Elementary information theory

D. S. JONES

D. S. JONES
University of Dundee

Elementary information theory

Preface

A one term's course on information theory has a number of advantages. It makes relatively few demands on the previous knowledge of the student and so can be placed anywhere convenient in the curriculum. Touching on the structure of language and having applications to computing it is suitable for students with a wide variety of interests. Important theorems can be reached without great effort and yet the techniques are sufficiently testing to stretch the student. The course can be followed, if desired, by more sophisticated units dealing with algebraic coding theory, cryptography, linguistics, error analysis, and optical communication, to choose a few examples.

The course presented in this book is largely self-contained. After a discussion of definitions there is a treatment of elementary coding theory including optimal binary codes. Another chapter deals with the capacity of a channel for discrete sources. Error correcting codes and continuous information are considered in further chapters. Exercises at various levels of difficulty are provided at the ends of chapters. Starred sections indicate topics from which a selection may be made to suit particular requirements.

My thanks are due to my wife Ivy for her continued support and forebearance, and to Mrs. D. Ross for managing to remain cheerful while typing the manuscript.

Dundee D. S. J.
January 1978

Contents

1 Elements of probability

The theory of information is based on notions drawn from probability. Indeed, some people regard information theory as a branch of applied probability. However, for much of the development the amount of probability theory needed is not large and the brief introduction given in this chapter should suffice for most purposes.

1.1. Probability

The starting point for the mathematical theory of probability is a real or imagined experiment such as tossing a coin, throwing a die, drawing a card from a deck, counting the number of road accidents on a day, tossing a coin ten times, and so on. After the experiment a certain *outcome* is observed, e.g. the coin toss resulted in a head, a 3 was thrown on the die, the ten of hearts was drawn from the pack.

If the number of outcomes of the experiment is finite then *finite probability* is involved; if the number of outcomes is countable then the situation is one of *discrete probability*.

Tossing a coin once is an example of finite probability because there are two possible outcomes, namely heads or tails, provided the coin is not permitted to stand on its edge. Similarly, throwing a die and drawing a card from a pack come within finite probability, there being 6 and 52 possible outcomes in the two cases. If a coin is tossed 10 times there are 1024 possible outcomes and again finite probability is concerned. On the other hand, if the experiment is tossing a coin until a head appears the possible outcomes are

$$\mathrm{H}, \mathrm{TH}, \mathrm{TTH}, \mathrm{TTTH}, \ldots$$

where H, T stand for head and tail respectively. Now the number of outcomes is infinite but countable so this experiment comes under the heading of discrete probability.

In advance of the experiment we do not know which outcome will occur, but we associate with an outcome a probability p. It is difficult to define precisely what is meant by p but the following conveys a rough idea. Suppose that the experiment is conducted n times and the particular outcome under consideration occurs f times. Then f/n is expected to approach p as $n \to \infty$.

Since f cannot be less than zero nor exceed n, it follows that p will be positive and lie somewhere between 0 and 1. Because $f = 0$ implies the non-occurrence of an outcome it is usual in discrete probability to interpret $p = 0$ as meaning that a particular outcome cannot occur. Similarly, $p = 1$ is taken to signify that an outcome is certain to occur because, if $f = n$, the same outcome appears after every experiment.

From now on, it will be understood that discrete probability (which includes finite probability) is under discussion unless otherwise is specified.

DEFINITION 1.1a *With each outcome* O_n *of all possible outcomes* $O_1, O_2, \ldots$ *of an experiment assume that there is an associated probability* $P(O_n)$ *which is positive, lies in* $[0, 1]$, *and is such that*

$$P(O_1) + P(O_2) + \cdots = 1.$$

Making $P(O_n)$ positive and in $[0, 1]$ is in conformity with properties of p described above. The last equation of Definition 1.1a is merely an assertion that it is certain that there must be one outcome from an experiment.

The notation $P(\;)$ will be used extensively to denote the probability of the occurrence of whatever is between the parentheses. Thus, in drawing a card from a pack, P(2 of diamonds) would mean the probability of picking the two of diamonds.

