Notite de curs: Statistica

Eugenia Paulescu

Facultatea de Fizica, Universitatea de Vest din Timisoara

Octombrie-2023

Aceste notite sunt pentru cursul de Metode Statistice pentru anul II, master in fizica, predate in anul universitar 2023/2024, in semestrul I. Voi fi recunoscatoare pentru orice feedback din partea studentilor sau altor cititori critici.

Fiecare curs are asociat un set de probleme, pe care-l gasiti in directorul TEME.

Cuprins

Bibliografie:

1. K.F.Riley, M.P.Hobson, S.J.Bence, Mathematical Methods for Physics and Engineering, Third Edition (2006)
2. Sheldon M. Ross, Probability and Statistics for Engineers and Scientists, Fifth Edition (2014)
3. Michael J. Crawley, Statistics, Second Edition (2015)

In order to learn about something, you must first collect data.
Statistics is the art of learning from data.
It is concerned with the collection of data, its description, and its analysis, which often leads to the drawing of conclusions. Statistics is concerned with the analysis of real experimental data. First, we discuss probability. To a mathematician, probability

is an entirely theoretical subject based on axioms. This axiomatic approach is important, and we discuss it briefly.

We begin our discussion with the terminology required, with particular reference to the convenient graphical representation of experimental results as Venn diagrams. The concepts of random variables and distributions of random variables are then introduced. We assert that the results of many experiments are random variables and that those results have some sort of regularity, which is represented by a distribution. Precise definitions of a random variable and a distribution are then given, as are the defining equations for some important distributions. We also derive some useful quantities associated with these distributions: mean, variance, moments.

Chapter 1 Probability

Bibliografie: Riley et al. (2006)

1.1 Venn diagrams

We call a single performance of an experiment a *trial* and each possible result an *outcome*. The *sample space* S of the experiment is then the set of all possible outcomes of an individual trial. For example, if we throw a six-sided die then there are six possible outcomes that together form the sample space of the experiment. At this stage we are not concerned with how likely a particular outcome might be, but rather will concentrate on the classification of possible outcomes. Some sample spaces are *finite* (e.g. the outcomes of throwing a die) whilst others are *infinite* (e.g. the outcomes of measuring people's heights). Most often, one is not interested in individual outcomes but in whether an outcome belongs to a given subset A (say) of the sample space S; these subsets are called *events*. For example, we might be interested in whether a person is taller or shorter than 180 cm, in which case we divide the sample space into just two events: namely, that the outcome (height measured) is (i) greater than 180 cm or (ii) less than 180 cm.

A common graphical representation of the outcomes of an experiment is the *Venn diagram*. A Venn diagram usually consists of a rectangle, the *interior* of which represents the *sample space*, together with one or more closed curves inside it. The interior of each *closed curve* then represents an *event*. Figure 1.1 shows a typical Venn diagram representing a sample space *S* and two events A and B. Every possible outcome is assigned to an appropriate region; in this example there are four regions to consider, marked *i* to *iv* in figure 1.1.

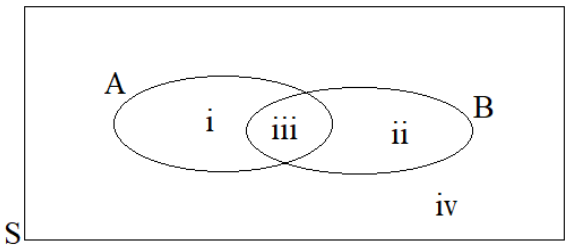

Figure 1.1

(i) outcomes that belong to event A but not to event B;

(ii) outcomes that belong to event B but not to event A;

(iii) outcomes that belong to both event A and event B;

(iv) outcomes that belong to neither event A nor event B.

Exercise 1. A six-sided die is thrown. Let event A be 'the number obtained is odd' and event B be 'the number obtained is divisible by 3'. Draw a Venn diagram to represent these events.

It is clear that the outcomes 1, 3, 5 belong to event A and that the outcomes 3, 6 belong to event B. Of these, 3 belongs to both A and B. The remaining outcomes, 2, 4, belong to neither A nor B. The appropriate Venn diagram is shown in figure 1.2.

