independent of F given H then the extra information that F is true does not alter the probability of E from that given H alone, and this gives the best intuitive idea as to what independence means. However, the restriction of this interpretation to the case where $\mathsf{P}(F|H) \neq 0$ makes the original equation slightly more general.

More generally, a sequence (E_n) of events is said to be *pairwise independent* given H if

$$\mathsf{P}(E_m E_n|H) = \mathsf{P}(E_m|H)\,\mathsf{P}(E_n|H) \text{ for } m \neq n$$

and is said to consist of *mutually independent* events given H if for every finite subset of them

$$\mathsf{P}(E_{n_1} E_{n_2} \ldots E_{n_k}|H) = \mathsf{P}(E_{n_1}|H)\,\mathsf{P}(E_{n_2}|H)\ldots\mathsf{P}(E_{n_k}|H).$$

You should be warned that pairwise independence does not imply mutual independence and that

$$\mathsf{P}(E_1 E_2 \ldots E_n|H) = \mathsf{P}(E_1|H)\,\mathsf{P}(E_2|H)\ldots\mathsf{P}(E_n|H)$$

is not enough to ensure that the finite sequence $E_1, E_2, \ldots, E_n$ consists of mutually independent events given H.

Naturally, if no conditioning event is explicitly mentioned, the probabilities concerned are conditional on Ω as defined above.

1.1.6 Some simple consequences of the axioms; Bayes' Theorem

We have already noted a few consequences of the axioms, but it is useful at this point to note a few more. We first note that it follows simply from P4 and P2 and the fact that $HH = H$ that

$$\mathsf{P}(E|H) = \mathsf{P}(EH|H)$$

and in particular

$$\mathsf{P}(E) = \mathsf{P}(E\Omega).$$

Next note that if, given H, E implies F, that is $EH \subset F$ and so $EFH = EH$, then by P4 and the aforementioned equation

$$\mathsf{P}(E|FH)\,\mathsf{P}(F|H) = \mathsf{P}(EF|H) = \mathsf{P}(EFH|H) = \mathsf{P}(EH|H) = \mathsf{P}(E|H).$$

From this and the fact that $\mathsf{P}(E|FH) \leqslant 1$ it follows that if, given H, E implies F, then

$$\mathsf{P}(E|H) \leqslant \mathsf{P}(F|H).$$

In particular, if E implies F then

$$\mathsf{P}(E|F)\,\mathsf{P}(F) = \mathsf{P}(E),$$
$$\mathsf{P}(E) \leqslant \mathsf{P}(F).$$

For the rest of this subsection, we can work in terms of 'unconditional' probabilities, although the results are easily generalized. Let (H_n) be a sequence of exclusive and exhaustive events, and let E be any event. Then

$$\mathsf{P}(E) = \sum_n \mathsf{P}(E|H_n)\mathsf{P}(H_n)$$

since by P4 the terms on the right-hand side are $\mathsf{P}(EH_n)$, allowing us to deduce the result from P3*. This result is sometimes called the *generalized addition law* or the *law of the extension of the conversation*.

The key result in the whole book is Bayes' Theorem. This is simply deduced as follows. Let (H_n) be a sequence of events. Then by P4

$$\mathsf{P}(H_n|E)\mathsf{P}(E) = \mathsf{P}(EH_n) = \mathsf{P}(H_n)\mathsf{P}(E|H_n),$$

so that provided $\mathsf{P}(E) \neq 0$

$$\mathsf{P}(H_n|E) \propto \mathsf{P}(H_n)\,\mathsf{P}(E|H_n).$$

This relationship is one of several ways of stating Bayes' Theorem, and is probably the best way in which to remember it. When we need the constant of proportionality, we can easily see from the above that it is $1/\mathsf{P}(E)$.

It should be clearly understood that there is nothing controversial about Bayes' Theorem as such. It is frequently used by probabilists and statisticians, whether or not they are Bayesians. The distinctive feature of Bayesian statistics is the application of the theorem in a wider range of circumstances than is usual in classical statistics. In particular, Bayesian statisticians are always willing to talk of the probability of a hypothesis, both unconditionally (its *prior probability*) and given some evidence (its *posterior probability*), whereas other statisticians will only talk of the probability of a hypothesis in restricted circumstances.

When (H_n) consists of exclusive and exhaustive events, we can combine the last two results to see that

$$\mathsf{P}(H_n|E) = \frac{\mathsf{P}(H_n)\,\mathsf{P}(E|H_n)}{\sum_m \mathsf{P}(H_m)\,\mathsf{P}(E|H_m)}.$$

A final result that we will find useful from time to time is the *generalized multiplication law*, which runs as follows. If $H_1, H_2, \ldots, H_n$ are any events then

$$\begin{aligned}\mathsf{P}(H_1H_2\ldots H_n) = \mathsf{P}(H_1)\,\mathsf{P}(H_2|H_1)\,P(H_3|H_1H_2)\ldots \\ \mathsf{P}(H_n|H_1H_2\ldots H_{n-1})\end{aligned}$$

provided all the requisite conditional probabilities are defined, which in practice they will be provided $\mathsf{P}(H_1H_2\ldots H_{n-1}) \neq 0$. This result is easily proved by repeated application of P4.

1.2 Examples on Bayes' Theorem

1.2.1 The Biology of Twins

Twins can be either monozygotic (M) (i.e. developed from a single egg) or dizygotic (D). Monozygotic twins often look very similar and then are referred to as identical twins, but it is not always the case that one finds very striking similarities between monozygotic twins, while some dizygotic twins can show marked resemblances. Whether twins are monozygotic or dizygotic is not, therefore, a matter which can be settled simply by inspection. However, it is always the case that monozygotic twins are of the same sex, whereas dizygotic twins can be of opposite sex. Hence, assuming that the two sexes are equally probable, if the sexes of a pair of twins are denoted GG, BB or GB (note GB is indistinguishable from BG)

$$\begin{aligned}\mathsf{P}(GG|M) = \mathsf{P}(BB|M) = \tfrac{1}{2},\ \mathsf{P}(GB|M) = 0,\\ \mathsf{P}(GG|D) = \mathsf{P}(BB|D) = \tfrac{1}{4},\ \mathsf{P}(GB|D) = \tfrac{1}{2}.\end{aligned}$$

It follows that

$$\begin{aligned}\mathsf{P}(GG) &= \mathsf{P}(GG|M)\mathsf{P}(M) + \mathsf{P}(GG|D)\mathsf{P}(D)\\ &= \tfrac{1}{2}\mathsf{P}(M) + \tfrac{1}{4}\{1 - \mathsf{P}(M)\}\end{aligned}$$

from which it can be seen that

$$\mathsf{P}(M) = 4\mathsf{P}(GG) - 1,$$

so that although it is not easy to be certain whether a particular pair are monozygotic or not, it is easy to discover the *proportion* of monozygotic twins in the whole population of twins simply by observing the sex distribution among *all* twins.