When the number of outcomes is finite, say n, and there is no reason to suppose that one outcome will appear in preference to any other

$$P(O_1) = P(O_2) = \cdots = P(O_n).$$

It then follows from Definition 1.1a that, in fact,

$$P(O_1) = P(O_2) = \cdots = P(O_n) = 1/n. \tag{1.1.1}$$

An ideal coin should not favour heads or tails and so (1.1.1) implies that

$$P(\mathrm{H}) = P(\mathrm{T}) = \tfrac{1}{2}.$$

If it should happen that $P(\mathrm{H}) = \frac{2}{3}$, $P(\mathrm{T}) = \frac{1}{3}$ the coin would not be ideal but *loaded* or *biased* in favour of heads. It will be assumed that coins are ideal unless otherwise stated.

A perfect die is one in which no face is preferentially treated so that, from (1.1.1),

$$P(1)=P(2)=\cdots=P(6)=\tfrac{1}{6}.$$

It is conventional to assume that in drawing a card from a pack (which contains 52 cards) any one is equally likely to be drawn so that

$$P(8 \text{ of clubs})=\tfrac{1}{52}.$$

Likewise, in bridge, where the pack is distributed in four hands of 13 cards, it is conventional to regard all distributions as equally likely. The verification of this assumption in bridge would require some 10^{30} experiments with well-shuffled packs and involve millions of man-years.

The probability of an *event* E can be determined once a rule has been provided which says for every possible outcome of an experiment whether or not the event E has occurred. There is an important distinction between outcomes and events. Outcomes are fixed by the experiment and are not within our control. In contrast, events can be chosen to suit our convenience.

DEFINITION 1.1b. *The probability of an event E is the sum of the probabilities of the outcomes in which E occurs.*

For example, if a coin is tossed and E is the tossing of a head

$$P(E)=P(\mathrm{H})=\tfrac{1}{2}.$$

In the experiment of tossing a coin twice, let E be the event that H appears only once. The possible outcomes are HH, HT, TH, TT; and

$$P(E)=P(\mathrm{HT})+P(\mathrm{TH})=\tfrac{1}{4}+\tfrac{1}{4}=\tfrac{1}{2}.$$

When the number of outcomes is finite and they are all equally likely, suppose k of them entail the event E. Then, from (1.1.1) and Definition 1.1b,

$$P(E)=k/n.$$

The probability of any outcome cannot be negative and so $P(E)\geqslant 0$. On the other hand, $p(e)\leqslant P(O_1)+P(O_2)+\cdots$ and so, from Definition 1.1a,

$$0\leqslant P(E)\leqslant 1. \qquad (1.1.2)$$

Of course, $1-\mathrm{P(E)}$ is the probability that E will not occur.

Let E_1 and E_2 be two events. Two new events can be defined from them. $E_1 \cup E_2$ is the event in which *either* E_1 *or* E_2 *or both occur.* $E_1 \cap E_2$ is the event in which *both* E_1 *and* E_2 *occur.* It may happen that if E_1 occurs, E_2 cannot and vice versa. Then E_1 and E_2 are said to be *mutually exclusive* and

$$P(E_1 \cap E_2) = 0. \tag{1.1.3}$$

This may be expressed as $E_1 \cap E_2 = 0$ on the understanding that 0 here stands for the null event and $P(0) = 0$.

An important result is given by the following theorem.

THEOREM 1.1.

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2)$$

and, if E_1 *and* E_2 *are mutually exclusive,*

$$P(E_1 \cup E_2) = P(E_1) + P(E_2).$$

Proof. In $P(E_1) + P(E_2)$ any outcome favourable to E_1 is counted once and any outcome favourable to E_2 is counted once. Therefore, outcomes favourable to both are counted twice. Hence, if these are subtracted once, as is done by taking away $P(E_1 \cap E_2)$, the outcomes favourable to either E_1 or E_2 or both are left. The first statement of the theorem is consequently demonstrated. The second assertion follows at once from (1.1.3) for mutually exclusive events.

According to (1.1.2), $P(E_1 \cap E_2) \geqslant 0$ and so we infer from Theorem 1.1 that

$$P(E_1 \cup E_2) \leqslant P(E_1) + P(E_2). \tag{1.1.4}$$

Let $E = E_1 \cup E_2$; then, from (1.1.4),

$$\begin{aligned} P(E_1 \cup E_2 \cup E_3) &= P(E \cup E_3) \leqslant P(E) + P(E_3) \\ &\leqslant P(E_1 \cup E_2) + P(E_3) \\ &\leqslant P(E_1) + P(E_2) + P(E_3) \end{aligned}$$

from (1.1.4) again. Obviously, the general formula is given by the following corollary.

