1

INTRODUCTION

Please, do read the Preface first!

Note. Capitalized 'Probability' and 'Statistics' (when they do not start a sentence) here refer to the *subjects*. Thus, Probability is the study of probabilities, and Statistics that of statistics. (Later, we shall also use 'Statistics' as opposed to 'statistics' in a different, technical, way.) 'Maths' is English for 'Math', 'Mathematics' made friendly.

1.1 Conditioning: an intuitive first look

One of the main aims of this book is to teach you to 'condition': to use conditional probabilities and conditional expectations effectively. Conditioning is an extremely powerful and versatile technique.

In this section and (more especially) the next, I want you to give your intuition free rein, using common sense and not worrying over much about rigour. All of the ideas in these sections will appear later in a more formal setting. Indeed, because fudging issues is *not* an aid to clear understanding, we shall study things with a level of precision very rare for books at this level; but that is for *later*. *For now, we use common sense,* and though we would later regard some of the things in these first two sections as 'too vague', 'non-rigorous', etc, I think it is right to begin the book as I do. **Intuition is much more important than rigour**, though we do need to know how to back up intuition with rigour.

In a subject in which it is easy to be misled by arguments which initially look plausible, our intuition needs constant honing. One is much more likely to make a mistake in elementary Probability than in (say) elementary Group Theory. We *all* make mistakes; and you should check out carefully everything *I* say in this book.

Always try to find counter-arguments to what I claim, shooting them down if I am indeed correct. Some notes on the need for care with intuition occupy the next subsection.

Since I shall later be developing the subject rather carefully, *it is important that you know from the beginning that it is one of the most enjoyable branches of mathematics*. I want you to get into the subject a little before we start building it up systematically. These first two sections contain some exercises for you – some of which (in the next section particularly) are a little challenging. In this connection, do remember that this is a 'first run through' this material: we shall return to it all later; and the later exercises when the book starts properly will give you more practice. If you peep ahead at the very extensive Section 4.1 for example, you will see one of the places where you get more practice with conditional probability.

Note. Several of our examples derive from games of chance. Probability did originate with the study of such games, and *their study is still of great value for developing one's intuition*. It also remains the case that *techniques developed for these 'frivolous' purposes continue to have great real-world benefit when applied to more important areas*. This is Wigner's 'unreasonable effectiveness of Mathematics' again.

A. Notes on the need for care with intuition. I mention some of the reasons why our intuition needs sharpening.

Aa. Misuse of the 'Law of Averages' in real-world discussions. All humans, even probabilists in moments of weakness, tend to misuse the 'Law of Averages' in everyday life. National Lotteries thrive on the fact that a person will think "I haven't won anything for a year, so I am more likely to win next week", which is, of course, nonsense. The (Un)Holy Lottery Machines behave *independently* on different weeks: they do not 'remember what they have done previously and try to balance out'.

A distinguished British newspaper even gave advice to people on how to choose a 'good' set of six numbers from the set $\{1, 2, \ldots, 49\}$ for the British National Lottery. (To win, you have to choose the same six numbers as the Lottery Machine.) It was said that the six chosen numbers should be 'randomly spread out' and should have an average close to 25. It is clear that the writer thought that Choice A of (say) the six numbers $8, 11, 19, 21, 37, 46$ is more likely to win than Choice B of the six numbers $1, 2, 3, 4, 5, 6$. Of course, on every week, these two choices have exactly the same chance of winning. Yet the average of all numbers chosen by the Lottery Machine over a year is very likely to be very close to 25; and this tends to 'throw' people.

One hears misuse of the 'Law of Averages' from sports commentators every week.

Some people tend to think that a new coin which has been tossed 6 times and landed 6 Heads, is more likely to land Tails on the next toss 'to balance things out'. They do not realize that what happens in future will swamp the past: if the coin is tossed a further

1000 times, what happened in the 6 tosses already done will essentially be irrelevant to the proportion of Heads in all 1006 tosses.

Now I know that *You* know all this. But (Exercise!) what do *you* say to the writer about the Lottery who says, "Since the average of all the numbers produced by the Lottery Machine over the year is very likely to be close to 25, then, surely, I would be better to stick with Choice A throughout a year than to stick with Choice B throughout the year." The British are of course celebrated for being 'bloody-minded', and the reality is that a surprisingly large number stick to Choice B!