1.2.2 A political example

The following example is a simplified version of the situation just before the time of the British national referendum as to whether the United Kingdom should remain part of the European Economic Community which was held in 1975. Suppose that at that date, which was shortly after an election which the Labour Party had won, the proportion of the electorate supporting Labour (L) stood at 52%, while the proportion supporting the Conservatives (C) stood at 48% (it being assumed for simplicity that support for all other parties was negligible, although this was far from being the case). There were many opinion polls taken at the time, so we can take it as known that 55% of Labour supporters and 85% of Conservative voters intended to vote 'Yes' (Y) and the remainder intended to vote 'No' (N). Suppose that knowing all this you met someone at the time who said that she intended to vote 'Yes', and you were interested in knowing which political party she supported. If this information were all you had available, you could reason as follows:

$$\begin{aligned}\mathsf{P}(L|Y) &= \frac{\mathsf{P}(Y|L)\mathsf{P}(L)}{\mathsf{P}(Y|L)\mathsf{P}(L) + \mathsf{P}(Y|C)\mathsf{P}(C)} \\ &= \frac{(0.55)(0.52)}{(0.55)(0.52) + (0.85)(0.48)} \\ &= 0.41.\end{aligned}$$

1.2.3 A warning

In the case of Connecticut v. Teal [see DeGroot *et al*. (1986, p. 9)], a case of alleged discrimination on the basis of a test to determine eligibility for promotion was considered. It turned out that of those taking the test 48 were black (B) and 259 were white (W), so that if we consider a random person taking the test

$$\mathsf{P}(B) = 48/307 = 0.16, \qquad \mathsf{P}(W) = 259/307 = 0.84.$$

Of the blacks taking the test, 26 passed (P) and the rest failed (F), whereas of the whites, 206 passed and the rest failed, so that altogether 232 people passed. Hence,

$$\mathsf{P}(B|P) = 26/232 = 0.11, \qquad \mathsf{P}(W|P) = 206/232 = 0.89.$$

There is a temptation to think that these are the figures which indicate the possibility of discrimination. Now there certainly is a case for saying that there was discrimination

in this case, *but* the figures that should be considered are

$$\mathsf{P}(P|B) = 26/48 = 0.54, \qquad \mathsf{P}(P|W) = 206/259 = 0.80.$$

It is easily checked that the probabilities here are related by Bayes' Theorem. It is worth while spending a while playing with hypothetical figures to convince yourself that the fact that $\mathsf{P}(B|P)$ is less than $\mathsf{P}(W|P)$ is irrelevant to the real question as to whether $\mathsf{P}(P|B)$ is less than $\mathsf{P}(P|W)$ – it might or might not be depending on the rest of the relevant information, that is on $\mathsf{P}(B)$ and $\mathsf{P}(W)$. The fallacy involved arises as the first of two well-known fallacies in criminal law which are both well summarized by Aitken (1996) (see also Aitken and Taroni, 2004, and Dawid, 1994) as follows:

> Suppose a crime has been committed. Blood is found at the scene for which there is no innocent explanation. It is of a type which is present in 1% of the population. The prosecutor may then state:
>
> 'There is a 1% chance that the defendant would have the crime blood type if he were innocent. Thus, there is a 99% chance that he is guilty'.
>
> Alternatively, the defender may state:
>
> 'This crime occurred in a city of 800,000 people. This blood type would be found in approximately 8000 people. The evidence has provided a probability of 1 in 8000 that the defendant is guilty and thus has no relevance.'

The first of these is known as the *prosecutor's fallacy* or the *fallacy of the transposed conditional* and, as pointed out above, in essence it consists in quoting the probability $\mathsf{P}(E|I)$ instead of $\mathsf{P}(I|E)$. The two are, however, equal if and only if the prior probability $\mathsf{P}(I)$ happens to equal $\mathsf{P}(\mathrm{E})$, which will only rarely be the case.

The second is the *defender's fallacy* which consists in quoting $\mathsf{P}(G|E)$ without regard to $\mathsf{P}(G)$. In the case considered by Aitken, the prior odds in favour of guilt are

$$\mathsf{P}(G)/\mathsf{P}(I) = 1/799\,999,$$

while the posterior odds are

$$\mathsf{P}(G|E)/\mathsf{P}(I|E) = 1/7\,999.$$

Such a large change in the odds is, in Aitken's words 'surely of relevance'. But, again in Aitken's words, 'Of course, it may not be enough to find the suspect guilty'.

As a matter of fact, Bayesian statistical methods are increasingly used in a legal context. Useful references are Balding and Donnelly (1995), Foreman, Smith and Evett (1997), Gastwirth (1988) and Fienberg (1989).

1.3 Random variables

1.3.1 Discrete random variables

As explained in Section 1.1, there is usually a set Ω representing the possibilities consistent with the sum total of data available to the individual or individuals concerned. Now suppose that with each elementary event ω in Ω, there is an integer $\widetilde{m}(\omega)$ which may be positive, negative or zero. In the jargon of mathematics, we have a function $\widetilde{m}$ mapping Ω to the set $\mathbb{Z}$ of all (signed) integers. We refer to the function as a *random variable* or an *r.v.*

A case arising in the context of the very first example we discussed, which was about tossing a red die and a blue die, is the integer representing the sum of the spots showing. In this case, ω might be 'red three, blue two' and then $\widetilde{m}(\omega)$ would be 5. Another case arising in the context of the second (political) example is the Labour majority (represented as a negative integer should the Conservatives happen to win), and here ω might be 'Labour 350, Conservative 250' in which case $\widetilde{m}(\omega)$ would be 100.

Rather naughtily, probabilists and statisticians tend not to mention the elementary event ω of which $\widetilde{m}(\omega)$ is a function and instead just write $\widetilde{m}$ for $\widetilde{m}(\omega)$. The reason is that what usually matters is the value of $\widetilde{m}$ rather than the nature of the elementary event ω, the definition of which is in any case dependent on the context, as noted earlier, in the discussion of elementary events. Thus, we write

$$\mathsf{P}(\widetilde{m} = m) \quad \text{for} \quad \mathsf{P}(\{\omega;\, \widetilde{m}(\omega) = m\})$$

for the probability that the random variable $\widetilde{m}$ takes the particular value m. It is a useful convention to use the same lower-case letter and drop the tilde ($\widetilde{\ }$) to denote a particular value of a random variable. An alternative convention used by some statisticians is to use capital letters for random variables and corresponding lower case letters for typical values of these random variables, but in a Bayesian context we have so many quantities that are regarded as random variables that this convention is too restrictive. Even worse than the habit of dropping mention of ω is the tendency to omit the tilde and so use the same notation for a random variable and for a typical value of it. While failure to mention ω rarely causes any confusion, the failure to distinguish between random variables and typical values of these random variables can, on occasion, result in real confusion. When there is any possibility of confusion, the tilde will be used in the text, but otherwise it will be omitted. Also, we will use

$$p(m) = \mathsf{P}(\widetilde{m} = m) = \mathsf{P}(\{\omega;\, \widetilde{m}(\omega) = m\})$$

for the probability that the random variable $\widetilde{m}$ takes the value m. When there is only one random variable we are talking about, this abbreviation presents few problems, but when we have a second random variable $\widetilde{n}$ and write

$$p(n) = \mathsf{P}(\widetilde{n} = n) = \mathsf{P}(\{\omega;\, \widetilde{n}(\omega) = n\})$$

then ambiguity can result. It is not clear in such a case what $p(5)$ would mean, or indeed what $p(i)$ would mean (unless it refers to $p(\widetilde{i} = i)$ where $\widetilde{i}$ is yet a third random variable). When it is necessary to resolve such an ambiguity, we will use

$$p_{\widetilde{m}}(m) = \mathsf{P}(\widetilde{m} = m) = \mathsf{P}(\{\omega;\, \widetilde{m}(\omega) = m\}),$$

so that, for example, $p_{\widetilde{m}}(5)$ is the probability that $\widetilde{m}$ is 5 and $p_{\widetilde{n}}(i)$ is the probability that $\widetilde{n}$ equals i. Again, all of this seems very much more confusing than it really is – it is usually possible to conduct arguments quite happily in terms of $p(m)$ and $p(n)$ and substitute numerical values at the end if and when necessary.

You could well object that you would prefer a notation that was free of ambiguity, and if you were to do so, I should have a lot of sympathy. But the fact is that constant references to $\widetilde{m}(\omega)$ and $p_{\widetilde{m}}(m)$ rather than to m and $p(m)$ would clutter the page and be unhelpful in another way.