COROLLARY 1.1.

$$P(E_1 \cup E_2 \cup \cdots) \leqslant P(E_1) + P(E_2) + \cdots$$

Example 1.1. In the experiment of tossing a coin twice let E_1 be the event that a head appears on the first toss and E_2 the event that a head occurs on the second toss. The possible outcomes are HH, HT, TH, and TT each of which has probability $\frac{1}{4}$ because they are all equally likely. The ones favourable to E_1 are HH and HT so that

$$P(E_1)=\tfrac{1}{4}+\tfrac{1}{4}=\tfrac{1}{2}.$$

Similarly, HH and TH entail E_2; so $P(E_2)=\frac{1}{2}$.

For $E_1 \cup E_2$ the relevant outcomes are HT, TH, and HH while $E_1 \cap E_2$ requires HH. Thus

$$P(E_1 \cup E_2)=\tfrac{3}{4}, \qquad P(E_1 \cap E_2)=\tfrac{1}{4}.$$

These results are consistent with Theorem 1.1 since

$$\tfrac{3}{4}=\tfrac{1}{2}+\tfrac{1}{2}-\tfrac{1}{4}.$$

1.2. Samples

Consider an alphabet of n letters $a_1, a_2, \ldots, a_n$. Pick r of them to form, say, $a_{j_1}, a_{j_2}, \ldots, a_{j_r}$. This is a *sample* of size r. Often the place from where the sample is drawn is called the population.

If each choice of letter in the sample is made from the entire alphabet any particular letter may appear more than once in the sample. This is known as *sampling with replacement* and, almost without exception, is the type of sampling which arises in the theory of information. On the other hand, if a letter once drawn cannot be picked again, the process is *sampling without replacement* and no letter can appear twice or more in a sample. There is no limit to the size of samples with replacement but, in samples without replacement, r cannot exceed n because the population is exhausted when $r=n$.

In sampling with replacement each choice of letter can be made in n ways so that the number of different samples of size r which can be selected is n^r. If each of the samples is equally probable the probability of one particular sample being chosen is $1/n^r$ from (1.1.1). Such a situation is often described as *random sampling*.

For sampling without replacement the first letter can be chosen in n ways. Once it has been selected there are $n-1$ possibilities open to the second letter. After the first and second letters have been fixed there are $n-2$ opportunities for the third. Consequently, the number of samples of size r $(r \leq n)$ is $n(n-1)\cdots(n-r+1)$. If the samples are equally probable the probability of one of them is $1/n(n-1)\cdots(n-r+1)$; again the term random sample is employed.

Keep r fixed and let $n \to \infty$. Then $n-j$ is approximately n for $j=0,\ldots,r-1$ and

$$n(n-1)\cdots(n-r+1) \sim n^r.$$

Thus the two methods of sampling are virtually equivalent for small samples from large populations.

1.3. Conditional probability

The probability of an event can alter as more is learned about it. Suppose a card is drawn from a pack then the probability that it is a queen is $\frac{4}{52}=\frac{1}{13}$. Suppose the further knowledge that the card drawn is a knave, queen, or king is available. Then the probability that it is a queen is $\frac{4}{12}$ or $\frac{1}{3}$. *Conditional probability* is the name given to probability when extra information is at our disposal.

To see how to calculate it consider a population of n people of whom b are blonde and f are female. Let E_1 be the event that a person chosen at random is blonde and E_2 the event that a person chosen at random is female. Then, from Section 1.1,

$$P(E_1)=b/n, \qquad P(E_2)=f/n.$$

Let f_b be the number of females who are blonde. Suppose it is known that the person chosen is female. Then the probability that she is blonde is f_b/f because f_b of the f possible outcomes are favourable. Let us agree to write this as $P(E_1 \mid E_2)$, meaning the probability that a person is blonde knowing that the person is female. Then

$$P(E_1 \mid E_2)=\frac{f_b}{f}=\frac{f_b}{n}\cdot\frac{n}{f}=\frac{P(E_1 \cap E_2)}{P(E_2)}.$$

This formula is the basis of the general definition of conditional probability (Definition 1.3).

Definition 1.3. *If $P(E_1)>0$ the conditional probability $P(E \mid E_1)$ is defined by*

$$P(E \mid E_1)=\frac{P(E \cap E_1)}{P(E_1)}.$$

$P(E \mid E_1)$ is the conditional probability of the event E given that the event E_1 has occurred.