Ab. Some common errors made in studying the subject. Paradoxically, *the most common mistake when it comes to actually doing calculations is to assume 'independence' when it is not present*, the 'opposite' of the common 'real-world' mistake mentioned earlier. We shall see a number of examples of this.

Another common error is to assume that things are equally likely when they are not. (The infamous 'Car and Goats' problem 15P even tripped up a very distinguished Math Faculty.) Indeed, I very deliberately avoid concentration on the 'equally likely' approach to Probability: it is an invitation to disaster.

Our intuition finds it very hard to cope with the sometimes perverse behaviour of ratios. The discussion at 179Gb will illustrate this spectacularly.

B. A first example on conditional probability. Suppose that 1 in 100 people has a certain disease. A test for the disease has 90% accuracy, which here means that 90% of those who do have the disease will test positively (suggesting that they *have* the disease) and 10% of those who do *not* have the disease will test positively.

One person is chosen at random from the population, tested for the disease, and the test gives a positive result. That person might be inclined to think: "I have been tested and found 'positive' by a test which is accurate 90% of the time, so there is a 90% chance that I have the disease." However, it is much more likely that the randomly chosen person does not have the disease and the test is in error than that he or she does have the disease and the test is correct. Indeed, we can reason as follows, using 'K' to signify 'thousand' (1000, not 1024) and 'M' for 'million'.

Let us suppose that there are 1M people in the population. Suppose that they are all tested. Then, amongst the 1M people,

about $1\text{M} \times 99\% = 990\text{K}$ would *not* have the disease, of whom
about $(1\text{M} \times 99\%) \times 10\% = 99\text{K}$ would test positively;

and

$1\text{M} \times 1\% = 10\text{K}$ *would* have the disease, of whom
about $(1\text{M} \times 1\%) \times 90\% = 9\text{K}$ would test positively.

So, a total of about $99\text{K} + 9\text{K} = 108\text{K}$ would test positively, of whom only 9K would actually have the disease. In other words, *only* $1/12$ *of those who would test positively actually have the disease.*

Because of this, we say that the *conditional probability* that a randomly chosen person does have the disease *given* that that person is tested with a positive result, is 1/12.

[[*Note.* At 6G below, we shall modify the interpretation of conditional probability just given, in which we have imagined sampling without replacement of the entire population, to one valid in all situations where we are forced to imagine instead 'sampling with replacement' to ensure 'independence'. The numerical value of the conditional probability is here unaffected.]]

I do not want to get involved in 'philosophical' questions at this stage, but, provided you understand that this book contains such things only to the minimal extent consistent with clarity, I mention briefly now a point which will occur in other forms much later on.

C. Discussion: Continuation of Example 3B. Now consider the situation where the experiment has actually been performed: a *real* person with an actual name – let's say it is Homer Simpson – has been chosen, and tested with a positive result. Can we tell Homer that the probability that he has the disease is 1/12? Do note that we are assuming that Homer is the person chosen at random; and that all we know about him is that his test proved positive. It is *not* the case (for example) that Homer is consulting his doctor because he fears he may have caught a sexually transmitted disease.

A Possible Frequency-School View. The problem is that there is no randomness in whether or not Homer has the disease: either he does have it, in which case the probability that he has it (conditional on any information) is 1; or he does not have it, in which case the probability that he has it (conditional on any information) is 0. All that we can say to Homer is that if every person were tested, then the fraction of those with positive results who would have the disease is 1/12; and in this sense he can be 11/12 'confident' that he does not have the disease. [The Note above on sampling with replacement applies here too, but does not really concern us now.] It is not very helpful to tell Homer only that the probability that he has the disease is either 0 or 1 but we don't know which.

The Bayesian-School View. If we take the contrasting view of the Bayesian School of Statistics, then we can interpret 'probability that a statement is true' as meaning 'degree of belief in that statement'; and then we *can* tell Homer that the probability (in this new sense) that he has the disease is 1/12.