We refer to the sequence $(p(m))$ as the *(probability) density (function)* or *pdf* of the random variable m (strictly $\widetilde{m}$). The random variable is said to have a distribution (of probability) and one way of describing a distribution is by its pdf. Another is by its *(cumulative) distribution function*, or *cdf* or *df*, defined by

$$F(m) = F_{\widetilde{m}}(m) = \mathsf{P}(\widetilde{m} \leqslant m) = \mathsf{P}(\{\omega;\, \widetilde{m}(\omega) \leqslant m\}) = \sum_{k \leqslant m} p_{\widetilde{m}}(k).$$

Because the pdf has the obvious properties

$$p(m) \geqslant 0, \qquad \sum_{m} p(m) = 1$$

the df is (weakly) increasing, that is

$$F(m) \leqslant F(m') \quad \text{if} \quad m \leqslant m'$$

and moreover

$$\lim_{m \to -\infty} F(m) = 0, \quad \lim_{m \to \infty} F(m) = 1.$$

1.3.2 The binomial distribution

A simple example of such a distribution is the *binomial distribution* (see Appendix A). Suppose, we have a sequence of trials each of which, independently of the others, results in success (S) or failure (F), the probability of success being a constant π (such trials are sometimes called Bernoulli trials). Then the probability of any particular sequence of n trials in which k result in success is

$$\pi^k (1 - \pi)^{n-k},$$

so that allowing for the $\binom{n}{k}$ ways in which k successes and $n-k$ failures can be ordered, the probability that a sequence of n trials results in k successes is

$$\binom{n}{k}\pi^k(1-\pi)^{n-k} \qquad (0 \leqslant k \leqslant n).$$

If then k (strictly $\widetilde{k}$) is a random variable defined as the number of successes in n trials, then

$$p(k) = \binom{n}{k}\pi^k(1-\pi)^{n-k}.$$

Such a distribution is said to be binomial of index n and parameter π, and we write

$$k \sim \mathrm{B}(n, \pi)$$

[or strictly $\widetilde{k} \sim \mathrm{B}(n, \pi)$].

We note that it follows immediately from the definition that if x and y are independent and $x \sim \mathrm{B}(m, \pi)$ and $y \sim \mathrm{B}(n, \pi)$ then $x + y \sim \mathrm{B}(m + n, \pi)$.

1.3.3 Continuous random variables

So far, we have restricted ourselves to random variables which take only integer values. These are particular cases of *discrete* random variables. Other examples of discrete random variables occur, for example, a measurement to the nearest quarter-inch which is subject to a distribution of error, but these can nearly always be changed to integer-valued random variables (in the given example simply by multiplying by 4). More generally, we can suppose that with each elementary event ω in Ω there is a real number $\widetilde{x}(\omega)$. We can define the (cumulative) distribution function, cdf or df of $\widetilde{x}$ by

$$F(x) = \mathsf{P}(\widetilde{x} \leqslant x) = \mathsf{P}(\{\omega;\, \widetilde{x}(\omega) \leqslant x\}).$$

As in the discrete case the df is (weakly) increasing, that is

$$F(x) \leqslant F(x') \quad \text{if} \quad x \leqslant x'$$

and moreover

$$\lim_{x\to-\infty} F(x) = 0, \qquad \lim_{x\to\infty} F(x) = 1.$$

It is usually the case that when $\widetilde{x}$ is not discrete there exists a function $p(x)$, or more strictly $p_{\widetilde{x}}(x)$, such that

$$F(x) = \int_{-\infty}^{x} p_{\widetilde{x}}(\xi)\mathrm{d}\xi$$

in which case $p(x)$ is called a (probability) density (function) or pdf. When this is so, x (strictly $\widetilde{x}$) is said to have a continuous distribution (or more strictly an absolutely continuous distribution). Of course, in the continuous case $p(x)$ is not itself interpretable directly as a probability, but for small δx

$$p(x)\delta x \cong \mathsf{P}(x < \widetilde{x} \leqslant x + \delta x) = \mathsf{P}(\{\omega; x < \widetilde{x}(\omega) \leqslant x + \delta x\}).$$

The quantity $p(x)\delta x$ is sometimes referred to as the *probability element*. Note that letting $\delta x \to 0$ this implies that

$$\mathsf{P}(\widetilde{x} = x) = 0$$

for every particular value x, in sharp contrast to the discrete case. We can also use the above approximation if y is some one-to-one function of x, for example

$$y = g(x).$$

Then if values correspond in an obvious way

$$\mathsf{P}(y < \widetilde{y} \leqslant y + \delta y) = \mathsf{P}(x < \widetilde{x} \leqslant x + \delta x)$$

which on substituting in the above relationship gives in the limit

$$p(y)\,|\mathrm{d}y| = p(x)\,|\mathrm{d}x|$$

which is the rule for *change of variable* in probability densities. (It is not difficult to see that, because the modulus signs are there, the same result is true if F is a strictly decreasing function of x). Another way of getting at this rule is by differentiating the obvious equation

$$F(y) = F(x)$$

[strictly $F_{\widetilde{y}}(y) = F_{\widetilde{x}}(x)$] which holds whenever y and x are corresponding values, that is $y = g(x)$. We should, however, beware that these results need modification if g is *not* a one-to-one function. In the continuous case, we can find the density from the df by differentiation, namely

$$p(x) = F'(x) = \mathrm{d}F(x)/\mathrm{d}x.$$

Although there are differences, there are many similarities between the discrete and the continuous cases, which we try to emphasize by using the same notation in both cases. We note that

$$F(x) = \sum_{\xi \leqslant x} p_{\tilde{x}}(\xi)$$

in the discrete case, but

$$F(x) = \int_{-\infty}^{x} p_{\tilde{x}}(\xi)\, \mathrm{d}\xi$$

in the continuous case. The discrete case is slightly simpler in one way in that no complications arise over change of variable, so that

$$p(y) = p(x)$$

if y and x are corresponding values, that is $y = g(x)$.

1.3.4 The normal distribution

The most important example of a continuous distribution is the so-called *normal* or Gaussian distribution. We say that z has a *standard normal* distribution if

$$p(z) = (2\pi)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}z^2\right)$$

and when this is so we write

$$z \sim \mathrm{N}(0, 1).$$

The density of this distribution is the familiar bell-shaped curve, with about two-thirds of the area between -1 and 1, 95% of the area between -2 and 2 and almost all of it between -3 and 3. Its distribution function is

$$\Phi(z) = \int_{-\infty}^{z} (2\pi)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}\zeta^2\right)\, \mathrm{d}\zeta.$$

More generally, we say that x has a normal distribution, denoted

$$x \sim \mathrm{N}(\mu, \phi)$$

if

$$x = \mu + z\sqrt{\phi}$$

where z is as aforementioned, or equivalently if

$$p(x) = (2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x-\mu)^2/\phi\right\}.$$

The normal distribution is encountered almost at every turn in statistics. Partly this is because (despite the fact that its density may seem somewhat barbaric at first sight) it is in many contexts the easiest distribution to work with, but this is not the whole story. The *Central Limit Theorem* says (roughly) that if a random variable can be expressed as a sum of a large number of components no one of which is likely to be much bigger than the others, these components being approximately independent, then this sum will be approximately normally distributed. Because of this theorem, an observation which has an error contributed to by many minor causes is likely to be normally distributed. Similar reasons can be found for thinking that in many circumstances we would expect observations to be approximately normally distributed, and this turns out to be the case, although there are exceptions. This is especially useful in cases where we want to make inferences about a population mean.

