A. The concepts of probability and independence

The main mathematical concept needed in genetics is that of probability. When we speak of the 'probability' of an event, we mean the frequency with which that event occurs in a long sequence of trials. Thus the probability that a six will turn up in a single throw of a six-sided die is approximately 1/6. For most dice it is not exactly 1/6, because the spots are marked with small depressions on the surface, so that the six face is lighter than the others and so finishes uppermost more often. The important points then are:

- (i) The probability of an event is defined as the frequency with which it occurs in a long sequence of trials: i.e. it is the number of 'successes' (e.g. sixes) divided by the total number of 'trials' (e.g. throws). A probability is therefore a number lying between o (the event never happens) and I (the event always happens).
- (ii) All statements of probability rest ultimately on empirical measurements. Thus we know that the probability of a six is 1/6, not merely because a die has six sides, but because such dice have been thrown a large number of times, and the six falls upper-most on about 1/6 of the throws.

To take a genetical example, it is approximately true that the probability that baby born in this country will be a boy is one half. Actually the fraction of all babies born that are boys varies somewhat. In England and Wales at the present time approximately 106 boys are born for every 100 girls; but to simplify the argument the proportion of boys will be taken to be exactly one half.

We can then ask the question: what is the probability that a family consisting of two children will consist of two boys? A simple but erroneous argument is as follows: there are three possible kinds of family—two boys, two girls, and one boy and one girl—so the

3 [57] SMI

probability of two boys is one third. The argument is false because families of one boy and one girl are more probable (i.e. occur more frequently) than the other two kinds of family.

There are two lines of argument which lead to the correct conclusion. The first is to argue that when birth order is taken into account, there are four kinds of family and not three—boy boy, boy girl, girl boy and girl girl. If we assume that these four are equally frequent, the probability of two boys is one quarter. Observation shows this answer to be correct, but why was it correct to assume that these four kinds of family are equally frequent?

The assumption that has been made becomes a little clearer if we express the argument in a different form, as follows. In one half of all families of two, the older child is a boy; and in one half of these, the second child is also a boy. Hence the probability of two boys is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. Similarly, the probability of each of the other kinds of family is $\frac{1}{4}$. In this argument, we have not only assumed that half of all births are boys, but also that this remains true of the second child when it is known that the first child was a boy. In other words, we have assumed that the sex of the second child is *independent* of that of the first.

The concept of independence is important. It is defined as follows. Let there be two events, A and B, with probabilities P(A) and P(B). The events are independent if the probability that both occur, P(A and B), is equal to $P(A) \times P(B)$. In other words, two events are independent if the frequency with which both events occur is equal to the product of the frequencies of the two events taken singly. In the example, event A is that the older child is a boy, and event B that the younger child is a boy; we assumed that the probability that both are boys is equal to $P(A) \times P(B)$.

The grounds for assuming that two events are independent are ultimately empirical. Thus if one quarter of all families of two children in fact consist of two boys, we can take this as evidence that the sexes of the first and second child are independent. Actually there are slightly more families of two boys than would be expected from the frequency of male births. There are many possible reasons for this; for example, some women may provide a uterine environment more favourable to the survival of male foetuses, others of

female foetuses. But the effect is a small one, and in what follows I shall assume that the sexes of successive births are independent.

To summarise:

- (i) The probability of an event is the frequency with which it occurs in a long sequence of trials. If two events are mutually exclusive and the only ones possible (e.g. head and tail in a single toss, or boy and girl in a single birth), and their probabilities are p and q respectively, then p+q=1.
- (ii) Two events are said to be independent if the probability that both occur (e.g. that in two tosses of a die, both are sixes) is equal to the product of the probabilities of the two events taken separately (i.e. $1/6 \times 1/6 = 1/36$).
- (iii) Statements about the probabilities and about the independence of events rest ultimately on empirical evidence.

B. The binomial theorem

Assuming that the sex ratio is 1:1 and that the sexes of successive children are independent, we can ask what is the probability, in a family of three children, of 1 boy and 2 girls. It is fairly easy to see the answer to this question, but as an introduction to more difficult questions it is worth taking it slowly.

