Concepts in Modern Mathematics, Ian Stewart

Chapter 17 The Theory of Probability

'Statistics is a branch of theology' –
A Cambridge research fellow

Probability theory has its origins in questions about gambling. In a game of cards, or dice, when do I have the best chance of winning? What are the odds?[1]

Because games are usually finite, the methods needed to handle such questions are *combinatorial*, that is, based on counting arguments. For example, to find the chance of throwing three consecutive heads with a coin one lists the possibilities

$$\begin{array}{cccc} HHH & HHT & HTH & HTT \\ THH & THT & TTH & TTT \end{array}$$

which are 8 in number. Exactly 1 is favourable, so the probability is 1/8.

This of course makes the assumption that throws of H or T are equally likely. Now we can't define 'equally likely' by saying 'probability 1/2' until we have defined what we mean by 'probability 1/2': and we can't do that without defining 'equally likely'. Or at least, so it seems.

If we try to get round it by doing experiments, we run into another difficulty. If H and T are equally likely, then in a long series of throws we would expect to have approximately equal numbers of H and T. Not exactly equal, of course: they couldn't possibly be equal in an odd number of throws anyway, and in an even number of throws there would probably be a small discrepancy. Toss a coin 20 times and see if you get exactly 10 heads. (If you do, try several more times and see how often it happens!)

What we would hope is that 'in the limit' the ratio of the number of Hs to the number of Ts should 'tend to' 1/2. The trouble is that this 'limit' is not a limit in the usual sense of analysis. It is conceivable that we might throw a sequence consisting entirely of Hs with a *fair* coin. It is, of course, unlikely. But to set up an idea of 'limit' which takes account of this possibility involves making precise what we mean by 'unlikely', which seems to require a definition of 'probability' again!

It wasn't until the 1930s that these difficulties were circumvented. This was achieved by developing an *axiomatic* probability theory. By divorcing the mathematics from its applications one can develop the mathematics without any logical qualms: *then* it can be tested experimentally to see if it fits the facts. Axiomatic probability theory succeeds for the same reason that axiomatic geometry succeeds.

Combinatorial Probability

For the moment assume that we know what 'equally likely' means. Then a rough working definition of the *probability* $p(E)$ of an *event* E is

$$p(E) = \frac{\text{number of ways in which } E \text{ can occur}}{\text{total number of possible occurrences}}$$

(provided all occurrences are equally likely).

Thus there are 36 ways of throwing 2 dice; and 5 of these give a total of 6 (namely $1+5, 2+4, 3+3, 4+2, 5+1$).

Therefore the probability that the total is 6 is

$$\frac{\text{number of ways of throwing 6}}{36}$$

which is 5/36.

Since the numbers involved are positive, and since the number of ways E can occur is at most equal to the total number of occurrences, we see that

$$0 \leq p(E) \leq 1.$$

If $p(E) = 0$ then E is impossible; if $p(E) = 1$ then E is certain.

The techniques of combinatorial probability centre around ways of combining events. Suppose we have two distinct events E and F. What is the probability that *either* E or F occurs?

Take the case of a die. E is the event '6 is thrown' and F the event '5 is thrown'. E or F is '5 or 6 is thrown' which obviously occurs 2 times out of 6. So

$$p(E \text{ or } F) = 1/3.$$

In general, let $N(E)$ and $N(F)$ be the number of ways in which E and F can occur, and T the total number of occurrences. Then

$$p(E \text{ or } F) = N(E \text{ or } F)/T.$$

What is $N(E \text{ or } F)$? Suppose the events E and F do not 'overlap'. (I'll return to this point.) Then

$$N(E \text{ or } F) = N(E)+N(F)$$

so that

$$\begin{aligned} p(E \text{ or } F) &= (N(E)+N(F))/T \\ &= (N(E))/T+(N(F))/T \\ &= p(E)+p(F). \end{aligned} \tag{1}$$

If, however, E and F do overlap then $N(E)+N(F)$ counts everything in the overlap *twice* whereas $N(E \text{ or } F)$ only counts it *once*.