1.3.5 Mixed random variables

While most commonly occurring random variables are discrete or continuous, there are exceptions, for example the time you have to wait until you are served in a queue, which is zero with a positive probability (if the queue is empty when you arrive), but otherwise is spread over a continuous range of values. Such a random variable is said to have a *mixed* distribution.

1.4 Several random variables

1.4.1 Two discrete random variables

Suppose that with each elementary event ω in Ω, we can associate a pair of integers $(\widetilde{m}(\omega), \widetilde{n}(\omega))$. We write

$$p(m,n) = \mathsf{P}(\widetilde{m} = m, \widetilde{n} = n) = \mathsf{P}(\{\omega;\, \widetilde{m}(\omega) = m, \widetilde{n}(\omega) = n\}).$$

Strictly speaking, $p(m, n)$ should be written as $p_{\widetilde{m},\widetilde{n}}(m, n)$ for reasons discussed earlier, but this degree of pedantry in the notation is rarely necessary. Clearly

$$p(m,n) \geqslant 0, \qquad \sum_m \sum_n p(m,n) = 1.$$

The sequence $(p(m, n))$ is said to be a *bivariate (probability) density (function)* or *bivariate pdf* and is called the *joint pdf* of the random variables m and n (strictly $\widetilde{m}$

and $\widetilde{n}$). The corresponding *joint distribution function*, *joint cdf* or *joint df* is

$$F(m, n) = \sum_{k \leqslant m} \sum_{l \leqslant n} p_{\widetilde{m},\widetilde{n}}(k, l).$$

Clearly, the density of m (called its *marginal density*) is

$$p(m) = \sum_{n} p(m, n).$$

We can also define a conditional distribution for n given m (strictly for $\widetilde{n}$ given $\widetilde{m} = m$) by allowing

$$\begin{aligned} p(n|m) &= \mathsf{P}(\widetilde{n} = n \mid \widetilde{m} = m) = \mathsf{P}(\{\omega;\, \widetilde{n}(\omega) = n\} \mid \{\omega;\, \widetilde{m}(\omega) = m\}) \\ &= p(m, n)/p(m), \quad \text{provided } p(m) \neq 0 \end{aligned}$$

to define the *conditional (probability) density (function)* or *conditional pdf*. This represents our judgement as to the chance that $\widetilde{n}$ takes the value n given that $\widetilde{m}$ is known to have the value m. If it is necessary to make our notation absolutely precise, we can always write

$$p_{\widetilde{m}|\widetilde{n}}(m|n),$$

so, for example, $p_{\widetilde{m}|\widetilde{n}}(4|3)$ is the probability that m is 4 given $\widetilde{n}$ is 3, but $p_{\widetilde{n}|\widetilde{m}}(4|3)$ is the probability that $\widetilde{n}$ is 4 given that $\widetilde{m}$ takes the value 3, but it should be emphasized that we will not often need to use the subscripts. Evidently

$$p(n|m) \geqslant 0, \qquad \sum p(n|m) = 1,$$

and

$$p(n) = \sum p(m, n) = \sum p(m)p(n|m).$$

We can also define a *conditional distribution function* or *conditional df* by

$$\begin{aligned} F(n|m) &= \mathsf{P}(\widetilde{n} \leqslant n \mid \widetilde{m} = m) = \mathsf{P}(\{\omega;\, \widetilde{n}(\omega) \leqslant n\} \mid \{\omega;\, \widetilde{m}(\omega) = m\}) \\ &= \sum_{k \leqslant n} p_{\widetilde{n}|\widetilde{m}}(k|m). \end{aligned}$$

1.4.2 Two continuous random variables

As in Section 1.4, we have begun by restricting ourselves to integer values, which is more or less enough to deal with any discrete cases that arise. More generally, we can

suppose that with each elementary event ω in Ω, we can associate a pair $(\widetilde{x}(\omega), \widetilde{y}(\omega))$ of real numbers. In this case, we define the *joint distribution function* or *joint df* as

$$F(x, y) = \mathsf{P}(\widetilde{x} \leqslant x,\ \widetilde{y} \leqslant y) = \mathsf{P}(\{\omega;\ \widetilde{x}(\omega) \leqslant x,\ \widetilde{y}(\omega) \leqslant y\}).$$

Clearly the df of x is

$$F(x, +\infty),$$

and that of y is

$$F(+\infty, y).$$

It is usually the case that when neither x nor y is discrete there is a function $p(x, y)$ such that

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} p_{\widetilde{x},\widetilde{y}}(\xi, \eta)\, \mathrm{d}\xi\ \mathrm{d}\eta,$$

in which case $p(x, y)$ is called a *joint (probability) density (function)* or *joint pdf*. When this is so, the joint distribution is said to be *continuous* (or more strictly to be absolutely continuous). We can find the density from the df by

$$p(x, y) = \partial^2 F(x, y)/\partial x \partial y.$$

Clearly,

$$p(x, y) \geqslant 0, \quad \iint p(x, y)\mathrm{d}x\ \mathrm{d}y = 1$$

and

$$p(x) = \int p(x, y)\, \mathrm{d}y.$$

The last formula is the continuous analogue of

$$p(m) = \sum_{n} p(m, n)$$

in the discrete case.

By analogy with the discrete case, we define the *conditional density* of y given x (strictly of $\widetilde{y}$ given $\widetilde{x} = x$) as

$$p(y|x) = p(x, y)/p(x),$$

provided $p(x) \neq 0$. We can then define the *conditional distribution function* by

$$F(y|x) = \int_{-\infty}^{y} p(\eta|x)\,\mathrm{d}\eta.$$

There are difficulties in the notion of conditioning on the event that $\widetilde{x} = x$ because this event has probability zero for every x in the continuous case, and it can help to regard the above distribution as the limit of the distribution which results from conditioning on the event that $\widetilde{x}$ is between x and $x + \delta x$, that is

$$\{\omega;\, x < \widetilde{x}(\omega) \leqslant x + \delta x\}$$

as $\delta x \to 0$.

1.4.3 Bayes' Theorem for random variables

It is worth noting that conditioning the random variable y by the value of x does not change the *relative* sizes of the probabilities of those pairs (x, y) that can still occur. That is to say, the probability $p(y|x)$ is proportional to $p(x, y)$ and the constant of proportionality is just what is needed, so that the conditional probabilities integrate to unity. Thus,

$$p(y|x) \geqslant 0, \qquad \int p(y|x)\,\mathrm{d}y = 1.$$

Moreover,

$$p(y) = \int p(x, y)\,\mathrm{d}x = \int p(x)p(y|x)\,\mathrm{d}x.$$

It is clear that

$$p(y|x) = p(x, y)/p(x) = p(y)p(x|y)/p(x),$$

so that

$$p(y|x) \propto p(y)p(x|y).$$

This is, of course, a form of *Bayes' Theorem*, and is in fact the commonest way in which it occurs in this book. Note that it applies equally well if the variables x and y are continuous or if they are discrete. The constant of proportionality is

$$1/p(x) = 1 \Big/ \int p(y)p(x|y)\,\mathrm{d}y$$

in the continuous case or

$$1/p(x) = 1 \Big/ \sum_y p(y)p(x|y)$$

in the discrete case.