We first ask what is the probability of a particular kind of family, taking into account birth order, with 1 boy and 2 girls—for example boy girl girl. The answer is clearly $\frac{1}{2} \times \frac{1}{2} = \frac{1}{8}$.

Next we ask how many kinds of families there are with 1 boy and 2 girls. The answer is three—boy girl girl, girl boy girl and girl girl boy. Hence the probability of a family with 1 boy and 2 girls is $3 \times \frac{1}{8} = \frac{3}{8}$.

This is an example of a theorem known as the binomial theorem. Before stating this in its general form, a more difficult example will be given. If an albino and a coloured mouse are crossed, the F_1 (i.e. the first generation offspring) are coloured, and in the F_2 (i.e. the second generation obtained by mating together two individuals from the F_1) we 'expect' 3 coloured to 1 white. By 'expect', we mean that if we count a large number of offspring from such crosses, one quarter will be white. Suppose we have a single litter of $5F_2$ mice;

what is the probability that it will consist of 3 coloured and 2 white mice?

As before, we calculate first the probability of a particular litter, allowing for birth order; for example, a litter in which the first two mice born were white and the last three coloured, represented as follows:

birth position I 2 3 4 5 colour
$$\circ$$
 \circ \bullet

The probability that the first two mice will be white is $\frac{1}{4} \times \frac{1}{4}$, and that the last three will be coloured is $\frac{3}{4} \times \frac{3}{4} \times \frac{3}{4}$. Hence the probability of this particular litter is $(\frac{1}{4})^2(\frac{3}{4})^3$.

We now want to know how many kinds of litter there are with two white mice, allowing for birth order. This is equivalent to asking 'In how many ways can I select two birth positions, corresponding to the two white mice, out of five?' The first position can be selected in 5 ways, and once this has been done, there are 4 ways in which the second position can be selected. This might suggest that there are $5 \times 4 = 20$ kinds of litter, but this is a mistake. In the twenty litters, we would count the litter represented above twice; we would count a litter in which the first birth position was selected first and the second birth position was selected second, and also a litter in which the second position was selected first and the first position was selected second. The same is true for all other kinds of litter—for example, $0 \bullet \bullet 0$. Thus the number of distinguishable kinds of litter is $\frac{1}{2}(5 \times 4) = 10$. You should satisfy yourself of the truth of this statement by listing the ten possible litters.

Hence the probability of a litter with 2 white and 3 coloured mice is

$$10 \times (\frac{1}{4})^2 \times (\frac{3}{4})^3 = \frac{135}{512}$$

Notice that the answer could be written

$$\frac{5!}{2!3!}(\frac{1}{4})^2(\frac{3}{4})^3,$$

where 5! denotes $5 \times 4 \times 3 \times 2 \times 1$, and is read as '5 factorial'. This suggests the following general theorem: If in each of

n independent trials, the probability of a success is p, then the probability of r successes is

$$\frac{n!}{r!(n-r)!}p^r(1-p)^{n-r}.$$

If, as is customary, we write 1 - p = q, we can reformulate this theorem as follows:

If in each of n independent trials the probability of a success is p, the probabilities of 0, 1, 2, ..., r, ..., n successes are given by successive terms of the expansion of $(q+p)^n$... i.e. by

$$q^n$$
, $nq^{n-1}p$, $\frac{n(n-1)}{2}q^{n-2}p^2$, ... $\frac{n!}{r!(n-r)!}q^{n-r}p^r$, ... p^n .

No formal proof of this theorem will be given. The proof resembles that given in the particular case just considered. Thus the probability that the first r trials out of n are successes and the rest failures is $p^r(1-p)^{n-r}$, and the number of ways of choosing r objects

(i.e. successful trials) out of
$$n$$
 (total trials) is $\frac{n!}{r!(n-r)!}$.