Suppose, for instance, that

E = 'A prime number is thrown'
F = 'An odd number is thrown'.

Then E occurs in three ways: 2, 3, 5. (Note: 1 is not prime.) And F occurs in three ways: 1, 3, 5. But E or F occurs in *four* ways: 1, 2, 3, 5. So

$$p(E) = 1/2 \qquad p(F) = 1/2 \qquad p(E \text{ or } F) = 2/3.$$

What happens in general is that

$$N(E \text{ or } F) = N(E)+N(F)-N(E \text{ and } F) \tag{2}$$

because subtracting $N(E \text{ and } F)$ puts right the double count in the overlap. In the above example, E and F occurs in two ways: 3, 5. So the equation gives

$$4 = 3+3-2$$

which is correct.

Dividing (2) by T we get

$$p(E \text{ or } F) = p(E)+p(F)-p(E \text{ and } F). \tag{3}$$

Enter Set Theory

We can express these ideas much better in terms of sets. The possible *outcomes* when throwing a die form a set

$$X = \{1, 2, 3, 4, 5, 6\}.$$

The events E and F are represented by *subsets* of X

$$E = \{2, 3, 5\}$$
$$F = \{1, 3, 5\}$$

as in Figure 172.

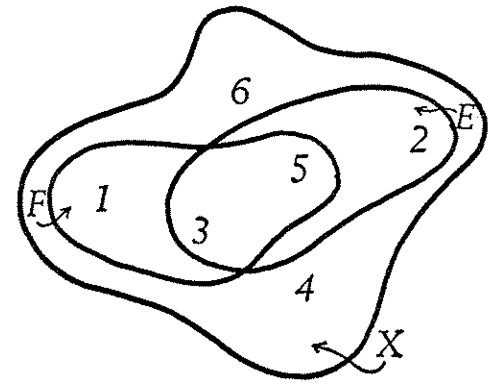

Figure 172

The event 'E or F' is the set $\{1, 2, 3, 5\}$ which is the *union* $E \cup F$. The event 'E and F' is the set $\{3, 5\}$ which is the intersection $E \cap F$. The probability p is a function defined on the set $\mathscr{E}$ of all subsets of X with target $\mathbf{R}$. In general we can say a little more about p: it has target $[0, 1]$, where this denotes the set of real numbers between 0 and 1.

Abstracting from this we obtain the idea of a *finite probability space*. This comprises

(i) a finite set X,
(ii) the set $\mathscr{E}$ of all subsets of X
(iii) a function $p: \mathscr{E} \to [0, 1]$ with the property that

$$p(E \cup F) = p(E)+p(F)-p(E \cap F)$$

for all $E, F \in \mathscr{E}$.

Axiomatic probability theory works entirely in terms of probability spaces. However, if one wishes to consider infinite probability spaces, the definition has to be made more subtly. In many applications it is necessary to have infinite sets X: for example the height of a man can be any real number (within certain limits) so there are infinitely many possibilities.

Independence

Another basic operation in probability theory deals with two trials in succession: what is the probability of event E occurring on the first trial, event F on the second? For example, we throw a die twice: what are the chances that we throw first a 5, then a 2?

Of the 36 possible combinations, only 1 is favourable: 5 followed by 2. So the probability is 1/36.

If E and F were the events considered in the previous section, then E can occur first in 3 ways, F second in 3 ways. We can pair any occurrence of E with any of F, giving $3\times 3 = 9$ favourable outcomes. So the probability of E followed by F is $9/36 = 1/4$.

In general we must suppose that there are T_1 possible outcomes of the first trial, of which $N(E)$ are occurrences of E; and T_2 in the second, $N(F)$ being occurrences of F. Then in the two trials together the total number of outcomes is $T_1\times T_2$, because any of the T_1 possibilities for the first can be followed by any of the T_2 possibilities for the second. In the same way the number of ways in which E can occur first, followed by F, is $N(E)\times N(F)$. So

$$\begin{aligned} p(E \text{ followed by } F) &= \frac{N(E)\times N(F)}{T_1T_2} \\ &= \frac{N(E)}{T_1}\times\frac{N(F)}{T_2} \\ &= p(E)\times p(F). \qquad (4) \end{aligned}$$

In this calculation we must assume that E and F are *independent*: that the outcome of the first trial does not alter the probabilities in the second one.