Remarks. The extent to which the difference between the schools in this case is a matter of 'Little-endians versus Big-endians' is up to you to decide for yourself later. (The dispute in *Gulliver's Travels* was over which way up to put an egg in an eggcup. If your primary concern is with eating the egg, ...)

In this book, I am certainly happy to tell Homer that the conditional probability that he has the disease is 1/12. I shall sell him a copy of the book so that he can decide what that statement means to him.

The logic which we used in Example 3B is the uncontroversial and incontrovertible logic of Bayes' Theorem – a *theorem* – in Probability. We shall

study the theorem as part of the full mathematical theory in Chapter 4. However, we shall develop many of its key ideas in this section.

D. Orientation: The Rules of Probability. Probability, the mathematical theory, is based on two simple rules: an Addition Rule and a Rule for Combining Conditional Probabilities, both of which we have in effect seen in our discussion of Example 3B. The Addition Rule is taken as *axiomatic* in the mathematical theory; the Rule for Combining Conditional Probabilities is really just a matter of *definition*. The subject is developed by application of logic to these rules. Especially, *no attempt is made to define 'probability' in the real world*. The *'long-term relative frequency' (LTRF)* idea is the key *motivation* for Probability, though, as we shall see, it is impossible to make it into a rigorous definition of probability.

The LTRF idea motivates the axioms; but once we have the axioms, we forget about the LTRF until its reappearance as a theorem, part of the Strong Law of Large Numbers.

► **E. LTRF motivation for Probability.** Let A be an event associated with some experiment $\mathcal{E}$, so that A might, or might not, occur when $\mathcal{E}$ is performed. Now consider a Super-experiment $\mathcal{E}^\infty$ which consists of an infinite number of independent performances of $\mathcal{E}$, 'independent' in that no performance is allowed to influence others. Write $N(A,n)$ for the number of occurrences of A in the first n performances of $\mathcal{E}$ within the Super-experiment. Then the LTRF idea is that

$$\frac{N(A,n)}{n} \text{ converges to } \mathbb{P}(A) \tag{E1}$$

in some sense, where $\mathbb{P}(A)$ is the probability of A, that is, the probability that A occurs within experiment $\mathcal{E}$. If our individual experiment $\mathcal{E}$ consists of tossing a coin with probability p of Heads *once*, then the LTRF idea is that if the coin is thrown *repeatedly*, then

$$\frac{\text{Number of Heads}}{\text{Number of tosses}} \to p \tag{E2}$$

in some sense. Here, A is the event that 'the coin falls Heads' in our individual experiment $\mathcal{E}$, and $p = \mathbb{P}(A)$. Since the coin has no memory, we believe (and postulate in mathematical modelling) that it behaves independently on different tosses.

The **certain event** Ω (Greek Omega), the event that 'something happens', occurs on every performance of experiment $\mathcal{E}$. The **impossible event** $\emptyset$ ('nothing happens') never occurs. The LTRF idea suggests that

$$\mathbb{P}(\Omega) = 1, \quad \mathbb{P}(\emptyset) = 0.$$

Note that in order to formulate and prove the Strong Law, we have to set up a model for the Super-experiment $\mathcal{E}^\infty$, and we have to be precise about 'in some sense'.

►► **F. Addition Rule for Two Events.** If A and B are events associated with our experiment $\mathcal{E}$, and these events are *disjoint* (or *exclusive*) in that it is impossible for A and B to occur simultaneously on any performance of the experiment, and if

$A \cup B$ is the event that 'A happens or B happens',

then, of course,

$$N(A \cup B, n) \;=\; N(A, n) + N(B, n).$$

The appropriateness of the set-theoretic 'union' notation will become clear later. If we 'divide by n and let n tend to ∞' we obtain LTRF motivation – but not proof – of the Addition Rule for Two Events:

$$\text{if } A \text{ and } B \text{ are disjoint, then } \quad \mathbb{P}(A \cup B) \;=\; \mathbb{P}(A) + \mathbb{P}(B). \tag{F1}$$

Later, we take (F1) as an axiom.