This 'binomial theorem' can be used to calculate the probability of any particular family from known parents. There is one additional trick which is often useful in calculating probabilities. Suppose for example that we want to know the probability that, in a litter of $8 F_2$ mice from a cross of an albino to a coloured mouse, there will be at least one white mouse. It would be very laborious to calculate the probabilities that there will be 1, 2, 3 ... up to 8 white mice, and add these probabilities together. Fortunately there is no need. We calculate the probability that there will be no white mice, i.e.

$$P(o) = (\frac{3}{4})^8 \simeq o \cdot 1001.$$

Then the probability that there is at least one white mouse follows directly, because

$$P \text{ (at least I)} = I - P(0) \simeq 0.8999.$$

In other words, when it is laborious to calculate the probability that something will happen, try calculating the probability that it won't.

c. Conditional probability

So far we have considered only events which are independent. But we often want to know the probability that two events will occur, when the probability of one of them depends on whether the other has or has not occurred.

What for example is the probability that two cards drawn at random from a pack are both spades? The probability that the first card drawn is a spade is $13/52 = \frac{1}{4}$. Once a spade has been drawn, there are 51 cards left in the pack, of which 12 are spades. Hence the probability that the second card is a spade, given that the first one is, is 12/51 and not $\frac{1}{4}$. Hence the probability that both are spades is $\frac{1}{4} \times 12/51 = 1/17$. (At this point, you should ponder the fact that the same conclusion would follow if both cards were drawn simultaneously).

This class of problem arises in contexts other than gambling. Suppose for example we want to know the probability that a girl who is a member of a family of three children has an older brother. Obviously this will depend on whether the girl is the first, second or third child in the family. To solve the problem, it will help to be clear what we mean by probability in this case. A probability refers to the frequency of 'successes' to 'trials'. In this case, we might imagine ourselves collecting all the girls in this country, and asking each one whether she belonged to a family of three children. Those who answered yes would constitute our population of 'trials'. Each of these would then be asked whether she had an older brother; those answering yes a second time would be 'successes'.

Of girls belonging to families of three children:

\frac{1}{3} would be the first child, and of these none would have an older brother;

\frac{1}{3} would be the second child, and of these half would have an older brother;

 $\frac{1}{3}$ would be the third child, and of these $\frac{3}{4}$ would have an older brother ($1 - \frac{1}{4} = \frac{3}{4}$ is the probability that, of two older sibs, at least one is a boy).

Thus the required probability is

$$\frac{1}{3} \times 0 + \frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times \frac{3}{4} = 5/12.$$

The procedure adopted here is made easier in complex cases by introducing a new notation. We write

P(A) = The probability (P) that A happens.

P(A|H) = The probability that A happens, given that H is the case.

Then if H_1 , H_2 and H_3 are three states of affairs of which one and only one is the case $(H_1, H_2 \text{ and } H_3 \text{ are then said to be 'mutually exclusive and the only possible'), then$

$$P(A) = P(A|H_1) \times P(H_1) + P(A|H_2) \times P(H_2) + P(A|H_3) \times P(H_3). \tag{4.1}$$

No formal proof of this theorem will be given. It is merely a shorthand way of writing the procedure adopted for the problem of the girl and her elder brother, with A meaning 'the girl has an elder brother', and H_1 , H_2 and H_3 meaning that the girl is, respectively, the first, second and third child out of three. The method can be applied for any finite number of possible conditions $H_1, H_2, ..., H_n$.

 H_1 , H_2 , etc., are sometimes referred to as 'hypotheses', but the term is unfortunate, since it suggests that they resemble a universal hypothesis such as Avogadro's hypothesis. If they did, it would be absurd to write $P(H_1)$, etc., since one cannot sensibly speak of the frequency with which Avogadro's hypothesis is the case. In the example considered, H_1 , etc., refer to particular states of affairs which are true in a certain proportion of cases.

The main utility of (4.1) will emerge in the next chapter, but a few examples will be given here. Returning to the F_2 between an albino and a coloured mouse, suppose that a single coloured F_2 mouse, whom we will call Minnie, is crossed to an albino, and a litter of 5 obtained. What is the probability that all five are coloured?

Clearly this will depend on whether Minnie is homozygous coloured (CC) or heterozygous for the albino gene (Cc). In the former case all her offspring will be coloured; in the latter, we expect I coloured to I albino. Thus we can write $P(\text{all 5 coloured}) = P(\text{all 5 coloured} | \text{Minnie is } CC) \times P(\text{Minnie is } CC) + P(\text{all 5 coloured} | \text{Minnie is } Cc) \times P(\text{Minnie is } Cc) = I \times \frac{1}{3} + (\frac{1}{2})^5 \times \frac{2}{3} = I7/48$.