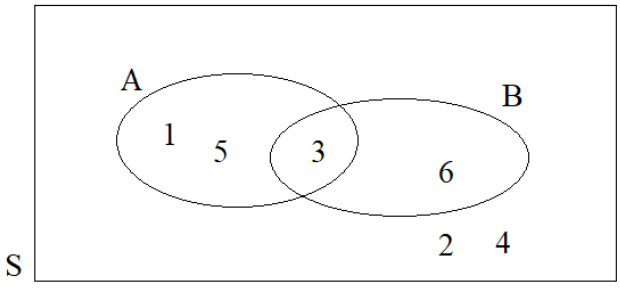

Figure 1.2

In the above example, one outcome, 3, is divisible by 3 and is odd, and so belongs to both A and B. This outcome is placed in region *iii* of figure 1.1, which is called the *intersection* of A and B and is denoted by $A \cap B$ (see figure 1.3(a)). If no events lie in the region of intersection then A and B are said to be *mutually exclusive* or disjoint. An event that contains no outcomes is called the *empty* event and denoted by $\emptyset$.

The event comprising all the elements that belong to either A or B, or to both, is called the *union* of A and B and is denoted by $A \cup B$ (see figure 1.3(b)). In the previous example, $A \cup B = \{1,3,5,6\}$.

It is sometimes convenient to talk about those outcomes that do *not belong* to a particular event. The set of outcomes that do not belong to A is called the *complement* of A and is denoted by $\bar{A}$ (see figure 1.3(c)); this can also be written as $\bar{A} = S - A$. It is clear that $A \cup \bar{A} = S$ and $A \cap \bar{A} = \varnothing$. The above notation can be extended in an obvious way, so that $A - B$ denotes the outcomes in A that do not belong to B. It is clear from figure 1.3(d) that $A - B$ can also be written as $A \cap \bar{B}$.

Finally, when all the outcomes in event B (say) also belong to event A, but A may contain, in addition, outcomes that do not belong to B, then B is called a *subset* of A, a situation that is denoted by $B \subset A$. In this case, the closed curve representing the event B is often drawn lying completely within the closed curve representing the event A.

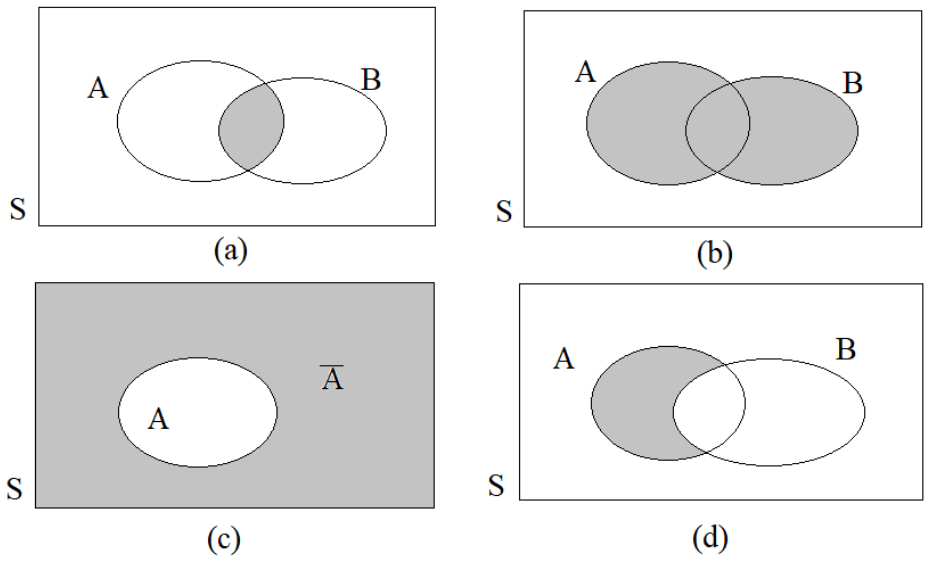

Figure 1.3

The operations $\cup$ and $\cap$ are extended straightforwardly to more than two events. If there exist *n* events $A_1, A_2, \ldots, A_n$ in some sample space *S*, then the event consisting of all those outcomes that belong to one or more of the A_i is the union of $A_1, A_2, \ldots, A_n$ and is denoted by

$$A_1 \cup A_2 \cup \ldots \cup A_n \tag{1.1}$$

Similarly, the event consisting of all the outcomes that belong to every one of the A_i is called the intersection of $A_1, A_2, \ldots, A_n$ and is denoted by

$$A_1 \cap A_2 \cap \ldots \cap A_n \tag{1.2}$$

If, for any pair of values i, j with $i \neq j$

$$A_i \cap A_j = \varnothing \tag{1.3}$$

then the events A_i are said to be *mutually exclusive* or *disjoint*.