This would not be the case if, say, the second event F was 'the total thrown is 4'. For if the first throw is 4 or more, the chance of success on the second is 0; if the first throw is 1, 2, or 3 the chance of success on the second is 1/6.

The notion of independence can be formulated in terms of probability spaces. In applications, one takes as hypothesis the independence of the real-world events to be considered, applies the theory, and tests the result by experiment.

Paradoxical Dice

Often our intuition about probabilities is wrong. Consider four dice A, B, C, D marked

A:	0	0	4	4	4	4
B:	3	3	3	3	3	3
C:	2	2	2	2	7	7
D:	1	1	1	5	5	5.

(The precise arrangement of faces does not matter.)

What is the probability that in a single throw die A will have a higher number showing than die B?

B always throws a 3. If A throws 4, which happens 4 times out of 6, he wins. If he throws 0, which happens 2 times out of 6, he loses. Therefore

A *beats* B *with probability* 2/3.

If B is thrown in competition with C it will win when C shows 2, lose when C shows 7. So

B *beats* C *with probability* 2/3.

If C plays against D matters are more complicated. With probability 1/2, D shows 1, and then C always wins; with probability 1/2 D shows 5 and C wins by showing 7 with probability 1/3. The probability that C will win is therefore

$$\tfrac{1}{2}.1+\tfrac{1}{2}.\tfrac{1}{3} = \tfrac{1}{2}+\tfrac{1}{6} = \tfrac{2}{3}.$$

Thus

C *beats* D *with probability* 2/3.

Finally, look at D versus A. If D shows 5, with probability 1/2, then D always wins. If D shows 1, with probability 1/2, then D wins if A shows 0, which has probability 1/3. The probability that D will win is

$$\tfrac{1}{2}.1+\tfrac{1}{2}.\tfrac{1}{3} = \tfrac{2}{3}.$$

Thus

D *beats* A *with probability* 2/3.

Now a die which wins more often than not is clearly 'better' than one which loses more often than it wins. In these terms,

A is better than B
B is better than C
C is better than D
and D is better than A.

There is nothing wrong with these calculations. If you play the game in practice, and let your opponent choose his die, then you can always choose another that gives you odds of 2:1 for a win.

We expect that A better than B better than C better than D should mean A better than D. We're wrong. In the present context the meaning of 'better than' depends on the choice of dice: we are really playing four *different* games. It is as if we had four people playing games: Alfred beats Bertram at tennis, Bertram beats Charlotte at chess, Charlotte beats Dierdre at badminton – and Dierdre beats Alfred at shove-halfpenny.

Those economists who believe that commodities can be ordered by majority preference might take note of this phenomenon.

Binomial Bias

Imagine a *biased* coin. Instead of coming down heads and tails with equal frequency, it has a preference for one particular side.

Such a coin provides a model for many probabilistic processes. If we are throwing a die and are interested only in whether a 6 turns up, we are effectively dealing with a biased coin such that $p(\text{head}) = 1/6$, $p(\text{tail}) = 5/6$. If we are looking at the sex of new-born babies, we have $p(\text{boy}) = 0{\cdot}52$, $p(\text{girl}) = 0{\cdot}48$.

In general we let

$$p = p(\text{head})$$
$$q = p(\text{tail})$$

and of course $p+q = 1$, because from (1) above

$$p(\text{head})+p(\text{tail}) = p(\text{head or tail}) = 1.$$

Using the theory of independent events we easily find the following list of probabilities for sequences of heads and tails:

H	p	HH	p^2	HHH	p^3
T	q	HT	pq	HHT	p^2q
		TH	pq	HTH	p^2q
		TT	q^2	HTT	pq^2
				THH	p^2q
				THT	pq^2
				TTH	pq^2
				TTT	q^3.