[[For example, if our individual experiment $\mathcal{E}$ is that of tossing a coin twice, then

$$\mathbb{P}(1 \text{ Head in all}) \;=\; \mathbb{P}(\mathtt{HT}) + \mathbb{P}(\mathtt{TH}), \tag{F2}$$

where, of course, HT signifies 'Heads on the 1st toss, Tails on the 2nd'.]]

If A is any event, we write

A^c for the event 'A does not occur'.

Then A and A^c are disjoint, and $A \cup A^c = \Omega$: precisely one of A and A^c occurs within our experiment $\mathcal{E}$. Thus, $1 = \mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$, so that

$$\mathbb{P}(A^c) \;=\; 1 - \mathbb{P}(A).$$

[[Hence, for our coin, $\mathbb{P}(\text{it falls Tails}) = q := 1 - p$.]]

► **G. The LTRF motivation for conditional probability.** Let A and B be events associated with our experiment $\mathcal{E}$, with $\mathbb{P}(A) \neq 0$. The LTRF motivation is that we regard the *conditional probability* $\mathbb{P}(B \mid A)$ *that* B *occurs given that* A *occurs* as follows. Suppose again that our experiment is performed 'independently' infinitely often. Then (the LTRF idea is that) $\mathbb{P}(B \mid A)$ is the *long-term proportion of those experiments on which* A *occurs that* B (also) *occurs, in other words, that both* A *and* B *occur.*
In other words, if

$A \cap B$ is the event that 'A and B occur simultaneously',

then we should have

$$\mathbb{P}(B \mid A) = \text{limit in some sense of } \frac{N(A \cap B, n)}{N(A, n)}$$
$$= \text{limit in some sense of } \frac{N(A \cap B, n)/n}{N(A, n)/n} = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

in our experiment $\mathcal{E}^\infty$. In the mathematical theory, we *define*

$$\mathbb{P}(B \mid A) := \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}. \tag{G1}$$

Suppose that our experiment $\mathcal{E}$ has actually been performed in the real world, and that we are told only that event A has occurred. Bayesians would say that $\mathbb{P}(B \mid A)$ is then our 'probability as degree of belief that B (also) occurred'. Once the experiment has been performed, whether or not B has occurred involves no randomness from the Frequentist standpoint. A Frequentist would have to quote: 'the long-term proportion of those experiments on which A occurs that B (also) occurs is (whatever is the numerical value of) $\mathbb{P}(B \mid A)$'; and in this sense $\mathbb{P}(B \mid A)$ represents our 'confidence' that B occurred in an actual experiment on which we are told that A occurred.

With this Frequentist view of probability, we should explain to Homer that if the experiment 'Pick a person at random and test him or her for the disease' were performed *independently* a very large number of times, then on a proportion $11/12$ of those occasions on which a person tested positively, he or she would *not* have the disease. To guarantee the independence of the performances of the experiment, we would have to pick each person from the *entire* population, so that the same person might be chosen many times. It is of course assumed that if the person is chosen many times, no record of the results of any previous tests is kept. This is an example of sampling with replacement. We shall on a number of occasions compare and contrast sampling with, and sampling without, replacement when we begin on the book proper.

►► **H. General Multiplication Rule.** We have for any 2 events A and B,

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B \mid A), \tag{H1}$$

this being merely a rearrangement of (G1).

For any 3 events A, B and C, we have, for the event $A \cap B \cap C$ that all of A, B and C occur simultaneously within our experiment $\mathcal{E}$,

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}\big((A \cap B) \cap C\big) = \mathbb{P}(A \cap B)\mathbb{P}(C \mid A \cap B),$$

whence

$$\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B \mid A)\mathbb{P}(C \mid A \cap B). \tag{H2}$$

The extension to 4 or more events is now obvious.

I. A decomposition result. Let A and B be any two events. The events $G := A \cap B$ and $H := A^c \cap B$ are disjoint, and $G \cup H = B$. (*Clarification.* We are decomposing B according to whether or not A occurs. If B occurs, then either 'A occurs and B occurs' or 'A does not occur and B occurs'.) We have

$$\mathbb{P}(B) \;=\; \mathbb{P}(G) + \mathbb{P}(H) \;=\; \mathbb{P}(A)\mathbb{P}(B \mid A) + \mathbb{P}(A^c)\mathbb{P}(B \mid A^c). \qquad \text{(I1)}$$

This, the simplest decomposition, is extremely useful.

