THE ALGEBRA OF PROBABLE INFERENCE

The Algebra of Probable Inference

by Richard T. Cox
PROFESSOR OF PHYSICS
THE JOHNS HOPKINS UNIVERSITY

BALTIMORE:
The Johns Hopkins Press

Distributed in Great Britain by Oxford University Press, London

Printed in the United States of America by Horn-Shafer Co., Baltimore

Library of Congress Catalog Card Number: 61-8039

to my wife Shelby

Preface

This essay had its beginning in an article of mine published in 1946 in the *American Journal of Physics*. The axioms of probability were formulated there and its rules were derived from them by Boolean algebra, as in the first part of this book. The relation between expectation and experience was described, although very scantily, as in the third part. For some years past, as I had time, I have developed further the suggestions made in that article. I am grateful for a leave of absence from my duties at the Johns Hopkins University, which has enabled me to bring them to such completion as they have here.

Meanwhile a transformation has taken place in the concept of entropy. In its earlier meaning it was restricted to thermodynamics and statistical mechanics, but now, in the theory of communication developed by C. E. Shannon and in subsequent work by other authors, it has become an important concept in the theory of probability. The second part of the present essay is concerned with entropy in this sense. Indeed I have proposed an even broader definition, on which the resources of Boolean algebra can be more strongly brought to bear. At the end of the essay, I have ventured some comments on Hume's criticism of induction.

Writing a preface gives a welcome opportunity to thank my colleagues for their interest in my work, especially Dr. Albert L. Hammond, of the Johns Hopkins Department of Philosophy, who was good enough to read some of the manuscript, and Dr. Theodore H. Berlin, now at the Rockefeller Institute in New York but recently with the Department of Physics at Johns Hopkins. For help with the manuscript it is a pleasure to thank Mrs. Mary B.

Rowe, whose kindness and skill as a typist and linguist have aided members of the faculty and graduate students for twenty-five years.

I have tried to indicate my obligations to other writers in the notes at the end of the book. Even without any such indication, readers familiar with *A Treatise on Probability* by the late J. M. Keynes would have no trouble in seeing how much I am indebted to that work. It must have been thirty years or so ago that I first read it, for it was almost my earliest reading in the theory of probability, but nothing on the subject that I have read since has given me more enjoyment or made a stronger impression on my mind.

The Johns Hopkins University R. T. C.
BALTIMORE, MARYLAND

Contents

THE ALGEBRA OF PROBABLE INFERENCE

I

Probability

1. *Axioms of Probable Inference*[1]

A probable inference, in this essay as in common usage, is one entitled on the evidence to partial assent. Everyone gives fuller assent to some such inferences than to others and thereby distinguishes degrees of probability. Hence it is natural to suppose that, under some conditions at least, probabilities are measurable. Measurement, however, is always to some extent imposed upon what is measured and foreign to it. For example, the pitch of a stairway may be measured as an angle, in degrees, or it may be reckoned by the rise and run, the ratio of the height of a step to its width. Either way the stairs are equally steep but the measurements differ because the choice of scale is arbitrary. It is therefore reasonable to leave the measurement of probability for discussion in later chapters and consider first what principles of probable inference will hold however probability is measured. Such principles, if there are any, will play in the theory of probable inference a part like that of Carnot's principle in thermodynamics, which holds for all possible scales of temperature, or like the parts played in mechanics by the equations of Lagrange and Hamilton, which have the same form no matter what system of coordinates is used in the description of motion.

It has sometimes been doubted that there are principles valid over the whole field of probable inference. Thus Venn wrote in his *Logic of Chance:*[2]

"In every case in which we extend our inferences by Induction or Analogy, or depend upon the witness of others, or trust to our own memory of the past, or come to a conclusion through conflicting arguments, or even make a long and complicated deduction by mathematics or logic, we have a result of which we can scarcely feel as certain as of the premises from which it was obtained. In all these cases then we are conscious of varying quantities of belief, but are the laws according to which the belief is produced and varied the same? If they cannot be reduced to one harmonious scheme, if in fact they can at best be brought to nothing but a number of different schemes, each with its own body of laws and rules, then it is vain to endeavour to force them into one science."