What is the probability that we throw a given number (0, 1, 2, or 3) of heads? We have to group together sequences with the same number of heads. Thus for 2 heads in 3 throws we get

HHT, HTH, THH, each with probability p^2q, which gives a total probability of $3p^2q$. Similar calculations give another table:

		number of heads			
		0	1	2	3
number	1	q	p		
of	2	q^2	$2pq$	p^2	
throws	3	q^3	$3pq^2$	$3p^2q$	p^3

The rows of this table should look familiar: compare the expansions

$$(q+p)^1 = q+p$$
$$(q+p)^2 = q^2+2pq+p^2$$
$$(q+p)^3 = q^3+3pq^2+3p^2q+p^3.$$

The terms on the right are exactly the entries in the table. The next row ought to come from

$$(q+p)^4 = q^4+4pq^3+6p^2q^2+4p^3q+p^4$$

and it is good practice to check that it does. In general, the entries in the nth row will be the terms of the expansion of

$$(q+p)^n.$$

This is not a coincidence, and it is not difficult to explain it. To expand, say, $(q+p)^5$ we must work out

$$(q+p)(q+p)(q+p)(q+p)(q+p).$$

The terms with exactly 3 qs come from products like this:

$$\begin{array}{ccccc} q & q & q & p & p \\ q & q & p & q & p \\ q & q & p & p & q \\ q & p & q & q & p \\ q & p & q & p & q \\ q & p & p & q & q \\ p & q & q & q & p \\ p & q & q & p & q \\ p & q & p & q & q \\ p & p & q & q & q. \end{array}$$

These correspond exactly to the 10 possible sequences of 3 tails and 2 heads:

T	T	T	H	H
T	T	H	T	H
T	T	H	H	T
T	H	T	T	H
T	H	T	H	T
T	H	H	T	T
H	T	T	T	H
H	T	T	H	T
H	T	H	T	T
H	H	T	T	T.

Obviously the same holds in general. If we write $\binom{n}{r}$ for the number of sequences of n *H*s and *T*s containing exactly r *H*s and $(n-r)$ *T*s, then the probability of getting exactly r heads in n throws is

$$\binom{n}{r} p^r q^{n-r}.$$

It isn't too hard to work out what $\binom{n}{r}$ is. If we choose the r positions for *H*s then everything is determined; so $\binom{n}{r}$ is just the number of ways of choosing r things from n. This, it can be shown, is given by

$$\binom{n}{r} = \frac{n(n-1)(n-2)\ldots(n-r+1)}{r(r-1)(r-2)\ldots 1}.$$

Thus for sequences of 2 heads and 3 tails we want

$$\binom{5}{2} = \frac{5\cdot 4}{2\cdot 1} = 10$$

which is correct.

The general expansion is

$$(q+p)^n = q^n + npq^{n-1} + \ldots + \binom{n}{r} p^r q^{n-r} + \ldots + p^n.$$

This is the *Binomial Theorem*, usually credited to Isaac Newton. It may or may not be coincidence that at one period Newton was Master of the Royal Mint.

The average number of heads obtainable in n throws can be calculated from this formula, and it turns out to be np. Thus the frequency with which heads occur is np/n, which is p. So we have come full circle to the idea of a probability as an 'average frequency of occurrence'. This theorem, which in a stronger form is called the *Law of Large Numbers*, shows how our mathematical model connects up with observations in the real world.

Random Walks

In the final section of this chapter I want to discuss another type of problem arising in probability theory. It has applications to questions about electrons bouncing around inside crystals, and particles floating in a liquid.

Imagine a particle starting at position $x = 0$ on the x-axis, at time $t = 0$. In time $t = 1$ it moves to the point $x = -1$ with probability 1/2, or to the point $x = +1$ with probability 1/2. If it is in position x at time t, then at time $t+1$ it moves either to $x-1$ or $x+1$, each with probability 1/2. What can we say about the particle's subsequent motion?