Ia. Example 3B revisited. In that example, let
B be 'chosen person has the disease'
A be 'chosen person tests positively'.
We want to find $\mathbb{P}(B \mid A)$. We are given that

$$\mathbb{P}(B) = 1\%, \quad \mathbb{P}(B^c) = 99\%, \quad \mathbb{P}(A \mid B) = 90\%, \quad \mathbb{P}(A \mid B^c) = 10\%.$$

We have, keeping the calculation in the same order as before,

$$\begin{aligned} \mathbb{P}(A) \;&=\; \mathbb{P}(B^c \cap A) + \mathbb{P}(B \cap A) \;=\; \mathbb{P}(B^c)\mathbb{P}(A \mid B^c) + \mathbb{P}(B)\mathbb{P}(A \mid B) \\ &=\; (0.99 \times 0.10) + (0.01 \times 0.90) \;=\; 0.108 \quad (= 108K/1M). \end{aligned}$$

We now know $\mathbb{P}(A \cap B)$ and $\mathbb{P}(A)$, so we can find $\mathbb{P}(B \mid A)$.

►► **J. 'Independence means Multiply'.** If A and B are two events, then we say that A and B are *independent* if

$$\mathbb{P}(A \cap B) \;=\; \mathbb{P}(A)\mathbb{P}(B), \qquad \text{(J1)}$$

one of several assertions which we shall meet that 'Independence means Multiply'. If $\mathbb{P}(A) = 0$, no comment is necessary. If $\mathbb{P}(A) \neq 0$, then we may rearrange (J1) as

$$\mathbb{P}(B \mid A) \;=\; \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} \;=\; \mathbb{P}(B),$$

which says that the information that A occurs on some performance of $\mathcal{E}$ does not affect 'our degree of belief that B occurs' on that same performance.

If we consider the experiment 'Toss a coin (with probability p of Heads) twice', then, we believe that, since the coin has no memory, the results of the two tosses will be independent. (The laws of physics would be very different if they are not!) Hence we have

$$\mathbb{P}(\mathtt{HT}) \;=\; \mathbb{P}(\text{Heads on first toss}) \times \mathbb{P}(\text{Tails on second}) \;=\; pq,$$

where q, the probability of Tails, is $1 - p$; and, using the Addition Rule as at 6(F2), we get the familiar answer that the probability of 'exactly one Head in all' is $pq + qp = 2pq$.

The Multiplication Rules for n independent events follow from the General Multiplication Rules at 7H similarly. If we toss a coin 3 times, the chance of getting HTT is, of course, pqq.

K. Counting. I now begin a discussion (continued in the next section) of various 'counting' and 'conditioning' aspects of the famous binomial-distribution result for coin tossing. The 'counting' approach may well be familiar to you; but, in the main, I want you to condition rather than to count.

▶ **Ka. Lemma.** *For non-negative integers r and n with $0 \le r \le n$, the number $\binom{n}{r}$, also denoted by ${}^nC_r$, of subsets of $\{1, 2, \ldots, n\}$ of size r is*

$$\binom{n}{r} = {}^nC_r = \frac{n!}{r!(n-r)!},$$

where, as usual,

$$n! := n(n-1)(n-2)\ldots 3.2.1, \qquad 0! := 1.$$

If $r < 0$ or $r > n$, we define ${}^nC_r := \binom{n}{r} := 0$.

Remarks. Here and everywhere, we adopt the standard convention in Maths that 'set' means 'unordered set': $\{1, 2, 3\} = \{3, 1, 2\}$. There are indeed ${}^4C_2 = 6$ subsets of size two of $\{1, 2, 3, 4\}$, namely, $\{1, 2\}$, $\{1, 3\}$, $\{1, 4\}$, $\{2, 3\}$, $\{2, 4\}$, $\{3, 4\}$. The empty set is the only subset of $\{1, 2, \ldots, n\}$ of size 0, even if $n = 0$.

The official rigorous proof would take one through the steps of the following lemma. Part (a) is not actually relevant for this purpose, but is crucial for other results.