Consider three events A, B and C with a Venn diagram such as is shown in figure 1.4. It will be clear that, in general, the diagram will be divided into eight regions and they will be of four different types. Three regions correspond to a single event; three regions are each the intersection of exactly two events; one region is the three-fold intersection of all three events; and finally one region corresponds to none of the events.

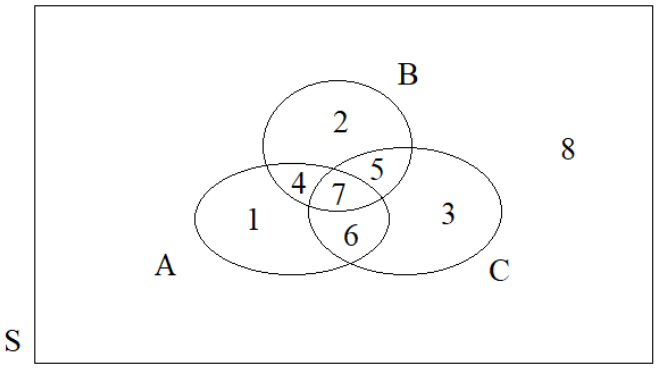

Figure 1.4

Let us now consider the numbers of different regions in a general *n*-event Venn diagram.

For one-event Venn diagrams there are two regions, for the two-event case there are four regions and, as we have just seen, for the three-event case there are eight. In the general *n*-event case there are 2^n regions.

The 2^n regions will break down into $n+1$ types, with the numbers of each type as follows

no events, $C_n^0 = 1$

one event but no intersections, $C_n^1 = n$

two-fold intersections, $C_n^2 = \frac{n(n-1)}{2}$

three-fold intersections, $C_n^3 = \frac{n(n-1)(n-2)}{3!}$

… an *n*-fold intersection, $C_n^n = 1$

That this makes a total of 2^n can be checked by considering the binomial expansion

$$2^n = (1+1)^n = 1 + n + \frac{n(n-1)}{2} + \frac{n(n-1)(n-2)}{3!} + \ldots + 1$$

The operations $\cap$ and $\cup$ obey the following algebraic laws:

commutativity, $A \cap B = B \cap A$, $A \cup B = B \cup A$

associativity, $(A \cap B) \cap C = A \cap (B \cap C)$, $(A \cup B) \cup C = A \cup (B \cup C)$

distributivity, $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$,

$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

idempotency, $A \cap A = A$, $A \cup A = A$

From Venn diagrams, it is simple to show that the following rules hold:

$$\begin{aligned} &\text{(i)} \quad \text{If } A \subset B \text{ then } \bar{A} \supset \bar{B} \\ &\text{(ii)} \quad \overline{A \cup B} = \bar{A} \cap \bar{B} \\ &\text{(iii)} \quad \overline{A \cap B} = \bar{A} \cup \bar{B} \end{aligned} \tag{1.4}$$

Statements (ii) and (iii) are known jointly as *de Morgan's laws*.

Exercise 2. There exist two events A and B such that

$$\overline{(X \cup A)} \cup \overline{(X \cup \bar{A})} = B$$

Find an expression for the event X in terms of A and B.

We begin by taking the complement of both sides of the above expression: applying de Morgan's laws we obtain

$$\bar{B} = (X \cup A) \cap (X \cup \bar{A})$$

We may then use the algebraic laws obeyed by $\cap$ and $\cup$ to yield

$$\bar{B} = X \cup (A \cap \bar{A}) = X \cup \varnothing = X$$

Thus, we find $X = \bar{B}$.

1.2 Probability

Venn diagrams are graphical representations of the possible outcomes of experiments. Moreover, how likely each outcome or event might be in any particular experiment? Most experiments show some regularity: the relative frequency of an event is approximately the same on each occasion that a set of trials is performed. For example, if we throw a die N times then we expect that a six will occur approximately N/6 times. The regularity of outcomes allows us to define the *probability*, Pr(A), as the expected relative frequency of event A in a large number of trials. If an experiment has a total of n_s outcomes in the sample space S, and n_A of these outcomes correspond to the event A, then the probability that event A will occur is

$$\Pr(A) = \frac{n_A}{n_S} \tag{1.5}$$

1.2.1 Axioms and theorems

From (1.5) we may deduce the following properties of the probability $\Pr(A)$.