Now consider a more difficult example. Suppose that in man blue eyes are recessive to brown—i.e. that BB and Bb have brown eyes and bb have blue eyes (this is only approximately true). Suppose also that in a particular human population 36 % of people have blue eyes, 16 % are homozygous for brown eyes, and 48 % are heterozygous. Given that a man has blue eyes and has a brother, what is the probability that the brother's eyes are also blue?

Clearly this depends on whether the man's parents were, for example, both bb, in which case the brother is certain to have blue eyes, or were for example both Bb, in which case the brother may well have brown eyes.

The first step is to set out a table, giving the different kinds of marriage, with their frequencies, and the proportion of blue-eyed children, as follows:

Father		Mother	Frequency of marriage	Proportion of blue-eyed children in family	blue-	Relative oportions of eyed children population
			f	Þ		$n = f \times p$
BB	×	BB	0.19×0.19	0		0
$egin{array}{c} BB \ bb \end{array}$	×	}	2×0·16×0·36	•		•
$egin{array}{c} BB \ Bb \end{array}$	×	= - }	2×0·16×0·48	0		•
Bb	×	Bb	o·48 × o·48	ł		0.0576
$egin{smallmatrix} Bb \ bb \end{matrix}$	×	}	2 × 0·48 × 0·36	$\frac{1}{2}$		0.1728
bb	×	bb	0·36 × 0·36	I		0·1296
				•	Total	0.3600

In writing down the frequencies of different kinds of marriages, we have assumed that the genotypes of husband and wife are independent as far as eye colour is concerned. This would not be the case if, for example, blue-eyed people tend to marry one another. But some assumption has to be made if the original question is to be answered.

It is easy to make mistakes when writing down tables of this kind; fortunately a number of checks are available. First, since there are 3 kinds of parent there are 9 kinds of marriage, and we have listed them all. Second, we should verify that the sum of the entries in the f column is unity. (A final check is provided by the fact that the frequency of blue-eyed children, 0.3600, equals the frequency of blue-eyed parents. But this is only the case, as will emerge in the next chapter, because we have assumed 'random mating', and have chosen frequencies for the parental genotypes which fit the 'Hardy-Weinberg' ratio.)

It follows from the table that of all blue-eyed children, a fraction $\frac{0.0576}{0.3600}$ have two brown-eyed parents, $\frac{0.1728}{0.3600}$ have one brown-eyed

parent, and $\frac{0.1296}{0.3600}$ have two blue-eyed parents.

We are now in a position to use (4.1) to calculate the probability, P, that if a blue-eyed man has a brother, the brother will also have blue eyes:

$$P = \frac{0.0576}{0.3600} \times \frac{1}{4} + \frac{0.1728}{0.3600} \times \frac{1}{2} + \frac{0.1296}{0.3600} \times I = 0.64.$$

A reason why one might wish to find such a probability will emerge later, when discussing twin diagnosis.

D. Inverse probability

Calculations of probability start by assuming the truth of some general propositions (e.g. that half the babies born are boys, and that the sexes of successive babies are independent) and calculating the probability (i.e. frequency) of some particular event (e.g. families of 2 boys and 2 girls). The problem of inverse probability is to start from the fact that a particular event or group of events takes place, and to calculate the probability that some general proposition is true. So formulated, the problem is clearly insoluble, and indeed is meaningless if by 'probability' we mean 'frequency of occurrence'. General propositions cannot be true in a certain proportion of cases.

Thus suppose for example we use the genes vestigial and aristapedia in Drosophila in an experiment intended to test Mendel's law of independent assortment, and obtain in the F_2 numbers closely agreeing with the 'expected' 9:3:3:1 ratio. We cannot then ascribe a numerical value to the probability that Mendel's law is true, because a probability is a measure of frequency, and Mendel's law is not sometimes true of *vestigial* and *aristapedia* and sometimes false.