1.4.4 Example

A somewhat artificial example of the use of this formula in the continuous case is as follows. Suppose y is the time before the first occurrence of a radioactive decay which is measured by an instrument, but that, because there is a delay built into the mechanism, the decay is recorded as having taken place at a time $x > y$. We actually have the value of x, but would like to say what we can about the value of y on the basis of this knowledge. We might, for example, have

$$\begin{aligned} p(y) &= \mathrm{e}^{-y} && (0 < y < \infty), \\ p(x|y) &= k\mathrm{e}^{-k(x-y)} && (y < x < \infty). \end{aligned}$$

Then

$$\begin{aligned} p(y|x) &\propto p(y)p(x|y) \\ &\propto \mathrm{e}^{(k-1)y} \qquad (0 < y < x). \end{aligned}$$

Often we will find that it is enough to get a result up to a constant of proportionality, but if we need the constant, it is very easy to find it because we know that the integral (or the sum in the discrete case) must be one. Thus, in this case

$$p(y|x) = \frac{(k-1)\mathrm{e}^{(k-1)y}}{\mathrm{e}^{(k-1)x} - 1} \qquad (0 < y < x).$$

1.4.5 One discrete variable and one continuous variable

We also encounter cases where we have two random variables, one of which is continuous and one of which is discrete. All the aforementioned definitions and formulae extend in an obvious way to such a case provided we are careful, for example, to use integration for continuous variables but summation for discrete variables. In particular, the formulation

$$p(y|x) \propto p(y)p(x|y)$$

for Bayes' Theorem is valid in such a case.

It may help to consider an example (again a somewhat artificial one). Suppose k is the number of successes in n Bernoulli trials, so $k \sim \mathrm{B}(n, \pi)$, but that the value

of π is unknown, your beliefs about it being uniformly distributed over the interval [0, 1] of possible values. Then

$$\begin{aligned} p(k|\pi) &= \binom{n}{k}\pi^k(1-\pi)^{n-k} \qquad (k = 0, 1, \ldots, n), \\ p(\pi) &= 1 \qquad\qquad\qquad\qquad\quad (0 \leqslant \pi \leqslant 1), \end{aligned}$$

so that

$$\begin{aligned} p(\pi|k) &\propto p(\pi)p(k|\pi) = \binom{n}{k}\pi^k(1-\pi)^{n-k} \\ &\propto \pi^k(1-\pi)^{n-k}. \end{aligned}$$

The constant can be found by integration if it is required. Alternatively, a glance at Appendix A will show that, given k, π has a beta distribution

$$\pi|k \sim \text{Be}(k+1, n-k+1)$$

and that the constant of proportionality is the reciprocal of the beta function $\text{B}(k+1, n-k+1)$. Thus, this beta distribution should represent your beliefs about π after you have observed k successes in n trials. This example has a special importance in that it is the one which Bayes himself discussed.

1.4.6 Independent random variables

The idea of independence extends from independence of events to independence of random variables. The basic idea is that y is independent of x if being told that x has any particular value does not affect your beliefs about the value of y. Because of complications involving events of probability zero, it is best to adopt the formal definition that x and y are independent if

$$p(x, y) = p(x)p(y)$$

for all values x and y. This definition works equally well in the discrete and the continuous cases (and indeed in the case where one random variable is continuous and the other is discrete). It trivially suffices that $p(x, y)$ be a product of a function of x and a function of y.

All the above generalizes in a fairly obvious way to the case of more than two random variables, and the notions of pairwise and mutual independence go through from events to random variables easily enough. However, we will find that we do not often need such generalizations.

1.5 Means and variances

1.5.1 Expectations

Suppose that m is a discrete random variable and that the series

$$\sum mp(m)$$

is absolutely convergent, that is such that

$$\sum |m|p(m) < \infty.$$

Then the sum of the original series is called the *mean* or *expectation* of the random variable, and we denote it

$$\mathsf{E}m \text{ (strictly } \mathsf{E}\widetilde{m}).$$

A motivation for this definition is as follows. In a large number N of trials, we would expect the value m to occur about $p(m)N$ times, so that the sum total of the values that would occur in these N trials (counted according to their multiplicity) would be about

$$\sum mp(m)N,$$

so that the average value should be about

$$\sum mp(m)N/N = \mathsf{E}m.$$

Thus, we can think of expectation as being, at least in some circumstances, a form of very long term average. On the other hand, there are circumstances in which it is difficult to believe in the possibility of arbitrarily large numbers of trials, so this interpretation is not always available. It can also be thought of as giving the position of the 'centre of gravity' of the distribution imagined as a distribution of mass spread along the x-axis.

More generally, if $g(m)$ is a function of the random variable and $\sum g(m)p(m)$ is absolutely convergent, then its sum is the expectation of $g(m)$. Similarly, if $h(m, n)$ is a function of two random variables m and n and the series $\sum\sum h(m, n)p(m, n)$ is absolutely convergent, then its sum is the expectation of $h(m, n)$. These definitions are consistent in that if we consider $g(m)$ and $h(m, n)$ as random variables with densities of their own, then it is easily shown that we get these values for their expectations.

In the continuous case, we define the expectation of a random variable x by

$$\mathsf{E}x = \int xp(x)\,\mathrm{d}x$$

provided that the integral is absolutely convergent, and more generally define the expectation of a function $g(x)$ of x by

$$\mathsf{E}g(x) = \int g(x)p(x)\,\mathrm{d}x$$

provided that the integral is absolutely convergent, and similarly for the expectation of a function $h(x, y)$ of two random variables. Note that the formulae in the discrete and continuous cases are, as usual, identical except for the use of summation in the one case and integration in the other.

1.5.2 The expectation of a sum and of a product

If x and y are any two random variables, independent or not, and a, b and c are constants, then in the continuous case

$$\begin{aligned}\mathsf{E}\{ax + by + c\} &= \iint (ax + by + c)p(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= a\iint xp(x, y)\,\mathrm{d}x\,\mathrm{d}y + b\iint yp(x, y)\,\mathrm{d}x\,\mathrm{d}y + c\iint p(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= a\int xp(x)\,\mathrm{d}x + b\int yp(y)\,\mathrm{d}y + c \\ &= a\mathsf{E}x + b\mathsf{E}y + c\end{aligned}$$

and similarly in the discrete case. Yet more generally, if $g(x)$ is a function of x and $h(y)$ a function of y, then

$$\mathsf{E}\{ag(x) + bh(y) + c\} = a\mathsf{E}g(x) + b\mathsf{E}h(y) + c.$$

We have already noted that the idea of independence is closely tied up with multiplication, and this is true when it comes to expectations as well. Thus, if x and y are independent, then

$$\begin{aligned}\mathsf{E}xy &= \iint xy\; p(x, y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \iint xy\; p(x)p(y)\,\mathrm{d}x\,\mathrm{d}y \\ &= \left(\int xp(x)\,\mathrm{d}x\right)\left(\int yp(y)\,\mathrm{d}y\right) \\ &= (\mathsf{E}x)(\mathsf{E}y)\end{aligned}$$

and more generally if $g(x)$ and $h(y)$ are functions of independent random variables x and y, then

$$\mathsf{E}g(x)h(y) = (\mathsf{E}g(x))(\mathsf{E}h(y)).$$

1.5.3 Variance, precision and standard deviation

We often need a measure of how spread out a distribution is, and for most purposes the most useful such measure is the variance $\mathsf{V}x$ of x, defined by

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

Clearly if the distribution is very little spread out, then most values are close to one another and so close to their mean, so that $(x - \mathsf{E}x)^2$ is small with high probability and hence $\mathsf{V}x$ is small. Conversely, if the distribution is well spread out then $\mathsf{V}x$ is large. It is sometimes useful to refer to the reciprocal of the variance, which is called the *precision*. Further, because the variance is essentially quadratic, we sometimes work in terms of its positive square root, the *standard deviation*, especially in numerical work. It is often useful that

$$\begin{aligned}\mathsf{V}x &= \mathsf{E}(x - \mathsf{E}x)^2 \\ &= \mathsf{E}\left(x^2 - 2(\mathsf{E}x)x + (\mathsf{E}x)^2\right) \\ &= \mathsf{E}x^2 - (\mathsf{E}x)^2.\end{aligned}$$

The notion of a variance is analogous to that of a moment of inertia in mechanics, and this formula corresponds to the *parallel axes theorem* in mechanics. This analogy seldom carries much weight nowadays, because so many of those studying statistics took it up with the purpose of avoiding mechanics.