(i) For any event *A* in a sample space *S*,

$$0 \le \Pr(A) \le 1 \tag{1.6}$$

If $\Pr(A) = 1$ then *A* is a *certainty*; if $\Pr(A) = 0$ then *A* is an *impossibility*.

(ii) For the entire sample space *S* we have

$$\Pr(S) = \frac{n_S}{n_S} = 1 \tag{1.7}$$

which simply states that we are certain to obtain one of the possible outcomes.

(iii) If *A* and *B* are two events in *S* then, from the Venn diagrams in figure 1.3, we see that

$$n_{A\cup B} = n_A + n_B - n_{A\cap B} \tag{1.8}$$

the final subtraction arising because the outcomes in the intersection of *A* and *B* are counted twice when the outcomes of *A* are added to those of *B*. Dividing both sides of (1.8) by n_s, we obtain the *addition rule* for probabilities

$$\Pr(A\cup B) = \Pr(A) + \Pr(B) - \Pr(A\cap B) \tag{1.9}$$

If *A* and *B* are mutually exclusive events $A\cap B = \varnothing$, then $\Pr(A\cap B) = 0$ and we obtain the special case

$$\Pr(A\cup B) = \Pr(A) + \Pr(B) \tag{1.10}$$

(iv) If $\bar{A}$ is the complement of *A* then $\bar{A}$ and *A* are mutually exclusive events. Thus, from (1.7) and (1.10) we have

$$1 = \Pr(S) = \Pr\left(A\cup\bar{A}\right) = \Pr(A) + \Pr\left(\bar{A}\right)$$

from which we obtain the *complement law*

$$\Pr\left(\bar{A}\right) = 1 - \Pr(A) \tag{1.11}$$

Exercise 3. Calculate the probability of drawing an *ace* or a *spade* from a pack of cards.

Let *A* be the event that an ace is drawn and *B* the event that a spade is drawn.

It follows that $\Pr(A)=\frac{4}{52}=\frac{1}{13}$ and $\Pr(B)=\frac{13}{52}=\frac{1}{4}$. The intersection of A and B consists of only the ace of spades and so $\Pr(A\cap B)=\frac{1}{52}$.

$$\Pr(A\cup B)=\Pr(A)+\Pr(B)-\Pr(A\cap B)=\frac{1}{13}+\frac{1}{4}-\frac{1}{52}=\frac{4}{13}$$

In this case it is just as simple to recognise that there are 16 cards in the pack that satisfy the required condition (13 spades plus three other aces) and so the probability is 16/52.

The above theorems can easily be extended to a greater number of events. For example, if $A_1, A_2, \ldots, A_n$ are mutually exclusive events then (1.10) becomes

$$\Pr(A_1\cup A_2\cup\ldots\cup A_n)=\Pr(A_1)+\Pr(A_2)+\ldots+\Pr(A_n) \qquad (1.12)$$

Exercise 4. A biased six-sided die has probabilities $\frac{1}{2}p$, $\frac{1}{3}p$, $\frac{1}{3}p$, $\frac{1}{3}p$, p, $2p$ of showing 1, 2, 3, 4, 5, 6 respectively. Calculate p.

Given that the individual events are mutually exclusive, (1.12) can be applied to give

$$\Pr(1\cup 2\cup 3\cup 4\cup 5\cup 6)=\frac{1}{2}p+\frac{1}{3}p+\frac{1}{3}p+\frac{1}{3}p+p+2p=\frac{9}{2}p$$

The union of all possible outcomes is clearly the sample space, S, and so

$$\Pr(S)=\frac{9}{2}p \quad \overset{(1.7)}{\Rightarrow} \quad 1=\frac{9}{2}p \quad \Rightarrow \quad p=\frac{2}{9}$$