When $P(E_1)=0$, $P(E \mid E_1)$ is apparently undefined but in the context in which $P(E_1)=0$ signifies that E_1 cannot occur it is meaningless to talk about the conditional probability of E when E_1 has occurred.

If Theorem 1.1 is applied to the events $E_1 \cap E$ and $E_2 \cap E$,

$$P(E_1 \cap E \cup E_2 \cap E) = P(E_1 \cap E) + P(E_2 \cap E) - P(E_1 \cap E \cap E_2 \cap E).$$

But the event $E_1 \cap E \cup E_2 \cap E$ is the same as $E_1 \cup E_2 \cap E$, and $E_1 \cap E \cap E_2 \cap E$ coincides with $E_1 \cap E_2 \cap E$. Hence

$$P(E_1 \cup E_2 \cap E) = P(E_1 \cap E) + P(E_2 \cap E) - P(E_1 \cap E_2 \cap E).$$

Division by $P(E)$ and use of Definition 1.3 gives

$$P(E_1 \cup E_2 \mid E) = P(E_1 \mid E) + P(E_2 \mid E) - P(E_1 \cap E_2 \mid E).$$

In other words, *Theorem* 1.1 *is unaffected by conditioning on an event.* As a consequence Corollary 1.1 holds under conditioning on an event.

From Definition 1.3

$$P(E_1 \mid E_2 \cap E_3) = \frac{P(E_1 \cap E_2 \cap E_3)}{P(E_2 \cap E_3)} = \frac{P(E_1 \cap E_2 \cap E_3)}{P(E_2 \mid E_3)P(E_3)}$$

whence

$$P(E_1 \cap E_2 \cap E_3) = P(E_1 \mid E_2 \cap E_3)P(E_2 \mid E_3)P(E_3). \tag{1.3.1}$$

There is no difficulty in generalizing (1.3.1) to more events, e.g.

$$\begin{aligned} P(E_1 \cap E_2 \cap E_3 \cap E_4) &= P(E_1 \mid E_2 \cap E_3 \cap E_4)P(E_2 \cap E_3 \cap E_4) \\ &= P(E_1 \mid E_2 \cap E_3 \cap E_4) \\ &\quad \times P(E_2 \mid E_3 \cap E_4)P(E_3 \mid E_4)P(E_4). \end{aligned} \tag{1.3.2}$$

Example 1.3. A carton contains 80 light bulbs of which 20 are defective. A bulb selected at random is found to be defective. What is the probability that a second bulb chosen at random is defective if the first is not replaced?

Let E_1 be the event that the first bulb is defective, E_2 the event that the second bulb is defective. We are asked for $P(E_2 \mid E_1)$.

Given that E_1 occurs there are 19 defective bulbs in 79 and so

$$P(E_2 \mid E_1)=\frac{19}{79}.$$

Since $P(E_1)=\frac{1}{4}$,

$$P(E_2 \cap E_1)=\frac{19}{79}\cdot\frac{1}{4}=\frac{19}{316}$$

which is the probability of two defectives on successive choices before any selection is made.

1.4. Independence

It may happen that the occurrence or otherwise of an event E_2 has no influence on the occurrence of E_1. In that case $P(E_1 \mid E_2)=P(E_1)$ because E_2 conveys no knowledge about E_1. It follows from Definition 1.3 that $P(E_1 \cap E_2)=P(E_1)P(E_2)$ in these circumstances. This is such an important concept that it deserves its own definition.

DEFINITION 1.4. *Two events E_1, E_2 are said to be statistically independent if, and only if,*

$$P(E_1 \cap E_2)=P(E_1)P(E_2).$$

To put it another way, two events are characterized as statistically independent when the probability of their joint occurrence is the product of their separate probabilities.

Example 1.4a. A card is drawn from a pack. The probability that it is a ten is 4/52 or 1/13. The probability that it is a spade is 13/52 or 1/4. The probability that it is the ten of spades is 1/52. Since 1/52 = (1/4)(1/13), there is agreement with the idea that the events of drawing a ten and of drawing a spade are statistically independent.

Example 1.4b. In a random permutation of a, b, c, d the event a precedes b is statistically independent of c precedes d.

Example 1.4c. In an experiment the probability that E occurs is p. If the experiment is repeated in an independent way the probability that E will occur on the second experiment is also p.

If E occurs on both experiments, the two cases are statistically independent and so Definition 1.4 implies that the probability of the double occurrence is p^2. Similarly, the probability of r occurrences on r independent experiments is p^r.