▶ **Kb. Lemma.** *(a) The number of* ordered *r-tuples $(i_1, i_2, \ldots, i_r)$ where each i_k is chosen from $\{1, 2, \ldots, n\}$ is n^r. (You will probably know from Set Theory that the set of all such r-tuples is the* Cartesian product *$\{1, 2, \ldots, n\}^r$.)*
(b) For $0 \le r \le n$, the number ${}^nP_r$ of ordered *r-tuples $(i_1, i_2, \ldots, i_r)$ where each i_k is chosen from $\{1, 2, \ldots, n\}$ and $i_1, i_2, \ldots, i_r$ are* distinct *is given by*

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

(c) The number of permutations of $\{1, 2, \ldots, n\}$ is $n!$.
(d) Lemma Ka is true.

Proof. In Part (a), there are n ways of choosing i_1, and, for each of these choices, n ways of choosing i_2, making $n \times n = n^2$ ways of choosing the ordered pair (i_1, i_2). For each of these n^2 choices of the ordered pair (i_1, i_2), there are n ways of choosing i_3; and so on.

In Part (b), there are n ways of choosing i_1, and, for each of these choices, $n-1$ ways of choosing i_2 (because we are now not allowed to choose i_1 again), making $n \times (n-1)$

ways of choosing the ordered pair (i_1, i_2). For each of these $n(n-1)$ choices of the ordered pair (i_1, i_2), there are $n-2$ ways of choosing i_3; and so on.

Part (c) is just the special case of Part (b) when $r = n$.

Now for Part (d). By Part (c), each subset of size r of $\{1, 2, \ldots, n\}$ gives rise to $r!$ ordered r-tuples $(i_1, i_2, \ldots, i_r)$ where $i_1, i_2, \ldots, i_r$ are the distinct elements of the set in some order. So it must be the case that $r! \times {}^nC_r = {}^nP_r$; and this leads to our previous formula for ${}^nC_r$. □

L. 'National Lottery' Proof of Lemma 9Ka. One can however obtain clearer intuitive understanding of Lemma 9Ka by using conditioning rather than counting as follows. Yes, a certain amount of intuition goes into the argument too.

A British gambler (who clearly knows no Probability) pays 1 pound to choose a subset of size r of the set $\{1, 2, \ldots, n\}$. (In Britain, $n = 49$, and $r = 6$.) The Lottery Machine later chooses 'at random' a subset of size r of the set $\{1, 2, \ldots, n\}$. If the machine chooses exactly the same subset as our gambler, then our gambler wins the 'jackpot'. It is clear that our gambler wins the 'jackpot' with probability $1/\binom{n}{r}$, and we can find $\binom{n}{r}$ from this probability.

Now, the probability that the first number chosen by the machine is one of the numbers in our gambler's set is clearly r/n. The *conditional probability* that the second number chosen by the machine is in our gambler's set *given* that the first is, is clearly $(r-1)/(n-1)$, because, given this information about the first, at the time the machine chooses its second number, there are $n-1$ 'remaining' numbers, $r-1$ of which are in our gambler's set. By Multiplication Rule 7(H1), the probability that the first *two* numbers chosen by the machine are in our gambler's set is

$$\frac{r}{n} \times \frac{r-1}{n-1}.$$

By extending the idea, we see that the probability that the machine chooses exactly the same set as our gambler is

$$\frac{r}{n} \times \frac{r-1}{n-1} \times \frac{r-2}{n-2} \times \cdots \times \frac{1}{n-r+1} = \frac{r!}{{}^nP_r} = \frac{r!(n-r)!}{n!}.$$

(In Britain, then, the probability of winning the jackpot is very roughly 1 in 14 million.) We have proved Lemma 9Ka.

For more on our gambler, see Exercise 17Rb below.

M. Binomial(n, p) distribution. Now, we can combine the ideas of 8J with Lemma 9Ka.

▶ **Ma. Lemma.** *If a coin with probability p of Heads is tossed n times, and we write Y for the total number of Heads obtained, then Y has the probability mass function of the binomial(n, p) distribution:*

$$\mathbb{P}(Y = r) = b(n, p; r) := \binom{n}{r} p^r (1-p)^{n-r}.$$