When the possible outcomes of a trial correspond to more than two events, and those events are not mutually exclusive, the calculation of the probability of the union of a number of events is more complicated, and the generalisation of the *addition law* (1.9) requires further work. Let us begin by considering the union of three events A_1, A_2 and A_3, which need not be mutually exclusive. We first define the event $B=A_2\cup A_3$ and, using the addition law (1.9), we obtain

$$\Pr\left(A_1 \cup A_2 \cup A_3\right) = \Pr\left(A_1 \cup B\right) = \Pr\left(A_1\right) + \Pr\left(B\right) - \Pr\left(A_1 \cap B\right) \qquad (1.13)$$

$$\Pr\left(A_1 \cap B\right) = \Pr\left(A_1 \cap \left(A_2 \cup A_3\right)\right) =$$

$$= \Pr\left(\left(A_1 \cap A_2\right) \cup \left(A_1 \cap A_3\right)\right) =$$

$$= \Pr\left(A_1 \cap A_2\right) + \Pr\left(A_1 \cap A_3\right) - \Pr\left(A_1 \cap A_2 \cap A_3\right) \qquad (1.14)$$

Substituting this expression (1.14), and that for $\Pr\left(B\right)$ obtained from (1.9), into (1.13) we obtain the probability addition law for three general events,

$$\Pr\left(A_1 \cup A_2 \cup A_3\right) = \Pr\left(A_1\right) + \Pr\left(A_2\right) + \Pr\left(A_3\right) - \Pr\left(A_2 \cap A_3\right) -$$

$$- \Pr\left(A_1 \cap A_2\right) - \Pr\left(A_1 \cap A_3\right) + \Pr\left(A_1 \cap A_2 \cap A_3\right) \qquad (1.15)$$

Exercise 5. Calculate the probability of drawing from a pack of cards one that is an ace or is a spade or shows an even number (2, 4, 6, 8, 10).

If *A* is the event that an ace is drawn, $\Pr\left(A\right) = \frac{4}{52}$. Similarly the event *B*, that a spade is drawn, has $\Pr\left(B\right) = \frac{13}{52}$. The possibility *C*, that the card is even (but not a picture card) has $\Pr\left(C\right) = \frac{20}{52}$. The two-fold intersections have probabilities

$$\Pr\left(A \cap B\right) = \frac{1}{52} \ , \quad \Pr\left(A \cap C\right) = 0 \ , \quad \Pr\left(B \cap C\right) = \frac{5}{52}$$

There is no three-fold intersection as events A and C are mutually exclusive. Hence

$$\Pr\left(A \cup B \cup C\right) = \frac{1}{52}\left[\left(4 + 13 + 20\right) - \left(1 + 0 + 5\right) + 0\right] = \frac{31}{52}$$

The probability for the union of the *n* general events, which may be proved by induction upon *n*, is

$$\Pr\left(A_1 \cup A_2 \cup \ldots A_n\right) = \sum_i \Pr\left(A_i\right) - \sum_{i,j} \Pr\left(A_i \cap A_j\right) +$$

$$+ \sum_{i,j,k} \Pr\left(A_i \cap A_j \cap A_k\right) - \ldots \left(-1\right)^{n+1} \Pr\left(A_1 \cap A_2 \cap \ldots \cap A_n\right) \qquad (1.16)$$

Each summation runs over all possible sets of subscripts, except those in which any two subscripts in a set are the same. The number of terms in the summation of probabilities of *m*-fold intersections of the *n* events is given by C_n^m (as discussed in section 1.1). We now illustrate this result with a worked example that has $n=4$ and includes a four-fold intersection.

Exercise 6. Find the probability of drawing from a pack a card that has at least one of the following properties:

A, it is an ace;

B, it is a spade;

C, it is a black honour card (ace, king, queen, jack or 10);

D, it is a black ace.

Measuring all probabilities in units of 1/52, the single-event probabilities are

$$\Pr(A)=4\ ,\quad \Pr(B)=13\ ,\quad \Pr(C)=10\ ,\quad \Pr(D)=2$$

The two-fold intersection probabilities, measured in the same units, are

$$\Pr(A\cap B)=1\ ,\quad \Pr(A\cap C)=2\ ,\quad \Pr(A\cap D)=2$$

$$\Pr(B\cap C)=5,\quad \Pr(B\cap D)=1,\quad \Pr(C\cap D)=2$$

The three-fold intersections have probabilities

$$\Pr(A\cap B\cap C)=1\ ,\ \Pr(A\cap B\cap D)=1,\ \Pr(A\cap C\cap D)=2,\ \Pr(B\cap C\cap D)=1$$

Finally, the four-fold intersection, requiring all four conditions to hold, is satisfied only by the ace of spades, and hence (again in units of 1/52)

$$\Pr(A\cap B\cap C\cap D)=1$$

Substituting in (1.16) gives

$$P=\frac{1}{52}\left[(4+13+10+2)-(1+2+2+5+1+2)+(1+1+2+1)-(1)\right]=\frac{20}{52}$$