What we can do is calculate the probability of getting this observed result if Mendel's law is true. We can go further, and calculate, again assuming Mendel's law is true, the probability of obtaining a result whose fit with the expected ratio is as bad as or worse than the one we actually obtained—this is what is usually calculated in statistical 'significance tests'. But there is a class of problems in which we adopt a kind of inverse argument to calculate the probability that some proposition is true. The proposition must not be a general one, but one which is sometimes true and sometimes false. An example will make the method clear.

Suppose that a woman of blood group O marries an AB man, and has a pair of boy twins of blood group B. If this is all we know, what is the probability that the twins are monovular—i.e. from a single egg?

Our difficulty is this: if we knew that a pair of twins from such a marriage were monovular, it would be easy to calculate the probability that both had the B blood group; but we have been asked to solve the inverse problem. However, let us do the easy part first. The genetics of the situation is as follows:

$$\begin{array}{ccc}
O & \times & A \\
\hline
O & & B
\end{array}$$

$$\begin{array}{cccc}
O & & O \\
\frac{1}{2}A & & \frac{1}{2}B
\end{array}$$
i.e. group A i.e. group B

Thus for binovular twins, the probability of two B children is $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$. For monovular twins, the probability that the first born is B is $\frac{1}{2}$; if he is, the second is sure to be. Hence the probability of two B twins is $\frac{1}{2}$.

Using the notation introduced on page 63, we will write these conclusions as follows:

$$P(2B|bin) = \frac{1}{4}; P(2B|mon) = \frac{1}{2}.$$

0/

What we want to know is P(mon|2B)—i.e. the probability that the twins are monovular if both are B. This problem we can only solve if we have some *a priori* knowledge of the frequency of monovular and binovular twins, knowing only that they are of the same sex.

This we can deduce from the observation that 32 % of all twin pairs are of unlike sex (the proportion varies from population to population). These 32 % are necessarily all binovular, and since equal numbers of binovular pairs will be of like and of unlike sex, a further 32 % of twin pairs must be binovular pairs of like sex, giving 64 % of binovular pairs in all. This leaves 36 % of monovular pairs. The argument is illustrated in fig. 20.

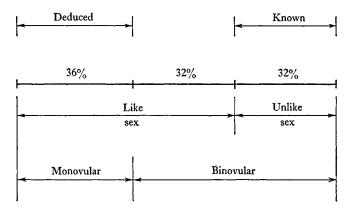


Fig. 20. Method of estimating the frequency of monovular twins.

It follows that of all like-sex twins, a fraction 0.36/0.68 are monovular, and 0.32/0.68 are binovular.

Hence, if we were to collect all the pairs of boy twins from marriages of O $\mathcal{Q} \times AB\mathcal{J}$ we would find

$$\frac{0.32}{0.68} \times \frac{1}{4}$$
 are binovular, and both B,
 $\frac{0.36}{0.68} \times \frac{1}{2}$ are monovular, and both B.

0.09

and

Hence, the probability that a pair of boy twins are monovular, given that they are both B, is

$$P(\text{mon}|2B) = \frac{0.36}{0.68} \times \frac{I}{2} \div \left(\frac{0.36}{0.68} \times \frac{I}{2} + \frac{0.32}{0.68} \times \frac{I}{4} \right) = \frac{9}{13}.$$

This is the solution to our problem. Notice that if we had written the *a priori* probabilities that a like-sex twin pair are monovular and binovular as P(mon) and P(bin) respectively, our solution has the form

$$P(\text{mon}|2B) = \frac{P(2B|\text{mon}) \times P(\text{mon})}{P(2B|\text{mon}) \times P(\text{mon}) + P(2B|\text{bin}) \times P(\text{bin})}.$$

This is a special case of Bayes' theorem. Thus suppose H_1 and H_2 are propositions of which one and only one is true, and B is some result whose probability depends on which proposition is true. The probability that both H_1 and B are true can be written $P(H_1 + B)$.