In discrete cases, it is sometimes useful that

$$\mathsf{V}x = \mathsf{E}x(x-1) + \mathsf{E}x - (\mathsf{E}x)^2.$$

1.5.4 Examples

As an example, suppose that $k \sim \mathrm{B}(n, \pi)$. Then

$$\mathsf{E}k = \sum_{k=0}^{n} k \binom{n}{k} \pi^k (1-\pi)^{n-k}.$$

After a little manipulation, this can be expressed as

$$\mathsf{E}k = n\pi \sum_{j=0}^{n-1} \binom{n-1}{j} \pi^j (1-\pi)^{n-1-j}.$$

Because the sum is a sum of binomial $\mathrm{B}(n-1,\pi)$ probabilities, this expression reduces to $n\pi$, and so

$$\mathsf{E}k = n\pi.$$

Similarly,

$$\mathsf{E}k(k-1) = n(n-1)\pi^2$$

and so

$$\begin{aligned}\mathscr{V}k &= n(n-1)\pi^2 + n\pi - (n\pi)^2 \\ &= n\pi(1-\pi).\end{aligned}$$

For a second example, suppose $x \sim \mathrm{N}(\mu,\phi)$. Then

$$\begin{aligned}\mathsf{E}x &= \int x(2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x-\mu)^2/\phi\right\} \mathrm{d}x \\ &= \mu + \int (x-\mu)(2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x-\mu)^2/\phi\right\} \mathrm{d}x.\end{aligned}$$

The integrand in the last expression is an odd function of $x-\mu$ and so vanishes, so that

$$\mathsf{E}x = \mu.$$

Moreover,

$$\mathscr{V}x = \int (x-\mu)^2(2\pi\phi)^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(x-\mu)^2/\phi\right\} \mathrm{d}x,$$

so that on writing $z = (x-\mu)/\sqrt{\phi}$

$$\mathscr{V}x = \phi \int z^2(2\pi)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}z^2\right) \mathrm{d}z.$$

Integrating by parts (using z as the part to differentiate), we get

$$\begin{aligned}\mathscr{V}x &= \phi \int (2\pi)^{-\frac{1}{2}} \exp\left(-\tfrac{1}{2}z^2\right) \mathrm{d}z \\ &= \phi.\end{aligned}$$

1.5.5 Variance of a sum; covariance and correlation

Sometimes we need to find the *variance of a sum* of random variables. To do this, note that

$$\begin{aligned}\mathscr{V}(x+y) &= \mathsf{E}\{x+y-\mathsf{E}(x+y)\}^2 \\ &= \mathsf{E}\{(x-\mathsf{E}x)+(y-\mathsf{E}y)\}^2 \\ &= \mathsf{E}(x-\mathsf{E}x)^2+\mathsf{E}(y-\mathsf{E}y)^2+2\mathsf{E}(x-\mathsf{E}x)(y-\mathsf{E}y) \\ &= \mathscr{V}x+\mathscr{V}y+2\mathscr{C}(x,y),\end{aligned}$$

where the *covariance* $\mathscr{C}(x, y)$ of x and y is defined by

$$\begin{aligned}\mathscr{C}(x,y) &= \mathsf{E}(x-\mathsf{E}x)(y-\mathsf{E}y) \\ &= \mathsf{E}xy-(\mathsf{E}x)(\mathsf{E}y).\end{aligned}$$

More generally,

$$\mathscr{V}(ax+by+c) = a^2\mathscr{V}x+b^2\mathscr{V}y+2ab\,\mathscr{C}(x,y)$$

for any constants a, b and c. By considering this expression as a quadratic in a for fixed b or vice versa and noting that (because its value is always positive) this quadratic cannot have two unequal real roots, we see that

$$(\mathscr{C}(x,y))^2 \leqslant (\mathscr{V}x)(\mathscr{V}y).$$

We define the correlation coefficient $\rho(x, y)$ between x and y by

$$\rho(x,y) = \frac{\mathscr{C}(x,y)}{\sqrt{(\mathscr{V}x)(\mathscr{V}y)}}.$$

It follows that

$$-1 \leqslant \rho(x,y) \leqslant 1$$

and indeed a little further thought shows that $\rho(x, y) = 1$ if and only if

$$ax+by+c=0$$

with probability 1 for some constants a, b and c with a and b having opposite signs, while $\rho(x, y) = -1$ if and only if the same thing happens except that a and b have the same sign. If $\rho(x, y) = 0$ we say that x and y are uncorrelated.

It is easily seen that if x and y are independent then

$$\mathscr{C}(x,y) = \mathsf{E}xy-(\mathsf{E}x)(\mathsf{E}y) = 0$$

from which it follows that independent random variables are uncorrelated.

The converse is *not* in general true, but it can be shown that *if* x and y have a bivariate normal distribution (as described in Appendix A), then they are independent if and only if they are uncorrelated.

It should be noted that if x and y are uncorrelated, and in particular if they are independent

$$\mathcal{V}(x \pm y) = \mathcal{V}x + \mathcal{V}y$$

(observe that there is a plus sign on the right-hand side even if there is a minus sign on the left).

1.5.6 Approximations to the mean and variance of a function of a random variable

Very occasionally, it will be useful to have an approximation to the mean and variance of a function of a random variable. Suppose that

$$z = g(x).$$

Then if g is a reasonably smooth function and x is not too far from its expectation, Taylor's theorem implies that

$$z \cong g(\mathsf{E}x) + (x - \mathsf{E}x)g'(\mathsf{E}x).$$

It, therefore, seems reasonable that a fair approximation to the expectation of z is given by

$$\mathsf{E}z = g(\mathsf{E}x)$$

and if this is so, then a reasonable approximation to $\mathcal{V}z$ may well be given by

$$\mathcal{V}z = \mathcal{V}x\{g'(\mathsf{E}x)\}^2.$$

As an example, suppose that

$$x \sim \mathrm{B}(n, \pi)$$

and that $z = g(x)$, where

$$g(x) = \sin^{-1}\sqrt{(x/n)},$$

so that

$$g'(x) = \tfrac{1}{2}n^{-1}[(x/n)\{1 - (x/n)\}]^{-\frac{1}{2}},$$

and thus $g'(\mathsf{E}x) = g'(n\pi) = \frac{1}{2}n^{-1}[\pi(1-\pi)]^{-\frac{1}{2}}$. The aforementioned argument then implies that

$$\mathsf{E}z \cong \sin^{-1}\sqrt{\pi}, \qquad \mathscr{V}z \cong 1/4n.$$

The interesting thing about this transformation, which has a long history [see Eisenhart *et al*. (1947, Chapter 16) and Fisher (1954)], is that, to the extent to which the approximation is valid, the variance of z does not depend on the parameter π. It is accordingly known as a *variance-stabilizing transformation*. We will return to this transformation in Section 3.2 on the 'Reference Prior for the Binomial Distribution'.

1.5.7 Conditional expectations and variances

If the reader wishes, the following may be omitted on a first reading and then returned to as needed.