We conclude this section on basic theorems by deriving a useful general expression for the probability $\Pr(A \cap B)$ that two events *A* and *B* both occur in the case where *A* (say) is the union of a set of *n* mutually exclusive events A_i. In this case

$$A \cap B = (A_1 \cap B) \cup (A_2 \cap B) \cup \ldots \cup (A_n \cap B)$$

where the events $A_i \cap B$ are also mutually exclusive. Thus, from the *addition law* (1.12) for mutually exclusive events, we find

$$\Pr(A \cap B) = \sum_i \Pr(A_i \cap B) \tag{1.17}$$

Moreover, in the special case where the events A_i exhaust the sample space *S*, we have $A \cap B = S \cap B = B$, and we obtain the *total probability law*

$$\Pr(B) = \sum_i \Pr(A_i \cap B) \tag{1.18}$$

1.2.2 Conditional probability

So far we have defined only probabilities of the form 'what is the probability that event *A* happens?'. In this section we turn to *conditional probability*, the probability that a particular event occurs given the occurrence of another, possibly related, event. For example, we may wish to know the probability of event *B*, drawing an ace from a pack of cards from which one has already been removed, given that event *A*, the card already removed was itself an ace, has occurred.

We denote this probability by $\Pr(B|A)$ and may obtain a formula for it by considering the probability $\Pr(A \cap B) = \Pr(B \cap A)$ that both *A* and *B* will occur. This may be written in two ways, i.e.

$$\Pr(A \cap B) = \Pr(A)\Pr(B|A)$$

$$= \Pr(B)\Pr(A|B)$$

From this we obtain

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)} \tag{1.19}$$

$$\Pr(B|A) = \frac{\Pr(B \cap A)}{\Pr(A)} \qquad (1.20)$$

In terms of Venn diagrams, we may think of $\Pr(B|A)$ as the probability of *B* in the *reduced sample space* defined by *A*. Thus, if two events *A* and *B* are mutually exclusive then

$$\Pr(A|B) = 0 = \Pr(B|A) \qquad (1.21)$$

When an experiment consists of drawing objects at random from a given set of objects, it is termed *sampling a population*. We need to distinguish between two different ways in which such a *sampling experiment* may be performed. After an object has been drawn at random from the set it may either be put aside or returned to the set before the next object is randomly drawn. The former is termed '*sampling without replacement*', the latter '*sampling with replacement*'.

Exercise 7. Find the probability of drawing two aces at random from a pack of cards (i) when the first card drawn is replaced at random into the pack before the second card is drawn, and (ii) when the first card is put aside after being drawn.

Let *A* be the event that the first card is an ace, and *B* the event that the second card is an ace. Now

$$\Pr(A \cap B) = \Pr(A)\Pr(B|A)$$

and for both (i) and (ii) we know that $\Pr(A) = \frac{4}{52} = \frac{1}{13}$.

(i) If the first card is replaced in the pack before the next is drawn then $\Pr(B|A) = \Pr(B) = \frac{4}{52} = \frac{1}{13}$ since *A* and *B* are independent events. We then have

$$\Pr(A \cap B) = \Pr(A)\Pr(B) = \frac{1}{13} \cdot \frac{1}{13} = \frac{1}{169}$$

(ii) If the first card is put aside and the second then drawn, *A* and *B* are not independent and $\Pr(B|A) = \frac{3}{51}$, with the result that

$$\Pr(A\cap B)=\Pr(A)\Pr(B|A)=\frac{1}{13}\cdot\frac{3}{51}=\frac{1}{221}$$

Two events A and B are *statistically independent* if $\Pr(A|B)=\Pr(A)$ (or equivalently if $\Pr(B|A)=\Pr(B)$) In words, the probability of A given B is then the same as the probability of A regardless of whether B occurs. For example, if we throw a coin and a die at the same time, we would normally expect that the probability of throwing a six was independent of whether a head was thrown. If A and B are statistically independent then

$$\Pr(A\cap B)=\Pr(A)\Pr(B) \qquad (1.22)$$

In fact (1.22) may be regarded as the definition of the *statistical independence* of two events.