Then
$$P(H_1+B) = P(H_1)P(B|H_1)$$
 and similarly
$$P(H_1+B) = P(B)P(H_1|B).$$
 Hence
$$P(H_1|B) = \frac{P(B|H_1)P(H_1)}{P(B)}$$
 and
$$P(B) = P(B|H_1)P(H_1) + P(B|H_2)P(H_2)$$
 and so
$$P(H_1|B) = \frac{P(B|H_1)P(H_1)}{P(B|H_1)P(H_1) + P(B|H_2)P(H_2)}$$
 (4.2)

which is Bayes' Theorem. The theorem can be extended to cases where three or more alternative *a priori* propositions exist. The theorem can be used to calculate the *a posteriori* probability of some proposition H_1 , in the light of additional evidence B, provided that

- (i) a priori probabilities of H_1 and not- H_1 , in the absence of knowledge about B, are known.
- (ii) the probability of B, given that H_1 or not- H_1 is the case, can be calculated.

The first point is crucial; in the examples considered, the frequency of monovular pairs among like-sexed twins was known.

The theorem can be applied to twin diagnosis even if the genotypes of the parents are not known, provided that the frequencies of genotypes in the population are known and that random mating can be assumed. For example, what is the probability that two blueeyed boy twins are monovular? With the usual notation

$$P(\text{mon}|2b) = \frac{P(2b|\text{mon})P(\text{mon})}{P(2b|\text{mon})P(\text{mon}) + P(2b|\text{bin})P(\text{bin})}.$$

Now, if we assume the same genotype frequencies as on page 64, 36 % of people have blue eyes.

Hence P(2b|mon) = 0.36.

We have already calculated that for a pair of brothers, and hence for a pair of binovular twins, if one has blue eyes, there is a probability of 0.64 that the other has. Hence

$$P(2b|\text{bin}) = 0.36 \times 0.64$$
.

And assuming as before that

$$P(\text{mon}) = 0.36/0.68$$
 and $P(\text{bin}) = 0.32/0.68$, we have

$$P(\text{mon}|2b) = 0.36 \times \frac{0.36}{0.68} \div \left(0.36 \times \frac{0.36}{0.68} + 0.36 \times 0.64 \times \frac{0.32}{0.68}\right)$$

= 0.638.

Before it was known that both twins were blue-eyed, the probability that they were monovular was 0.36/0.68 = 0.530; the additional evidence has raised the probability to 0.638.

Examples

- Two unbiased dice are thrown. What is the probability that the numbers showing (a) add up to 9; (b) differ by 2; (c) are different?
 - 2 A red and a green die are thrown. What is the probability that
- (a) the number on the red die is even and the number on the green die is less than 3;
- (b) the number on the red die is less than three or the number on the green die is more than three;
- (c) the number on the red die is 5 given that the sum of the spots on the two dice is 9 or more?
 - 3 5 cards are drawn from a normal pack of 52. What is the

probability that (a) they are all the same suit; (b) they include 4 aces?

- 4 Albinism in mice is due to a recessive gene. An albino is crossed to a pure-bred coloured mouse, and a second generation (F_2) litter of 4 mice is obtained (expectation, 3 coloured: 1 albino). What is the probability that the litter will contain
 - (a) 3 coloured and 1 albino;
 - (b) at least 1 albino;
 - (c) 1 albino, 2 heterozygous coloured, 1 pure-bred coloured;
 - (d) 2 albino females and 2 coloured males?
- 5 What proportion of girls from families of 4 children have at least two older brothers?
- 6 (a) A coloured mouse called Minnie from an F_2 litter similar to that described in question 4 is crossed to an albino male. She has a litter of 5. What is the probability that at least one of them is an albino?
- (b) In fact, Minne's litter consists of 5 coloured mice. What is the probability that Minnie carries an albino gene?
- 7 Of three prisoners, Matthew, Mark and Luke, two are to be executed, but Matthew does not know which. He therefore asks the jailer 'Since either Mark or Luke are certainly going to be executed, you will give me no information about my own chances if you give me the name of one man, either Mark or Luke, who is going to be executed.' Accepting this argument, the jailer truthfully replied 'Mark will be executed'. Thereupon, Matthew felt happier, because before the jailer replied his own chances of execution were 2/3, but afterwards there are only two people, himself and Luke, who could be the one not to be executed, and so his chance of execution is only \frac{1}{2}.

Is Matthew right to feel happier?

(This should be called the Serbelloni problem since it nearly wrecked a conference on theoretical biology at the villa Serbelloni in the summer of 1966; it yields at once to common sense or to Bayes' theorem.)