We define the *conditional expectation* of y given x by

$$\mathsf{E}(y|x) = \int yp(y|x)\,\mathrm{d}y$$

in the continuous case and by the corresponding sum in the discrete case. If we wish to be pedantic, it can occasionally be useful to indicate what we are averaging over by writing

$$\mathsf{E}_{\widetilde{y}|\widetilde{x}}(\widetilde{y}|x)$$

just as we can write $p_{\widetilde{y}|\widetilde{x}}(y|x)$, but this is rarely necessary (though it can slightly clarify a proof on occasion). More generally, the conditional expectation of a function $g(y)$ of y given x is

$$\mathsf{E}(g(y)|x) = \int g(y)p(y|x)\,\mathrm{d}y.$$

We can also define a conditional variance as

$$\begin{aligned}\mathscr{V}(y|x) &= \mathsf{E}[\{y - \mathsf{E}(y|x)\}^2 \,|\, x] \\ &= \mathsf{E}(y^2|x) - \{\mathsf{E}(y|x)\}^2.\end{aligned}$$

Despite some notational complexity, this is easy enough to find since after all a conditional distribution is just a particular case of a probability distribution. If we are really pedantic, then $\mathsf{E}(\widetilde{y}|x)$ is a real number which is a function of the real number x, while $\mathsf{E}(\widetilde{y}|\widetilde{x})$ is a random variable which is a function of the random variable $\widetilde{x}$, which takes the value $\mathsf{E}(\widetilde{y}|x)$ when $\widetilde{x}$ takes the value x. However, the distinction, which is hard to grasp in the first place, is usually unimportant.

We may note that the formula

$$p(y) = \int p(y|x)p(x)\mathrm{d}x$$

could be written as

$$p(y) = \mathsf{E}p(y|\widetilde{x})$$

but we must be careful that it is an expectation over values of $\widetilde{x}$ (i.e. $\mathsf{E}_{\widetilde{x}}$) that occurs here.

Very occasionally we make use of results like

$$\begin{aligned}\mathsf{E}\widetilde{y} &= \mathsf{E}_{\widetilde{x}}\{\mathsf{E}_{\widetilde{y}|\widetilde{x}}(\widetilde{y}|\widetilde{x})\},\\ \mathscr{V}\widetilde{y} &= \mathsf{E}_{\widetilde{x}}\mathscr{V}_{\widetilde{y}|\widetilde{x}}(\widetilde{y}|\widetilde{x}) + \mathscr{V}_{\widetilde{x}}\mathsf{E}_{\widetilde{y}|\widetilde{x}}(\widetilde{y}|\widetilde{x}).\end{aligned}$$

The proofs are possibly more confusing than helpful. They run as follows:

$$\begin{aligned}\mathsf{E}\{\mathsf{E}(\widetilde{y}|\widetilde{x})\} &= \int \mathsf{E}(\widetilde{y}|x)p(x)\mathrm{d}x\\ &= \int \left(\int yp(y|x)\,\mathrm{d}y\right)p(x)\mathrm{d}x\\ &= \iint yp(x,\,y)\,\mathrm{d}y\mathrm{d}x\\ &= \int yp(y)\,\mathrm{d}y\\ &= \mathsf{E}\widetilde{y}.\end{aligned}$$

Similarly, we get the generalization

$$\mathsf{E}\{\mathsf{E}(g(\widetilde{y})|\widetilde{x})\} = \mathsf{E}g(\widetilde{y})$$

and in particular

$$\mathsf{E}\{\mathsf{E}(\widetilde{y}^2|\widetilde{x})\} = \mathsf{E}\widetilde{y}^2,$$

hence

$$\begin{aligned}\mathsf{E}\mathscr{V}(\widetilde{y}|\widetilde{x}) &= \mathsf{E}\{\mathsf{E}(\widetilde{y}^2|\widetilde{x})\} - \mathsf{E}\{\mathsf{E}(\widetilde{y}|\widetilde{x})\}^2\\ &= \mathsf{E}\widetilde{y}^2 - \mathsf{E}\{\mathsf{E}(\widetilde{y}|\widetilde{x})\}^2\end{aligned}$$

while

$$\begin{aligned}\mathscr{V}\mathsf{E}(\widetilde{y}|\widetilde{x}) &= \mathsf{E}\{\mathsf{E}(\widetilde{y}|\widetilde{x})\}^2 - [\mathsf{E}\{\mathsf{E}(\widetilde{y}\widetilde{|}x)\}]^2 \\ &= \mathsf{E}\{\mathsf{E}(\widetilde{y}|\widetilde{x})\}^2 - \mathsf{E}(\widetilde{y})^2\end{aligned}$$

from which it follows that

$$\mathsf{E}\mathscr{V}(\widetilde{y}|\widetilde{x}) + \mathscr{V}\mathsf{E}(\widetilde{y}|\widetilde{x}) = \mathsf{E}\widetilde{y}^2 - (\mathsf{E}\widetilde{y})^2 = \mathscr{V}\widetilde{y}.$$

1.5.8 Medians and modes

The mean is not the only measure of the centre of a distribution. We also need to consider the *median* from time to time, which is defined as any value x_0 such that

$$\mathsf{P}(\widetilde{x} \leqslant x_0) \geqslant \tfrac{1}{2} \quad \text{and} \quad \mathsf{P}(\widetilde{x} \geqslant x_0) \geqslant \tfrac{1}{2}.$$

In the case of most continuous random variables there is a unique median such that

$$\mathsf{P}(\widetilde{x} \geqslant x_0) = \mathsf{P}(\widetilde{x} \leqslant x_0) = \tfrac{1}{2}.$$

We occasionally refer also to the *mode*, defined as that value at which the pdf is a maximum. One important use we shall have for the mode will be in methods for finding the median based on the approximation

$$\text{mean} - \text{mode} = 3\,(\text{mean} - \text{median}\,)$$

or equivalently

$$\text{median} = (2\,\text{mean} + \text{mode}\,)/3$$

(see the preliminary remarks in Appendix A).

1.6 Exercises on Chapter 1

1. A card came is played with 52 cards divided equally between four players, North, South, East and West, all arrangements being equally likely. Thirteen of the cards are referred to as trumps. If you know that North and South have ten trumps between them, what is the probability that all three remaining trumps are in the same hand? If it is known that the king of trumps is included among the other three, what is the probability that one player has the king and the other the remaining two trumps?

2. (a) Under what circumstances is an event A independent of itself?

   (b) By considering events concerned with independent tosses of a red die and a blue die, or otherwise. give examples of events A, B and C which are not independent, but nevertheless are such that every pair of them is independent.

   (c) By considering events concerned with three independent tosses of a coin and supposing that A and B both represent tossing a head on the first trial, give examples of events A, B and C which are such that $\mathsf{P}(ABC) = \mathsf{P}(A)\mathsf{P}(B)\mathsf{P}(C)$ although no pair of them is independent.

3. Whether certain mice are black or brown depends on a pair of genes, each of which is either B or b. If both members of the pair are alike, the mouse is said to be homozygous, and if they are different it is said to be heterozygous. The mouse is brown only it it is homozygous bb. The offspring of a pair of mice have two such genes, one from each parent, and if the parent is heterozygous, the inherited gene is equally likely to be B or b. Suppose that a black mouse results from a mating between two heterozygotes.

   (a) What are the probabilities that this mouse is homozygous and that it is heterozygous?

   Now suppose that this mouse is mated with a brown mouse, resulting in seven offspring, all of which turn out to be black.

   (b) Use Bayes' Theorem to find the probability that the black mouse was homozygous BB.

   (c) Recalculate the same probability by regarding the seven offspring as seven observations made sequentially, treating the posterior after each observation as the prior for the next (cf. Fisher, 1959, Section II.2).