We now derive two results that often prove useful when working with conditional probabilities. Let us suppose that an event A is the union of n mutually exclusive events A_i. If B is some other event then from (1.17) we have

$$\Pr(A\cap B)=\sum_i \Pr(A_i\cap B)$$

Dividing both sides of this equation by $\Pr(B)$ and using (1.19), we obtain

$$\Pr(A|B)=\sum_i \Pr(A_i|B) \qquad (1.23)$$

which is the *addition law* for *conditional probabilities*.

Furthermore, if the set of mutually exclusive events A_i exhausts the sample space S then, from the total probability law (1.18), the probability $\Pr(B)$ of some event B in S can be written as

$$\Pr(B)=\sum_i \Pr(A_i\cap B)=\sum_i \Pr(A_i)\Pr(B|A_i) \qquad (1.24)$$

Exercise 10. A collection of traffic islands connected by a system of one-way roads is shown in figure 1.5. At any given island a car driver chooses a direction at random

from those available. What is the probability that a driver starting at *O* will arrive at *B*?

In order to leave *O* the driver must pass through one of A_1, A_2, A_3 or A_4, which thus form a complete set of mutually exclusive events. Since at each island (including *O*) the driver chooses a direction at random from those available, we have that $\Pr(A_i)=\frac{1}{4}$, for $i=1,2,3,4$ From figure 1.5, we see also that

$$\Pr(B|A_1)=\frac{1}{3}\ ,\quad \Pr(B|A_2)=\frac{1}{3}\ ,\quad \Pr(B|A_3)=0\ ,\quad \Pr(B|A_4)=\frac{2}{4}=\frac{1}{2}$$

Thus, using the total probability law (1.24), we find that the probability of arriving at *B* is given by

$$\Pr(B)=\sum_i \Pr(A_i)\Pr(B|A_i)=\frac{1}{4}\left(\frac{1}{3}+\frac{1}{3}+0+\frac{1}{2}\right)=\frac{7}{24}$$

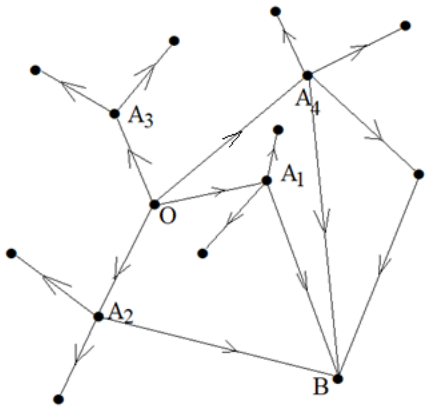

Figure 1.5 A collection of traffic islands connected by one-way roads.

1.2.3 Bayes' theorem

In the previous section we saw that the probability that both an event *A* and a related event *B* will occur can be written either as $\Pr(A)\Pr(B|A)$ or $\Pr(B)\Pr(A|B)$. Hence

$$\Pr(A)\Pr(B|A)=\Pr(B)\Pr(A|B)$$

from which we obtain *Bayes' theorem*,

$$\Pr(A|B) = \frac{\Pr(A)}{\Pr(B)} \Pr(B|A) \qquad (1.25)$$

This theorem clearly shows that $\Pr(B|A) \neq \Pr(A|B)$, unless $\Pr(A) = \Pr(B)$. It is sometimes useful to rewrite $\Pr(B)$, if it is not known directly, as

$$\Pr(B) = \Pr(A)\Pr(B|A) + \Pr(\bar{A})\Pr(B|\bar{A})$$

so that Bayes' theorem becomes

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(A)\Pr(B|A) + \Pr(\bar{A})\Pr(B|\bar{A})} \qquad (1.26)$$

Exercise 9: Suppose that the blood test for some disease is reliable in the following sense: for people who are infected with the disease the test produces a positive result in 99.99% of cases; for people not infected a positive test result is obtained in only 0.02% of cases. Furthermore, assume that in the general population one person in 10000 people is infected. A person is selected at random and found to test positive for the disease. What is the probability that the individual is actually infected?

Let *A* be the event that the individual is infected and *B* be the event that the individual tests positive for the disease. Using Bayes' theorem the probability that a person who tests positive is actually infected is

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\Pr(A)\Pr(B|A) + \Pr(\bar{A})\Pr(B|\bar{A})}$$

Now, $\Pr(A) = \frac{1}{10000} = 1 - \Pr(\bar{A})$

and we are told that $\Pr(B|A) = \frac{9999}{10000}$ and $\Pr(B|\bar{A}) = \frac{2}{10000}$. Thus we obtain

$$\Pr(A|B) = \frac{\frac{1}{10000} \cdot \frac{9999}{10000}}{\frac{1}{10000} \cdot \frac{9999}{10000} + \frac{9999}{10000} \cdot \frac{2}{10000}} = \frac{1}{3}$$

Thus, there is only a one in three chance that a person chosen at random, who tests positive for the disease, is actually infected.