When, in r experiments, E occurs only s times ($s \leqslant r$) the probability of a particular sequence is $p^s(1-p)^{r-s}$ since $1-p$ is the probability of non-occurrence of E on an experiment (Section 1.1). The number of ways in which E can occur s times is the number of ways of selecting s slots from r, i.e. $r!/s!(r-s)!$. Hence Definition 1.1b implies that E can occur exactly s times in r independent experiments with probability

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

Note that this is consistent with the preceding paragraph when $s=r$ because $0!=1$.

Example 1.4d. The probability that a lawyer has a car accident in one year is p_1 and for a miner is p_2. If there are 5 times as many miners as lawyers, find the probability that one person selected at random from the combined group will have an accident in the second year if the person has had one in the first year.

Let E_1, E_2 be the events of an accident in the first and second years respectively. Since the lawyers form $\frac{1}{6}$ of the group and the miners $\frac{5}{6}$

$$P(E_1)=\tfrac{1}{6}p_1+\tfrac{5}{6}p_2.$$

The probability of a lawyer having an accident in both years is $p_1{}^2$ according to Example 1.4c. For miners the relevant probability is $p_2{}^2$. Hence

$$P(E_1 \cap E_2)=\tfrac{1}{6}p_1{}^2+\tfrac{5}{6}p_2{}^2.$$

Therefore

$$P(E_2 \mid E_1)=\frac{p_1{}^2+5p_2{}^2}{p_1+5p_2}$$

is the desired conditional probability.

As a numerical illustration take $p_1=0.6$, $p_2=0.06$. Then $P(E_1)=0.15$, $P(E_1 \cap E_2) \doteq 0.063$, $P(E_2 \mid E_1)=0.42$. Thus, knowing that a person has had an accident one year increases the odds (by almost a factor of 3) that the person will have a second accident, indicating that the person chosen at random has a certain proneness to accidents. This is in contrast to the probability of two successive accidents being quite small when there is no advance knowledge of an accident in one year. Statistical independence is lacking here.

The notion of statistical independence can be extended to more than two events. For example, three events are (mutually) statistically independent if and only if

$$P(E_j \cap E_k) = P(E_j)P(E_k) \qquad (j \neq k)$$
$$P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2)P(E_3),$$

i.e. not only are the events independent in pairs but also the probability of the triple is the product of the three probabilities. Similarly, for the independence of four events, they must be independent in pairs, in triples and the probability of the four must be the product of the four probabilities. The generalization to n events is immediate.

1.5. The law of large numbers

In many situations it is convenient to classify the result of an experiment as either a success (S) or a failure (F). What constitutes a success is not of concern but it can be chosen to suit our own purposes. Repeat the experiment until it has been carried out n times, each repetition being independent of the others, and count the number of times that S occurs—say, S_n. Then, if p is the probability of S on a single experiment, the behaviour of S_n is governed by the Law of Large Numbers.

LAW OF LARGE NUMBERS. *Given arbitrarily small $\varepsilon > 0$ and $\delta > 0$, then*

$$P\left(\left|\frac{S_n}{n} - p\right| < \varepsilon\right) > 1 - \delta$$

for sufficiently large n.

This is a standard result which can be found in textbooks on probability and no proof will be given here.

If it were legitimate to place ε and δ equal to zero the law would state that it was certain that S_n/n was p. Because this is illegal the most that the law suggests is that S_n/n approaches p as $n \to \infty$, in conformity with the earlier rough idea of probability in

Section 1.1. But one must not be read more into the law than is actually there. The law is a statement about probability and asserts that it is almost certain for large n that S_n/n will be near p, but not that it is absolutely certain. In other words, S_n/n can fluctuate quite widely from p (and in practice usually does) but only for rare values of n.

Example 1.5. In an infinite decimal let the occurrence of 5 be regarded as a success and any other digit as a failure. If all digits are equally likely $p = 1/10$. Then the law of large numbers says that, of the first n figures, $n/10$ will be 5 with high probability as $n \rightarrow \infty$.

Similarly, in an infinite sequence of binary digits, the first n figures will contain pn zeros with high probability as $n \rightarrow \infty$ if p is the probability of 0 occurring at any place.

Exercises

1.1. Two dice are thrown. List the possible outcomes of the experiment. Do you think that the sum of two faces is as likely to be 3 as to be 7?