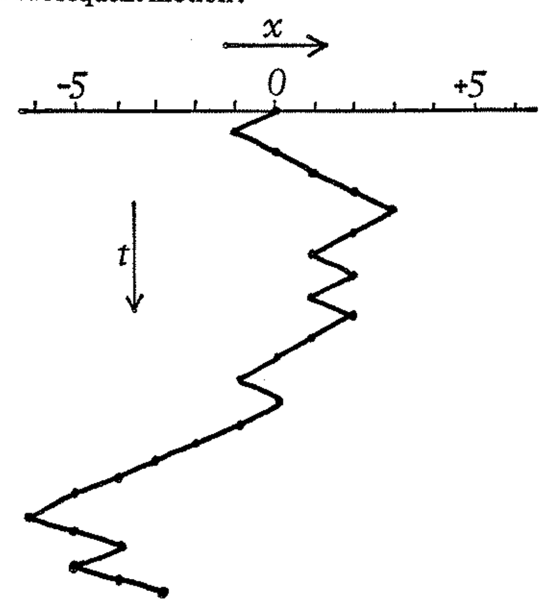

Figure 173

For example, it may move left and right according to the sequence

LRRRRLLRLRLLLRLLLLLLRRLRR

in which case its motion is as in Figure 173, with the path stretched a little in the t direction for clarity. This is a fairly typical path: if you want, you can use a coin to decide between L and R and construct other paths for yourself.

Instead of a line we could consider a plane: now the particle moves either 1 unit up, down, left, or right; each with probability 1/4. Or a 3-dimensional walk, with 6 possible directions each having probability 1/6.

Especially interesting is the question: given any other point X, what is the probability that eventually (we don't mind how long it takes) the particle reaches X?

One would expect this to decrease as X gets further away from the origin. In fact it does nothing of the kind: the probability remains the same for *all* X. In a random walk every point is as good as every other point, in the long run.

For a random walk in 1 or 2 dimensions, this probability is 1. It is almost certain that the particle will reach any given point X. (I say 'almost' because it may not do so: it could rush off to the right and never be seen again. But this has probability 0. With infinite processes, it is not quite true[2] that probability 1 means 'certain' and 0 'impossible'.)

But in 3 dimensions, the probability is only 0·24.

If you were lost in 1- or 2- dimensional space, and wandered about at random, then with probability 1 you would eventually find your way home. In 3 dimensions your chances of getting home are less than 1 in 4.

However, in all cases, the average time it would take you to arrive home is *infinite*. More precisely, pick any time t_0 – it might be 5 seconds or 3000 years. Then if you keep wandering, on most occasions you will be away from home for a time larger than t_0.

Chapter 18 Computers and Their Uses

'A hydrodynamicist was reading a research paper translated from the Russian, and was puzzled by references to a "water sheep". It transpired that the paper had been translated using a computer: the phrase in question should have read "hydraulic ram"' – Cautionary tale

Strictly speaking, computing is not part of mathematics, but a discipline in its own right. The computer is not a *concept* of modern mathematics: it is a *product* of modern technology. Nevertheless most modern mathematics courses in schools include a certain amount of computing, and quite rightly so; for the computer is a powerful tool of great practical importance in the applications of mathematics to the modern world.

On the whole computers play no role in theoretical mathematics. In order to be able to put a problem on to a computer one must *in principle* know exactly how to perform the steps necessary to solve it. From the theoretical viewpoint this means that the problem is as good as solved already, especially if the main concern is *method*. But to get results (which in practical applications are of course the main desideratum) a method which works in principle is not enough: it must also work in practice. The importance of the computer lies in its ability to bridge the gap between principle and practice.

The computer is also of interest to the mathematician because of the mathematical ideas which underly its construction.

I want, in this chapter, to give some small idea of both the mathematical and practical considerations behind the design and use of computers. For the technical details the reader must consult a specialist book.[1]

Binary Notation

Basically the computer is a calculating machine. That is to say it is given data, in the form (usually) of numbers, told how to