At a first glance, this answer may seem a little surprising, but the reason for the counterintuitive result is that the probability that a randomly selected person is not infected is 9999/10000, which is very high. Thus, the 0.02% chance of a positive test for an uninfected person becomes significant.

We note that (1.26) may be written in a more general form if S is not simply divided into A and $\bar{A}$ but, rather, into any set of mutually exclusive events A_i that exhaust S. Using the total probability law (1.24), we may then write

$$\Pr(B) = \sum_i \Pr(A_i)\Pr(B|A_i)$$

so that Bayes' theorem takes the form

$$\Pr(A|B) = \frac{\Pr(A)\Pr(B|A)}{\sum_i \Pr(A_i)\Pr(B|A_i)} \qquad (1.27)$$

where the event A need not coincide with any of the A_i.

Exercise 10: An insurance company believes that people can be divided into two classes — those that are accident prone and those that are not. Their statistics show that an accident-prone person will have an accident at some time within a fixed 1-year period with probability 0.4, whereas this probability decreases to 0.2 for a non-accident-prone person. If we assume that 30 percent of the population is accident prone, what is the probability that a new policy holder will have an accident within a year of purchasing a policy?

We obtain the desired probability by first conditioning on whether or not the policy holder is accident prone. Let A_1 denote the event that the policy holder will have an accident within a year of purchase; and let A denote the event that the policy holder is accident prone. Hence, the desired probability, $P(A_1)$, is given by

$$\begin{aligned} P(A_1) &= P(A)P(A_1|A) + P(\bar{A})P(A_1|\bar{A}) \\ &= 0.3\times 0.4 + 0.7\times 0.2 = 0.26 \end{aligned}$$

Exercise 11: Reconsider the example with the insurance company and suppose that a new policy holder has an accident within a year of purchasing his policy. What is the probability that he is accident prone?

Initially, at the moment when the policy holder purchased his policy, we assumed there was a 30 percent chance that he was accident prone. That is, $P(A) = 0.3$. However, based on the fact that he has had an accident within a year, we now reevaluate his probability of being accident prone as follows:

$$P(A|A_1) = \frac{P(A \cap A_1)}{P(A_1)} = \frac{P(A)P(A_1|A)}{P(A_1)} = \frac{0.3 \times 0.4}{0.26} = 0.4615$$

Exercise 12: Twins can either be identical or fraternal. Identical, also called *monozygotic*, twins form when a single fertilized egg splits into two genetically identical parts. Consequently, identical twins always have the same set of genes. Fraternal, also called *dizygotic*, twins develop when two separate eggs are fertilized and implant in the uterus. The genetic connection of fraternal twins is no more or less the same as siblings born at separate times. A Los Angeles county scientist wishing to know the current fraction of twin pairs born in the county that are identical twins has assigned a county statistician to study this issue. The statistician initially requested each hospital in the county to record all twin births, indicating whether the resulting twins were identical or not. The hospitals, however, told her that to determine whether newborn twins were identical was not a simple task, as it involved the permission of the twins's parents to perform complicated and expensive DNA studies that the hospitals could not afford. After some deliberation, the statistician just asked the hospitals for data listing all twin births along with an indication as to whether the twins were of the same sex. When such data indicated that approximately 64 percent of twin births were same-sexed, the statistician declared that approximately 28 percent of all twins were identical. How did she come to this conclusion?

The statistician reasoned that identical twins are always of the same sex, whereas fraternal twins, having the same relationship to each other as any pair of siblings, will have probability ½ of being of the same sex. Letting *I* be the event that a pair of twins are identical, and *SS* be the event that a pair of twins are of the same sex, she computed the probability $P(SS)$ by conditioning on whether the twin pair was identical. This gave

$$P(SS) = P(SS|I)P(I) + P(SS|\bar{I})P(\bar{I})$$

$$P(SS) = 1 \times P(I) + \frac{1}{2} \times (1 - P(I)) = \frac{1}{2} + \frac{1}{2}P(I)$$

which, using that $P(SS) \approx 0.64$ yielded the result $P(I) \approx 0.28$