1.2. From five digits 1, 2, 3, 4, 5 one is chosen and then a second selection is made from the remaining four digits. Find the probability that an odd digit will be chosen (a) the first time, (b) the second time, (c) both times.

1.3. Let every permutation of the four symbols a_1, a_2, a_3, a_4 be equally probable. Let E_j be the event that a_j appears in the jth position. Verify that

$$P(E_1 \cup E_3) = P(E_1) + P(E_3) - P(E_1 \cap E_3).$$

1.4. Two dice are thrown. E_1 is the event that the sum of the faces is odd. E_2 is the event that at least one 1 is thrown. Describe $E_1 \cup E_2$ and $E_1 \cap E_2$; find their probabilities.

1.5. A coin is tossed until the same result appears twice in succession. With every possible outcome requiring n tosses associate the probability $1/2^n$. Find the probability that the experiments ends (a) before the sixth toss, (b) after an even number of tosses.

1.6. How many different sets of initials can be formed if every person has one surname and (a) exactly two forenames, (b) at most two forenames. Deduce that in case (b) some people have the same initials in a town of 20 000 inhabitants.

1.7. Three dice are thrown. If no two faces are the same, what is the probability that one is a 3?

1.8. The probability that a man will live 10 more years is 0.4 and the probability that his wife will live 10 more years is 0.5. Find the probability (a) they will both live for 10 years, (b) at least one will live for 10 years, (c) neither will live for 10 years.

1.9. Of three cards one is marked 1 on both sides, one has 0 on both sides, and the third has 1 on one side and 0 on the other. A card is selected at random and found to have 1 on one side. What is the probability that there is 1 on the other side?

1.10. A television advertisement for perfume is seen by 40 per cent of the nation. If the probability is 0.1 that a person who sees the advertisement buys the perfume what is the probability that a person picked at random will have seen the advertisement and bought the perfume?

1.11. The probability that a child is born a boy is $\frac{1}{2}$ and the probability that a family has exactly k children is p_k with $p_0+p_1+\cdots=1$. What is the probability that a family has boys but no girls? If it is known that the family has no girls, what is the probability that it has only one child?

1.12. In Exercise 1.11, $p_k=ap^k$ for $k\geqslant 1$, a being a positive constant and $0<p<1$. Show that the probability that a family contains m boys $(m\geqslant 1)$ is $2ap^m/(2-p)^{m+1}$. Given that a family includes at least one boy, what is the probability there are two or more?

1.13. Blondes are always on time for appointments, redheads are always late, and brunettes toss a coin for each appointment to decide whether to be prompt or late. The numbers of blondes, redheads, and brunettes are in the ratio 1 : 1 : 2. If a female arrives on time what is the probability that she is (a) blonde, (b) redhead, (c) brunette. If she arrives promptly for three successive appointments, what is the probability that she is a brunette?

1.14. The events $E_1, E_2, \ldots, E_n$ are statistically independent and $P(E_k)=p_k$. Find the probability that none of the events occurs.

2 Basic concepts

Before introducing some of the definitions of information theory, it is desirable to remove one possible cause of misapprehension. Possible combinations of the letters a, n, and t are tan, ant, nat. These words may have meaning and significance for readers but their impact on individuals will vary, depending on the reader's subjective reaction. Subjective information conveyed in this way is impossible to quantify in general. Therefore the meaning of groups of symbols is excluded from the theory of information; each symbol is treated as an entity in its own right and how any particular grouping is interpreted by an individual is ignored. Information theory is concerned with how symbols are affected by various processes but not with information in its most general sense.

2.1. Self-information

Let S be a system of events $E_1, E_2, \ldots, E_n$ in which $P(E_k) = p_k$ with $0 \leqslant p_k \leqslant 1$ and

$$p_1 + p_2 + \cdots + p_n = 1.$$

Then we introduce the following definition

DEFINITION 2.1. *The self-information of the event E_k is written $I(E_k)$ and defined by*

$$I(E_k) = -\log p_k.$$

The base of the logarithm is not specified in the definition. For most of our work it will not matter what base is chosen since a change of base merely alters the scale of units. The most common bases encountered are 2 and e. With base 2, I is measured in *bits* (an abbreviation of binary digits) whereas, in base e, the units of I are *nats* (to indicate that a natural logarithm is involved). The number of nats is 0.693 times the number of bits. Normally, no special choice of base will be made (other than requiring it to exceed unity) but when the base 2 is employed this will be shown