4. The example on Bayes' Theorem in Section 1.2 concerning the biology of twins was based on the assumption that births of boys and girls occur equally frequently, and yet it has been known for a very long time that fewer girls are born than boys (cf. Arbuthnot, 1710). Suppose that the probability of a girl is p, so that

$$\mathsf{P}(GG|M) = p,\ \mathsf{P}(BB|M) = 1 - p, \quad \mathsf{P}(GB|M) = 0,$$
$$\mathsf{P}(GG|D) = p^2,\ \mathsf{P}(BB|D) = (1 - p)^2,\ \mathsf{P}(GB|D) = 2p(1 - p).$$

   Find the proportion of monozygotic twins in the whole population of twins in terms of p and the sex distribution among all twins.

5. Suppose a red and a blue die are tossed. Let x be the sum of the number showing on the red die and twice the number showing on the blue die. Find the density function and the distribution function of x.

6. Suppose that $k \sim \mathrm{B}(n, \pi)$ where n is large and π is small but $n\pi = \lambda$ has an intermediate value. Use the exponential limit $(1 + x/n)^n \to \mathrm{e}^x$ to show that $\mathsf{P}(k = 0) \cong \mathrm{e}^{-\lambda}$ and $\mathsf{P}(k = 1) \cong \lambda\mathrm{e}^{-\lambda}$. Extend this result to show that k is such

that

$$p(k) \cong \frac{\lambda^k}{k!} \exp(-\lambda)$$

that is, k is approximately distributed as a Poisson variable of mean λ (cf. Appendix A).

7. Suppose that m and n have independent Poisson distributions of means λ and μ respectively (see question 6) and that $k = m + n$.
   (a) Show that $\mathsf{P}(k = 0) = \mathrm{e}^{-(\lambda+\mu)}$ and $\mathsf{P}(k = 1) = (\lambda + \mu)\mathrm{e}^{-(\lambda+\mu)}$.
   (b) Generalize by showing that k has a Poisson distribution of mean $\lambda + \mu$.
   (c) Show that conditional on k, the distribution of m is binomial of index k and parameter $\lambda/(\lambda + \mu)$.
8. Modify the formula for the density of a one-to-one function $g(x)$ of a random variable x to find an expression for the density of x^2 in terms of that of x, in both the continuous and discrete case. Hence, show that the square of a standard normal density has a chi-squared density on one degree of freedom as defined in Appendix A.
9. Suppose that $x_1, x_2, \ldots, x_n$ are independently and all have the same continuous distribution, with density $f(x)$ and distribution function $F(x)$. Find the distribution functions of

$$M = \max\{x_1, x_2, \ldots, x_n\} \quad \text{and} \quad m = \min\{x_1, x_2, \ldots, x_n\}$$

   in terms of $F(x)$, and so find expressions for the density functions of M and m.
10. Suppose that u and v are independently uniformly distributed on the interval [0, 1], so that the divide the interval into three sub-intervals. Find the joint density function of the lengths of the first two sub-intervals.
11. Show that two continuous random variables x and y are independent (i.e. $p(x, y) = p(x)p(y)$ for all x and y) if and only if their joint distribution function $F(x, y)$ satisfies $F(x, y) = F(x)F(y)$ for all x and y. Prove that the same thing is true for discrete random variables. [This is an example of a result which is easier to prove in the continuous case.]
12. Suppose that the random variable x has a negative binomial distribution $\mathrm{NB}(n, \pi)$ of index n and parameter π, so that

$$p(x) = \binom{n + x - 1}{x} \pi^n (1 - \pi)^x$$

    Find the mean and variance of x and check that your answer agrees with that given in Appendix A.

13. A random variable X is said to have a chi-squared distribution on ν degrees of freedom if it has the same distribution as

$$Z_1^2 + Z_2^2 + \cdots + Z_\nu^2$$

where $Z_1, Z_2, \ldots, Z_\nu$ are independent standard normal variates. Use the facts that $\mathsf{E}Z_i = 0$, $\mathsf{E}Z_i^2 = 1$ and $\mathsf{E}Z_i^4 = 3$ to find the mean and variance of X. Confirm these values using the probability density of X, which is

$$p(X) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} X^{\nu/2-1} \exp(-\tfrac{1}{2}X) \qquad (0 < X < \infty)$$

(see Appendix A).

14. The *skewness* of a random variable x is defined as $\gamma_1 = \mu_3/(\mu_2)^{\frac{3}{2}}$ where

$$\mu_n = \mathsf{E}(x - \mathsf{E}x)^n$$

(but note that some authors work in terms of $\beta_1 = \gamma_1^2$). Find the skewness of a random variable X with a binomial distribution $B(n, \pi)$ of index n and parameter π.

15. Suppose that a continuous random variable X has mean μ and variance ϕ. By writing

$$\phi = \int (x-\mu)^2 p(x)\mathrm{d}x \geqslant \int_{\{x;\,|x-\mu|\geqslant c\}} (x-\mu)^2 p(x)\mathrm{d}x$$

and using a lower bound for the integrand in the latter integral, prove that

$$\mathsf{P}(|x-\mu| \geqslant c) \leqslant \frac{\phi}{c^2}.$$

Show that the result also holds for discrete random variables. [This result is known as Čebyšev's Inequality (the name is spelt in many other ways, including Chebyshev and Tchebycheff).]

16. Suppose that x and y are such that

$$\begin{aligned}\mathsf{P}(x=0, y=1) = \mathsf{P}(x=0, y=-1) = \mathsf{P}(x=1, y=0)\\ = \mathsf{P}(x=-1, y=0) = \tfrac{1}{4}.\end{aligned}$$

Show that x and y are uncorrelated but that they are *not* independent.

17. Let x and y have a bivariate normal distribution and suppose that x and y both have mean 0 and variance 1, so that their marginal distributions are standard

normal and their joint density is

$$p(x, y) = \left\{2\pi\sqrt{(1-\rho^2)}\right\}^{-1} \exp\left\{-\tfrac{1}{2}(x^2 - 2\rho xy + y^2)/(1-\rho^2)\right\}.$$

Show that if the correlation coefficient between x and y is ρ, then that between x^2 and y^2 is ρ^2.

18. Suppose that x has a Poisson distribution (see question 6) P(λ) of mean λ and that, for given x, y has a binomial distribution B(x, π) of index x and parameter π.

    (a) Show that the unconditional distribution of y is Poisson of mean

    $$\lambda\pi = \mathsf{E}_{\widetilde{x}}\mathsf{E}_{\widetilde{y}|\widetilde{x}}(\widetilde{y}|\widetilde{x}).$$

    (b) Verify that the formula

    $$\mathscr{V}\,\widetilde{y} = \mathsf{E}_{\widetilde{x}}\mathscr{V}_{\widetilde{y}|\widetilde{x}}(\widetilde{y}|\widetilde{x}) + \mathscr{V}_{\widetilde{x}}\mathsf{E}_{\widetilde{y}|\widetilde{x}}(\widetilde{y}|\widetilde{x})$$

    derived in Section 1.5 holds in this case.

19. Define

$$I = \int_0^\infty \exp(-\tfrac{1}{2}z^2)\,\mathrm{d}z$$

and show (by setting $z = xy$ and then substituting z for y) that

$$I = \int_0^\infty \exp(-\tfrac{1}{2}(xy)^2)\,y\,dx = \int_0^\infty \exp(-\tfrac{1}{2}(zx)^2)\,z\,\mathrm{d}x.$$

Deduce that

$$I^2 = \int_0^\infty \int_0^\infty \exp\{-\tfrac{1}{2}(x^2+1)z^2\}\,z\,\mathrm{d}z\,\mathrm{d}x.$$

By substituting $(1+x^2)z^2 = 2t$, so that $z\,\mathrm{d}z = \mathrm{d}t/(1+x^2)$ show that $I = \sqrt{\pi/2}$, so that the density of the standard normal distribution as defined in Section 1.3 does integrate to unity and so is indeed a density. (This method is due to Laplace, 1812, Section 24.)