In this passage, the first of three sentences distinguishes types of inference which common usage calls probable, the second asks whether inferences of these different kinds are subject to the same laws and the third implies that they are not. Nevertheless, if we look for them, we can find likenesses among these examples and likenesses also between these and others which would be accepted as proper examples of probability by all the schools of thought on the subject. Venn himself belonged to the school of authors who define probability in statistical terms and restrict its meaning to examples in which it can be so defined.[3] By their definition, they estimate the probability that an event will occur under given circumstances from the relative frequencies with which it has occurred and failed to occur in past instances of the same circumstances. Every instance in which it has occurred strengthens the argument that it will occur in a new instance and every contrary instance strengthens the contrary argument. Thus, whenever they estimate a probability in the restricted sense their definition allows and the way their theory prescribes, they "come to a conclusion through conflicting arguments," as do the advocates of other definitions and theories. The argument, moreover, which makes one inference more probable makes the contradictory inference less probable and thus the two probabilities stand in a mutual relation. In this all schools can agree and

it may be taken as an axiom on any definition of probability that:

> The probability of an inference on given evidence determines the probability of its contradictory on the same evidence. (1.i)

Continuing with Venn's list of varieties of probable inference, let us consider the probability of the right result in "a long and complicated deduction in mathematics" and compare it with the probability of a long run of luck at cards or dice, a classical example in the theory of probability. In any game of chance, a long run of luck is, of course, less probable than a short one, because the run may be broken by a mischance at any single toss of a die or drawing of a card. Similarly, in a commonplace example of mathematical deduction, a long bank statement is less likely to be right at the end than a short one, because a mistake in any single addition or subtraction will throw it out of balance. Clearly we are concerned here with one principle in two examples. A mathematical deduction involving more varied operations in its successive steps or a chain of reasoning in logic would provide only another example of the same principle.

The uncertainties of testimony and memory, also cited by Venn, come under this principle as well. Consider, for example, the probability of the assertion, made by Sir John Maundeville in his *Travels*, that Noah's Ark may still be seen on a clear day, resting where it was left by the receding waters of the Flood, on the top of Mount Ararat. For this assertion to be probable on Sir John's testimony, it must first of all be probable that he made it from his recollection rather than his fancy. Then, on the assumption that he wrote as he remembered what he saw or heard told, it must be probable also that his memory could be trusted against a lapse such as might have occurred during the long years after he left the region of Mount Ararat and before he found in his writing a solace from his "rheumatic gouts" and his "miserable rest." Finally, on the assumption that his testimony was honest and his memory sound, it must be probable that he or those on whom he depended could be sure that they had truly seen Noah's

Ark, a matter made somewhat doubtful by his other statement that the mountain is seven miles high and has been ascended only once since the Flood.

Every assertion which, like this one, involves the transmission of knowledge by a witness or its retention in the memory is, on this account, a conjunction of two or more assertions, each of which contributes to the uncertainty of the joint assertion. For this reason, it comes under the same principle which we saw involved in the probability of a run of luck at cards and which can be stated in the following axiom:

> The probability on given evidence that both of two inferences are true is determined by their separate probabilities, one on the given evidence, the other on this evidence with the additional assumption that the first inference is true. (1.ii)

Thus the uncertainties of testimony and memory, of long and complicated deductions and conflicting arguments—all the specific examples in Venn's list—have traits in common with one another and with the classical examples provided by games of chance.

The more general subjects of induction and analogy, also mentioned in the quotation from Venn, must be reserved for discussion in later chapters, but the examples already considered may serve to launch an argument that all kinds of probable inference can be "reduced to one harmonious scheme."[4]

For this reduction, the argument will require only the two axioms just given, when they are implemented by the logical rules of Boolean algebra.[5]

2. *The Algebra of Propositions*

Ordinary algebra is the algebra of quantities. In our use of it here, quantities will be denoted by italic letters, as a, b, A, B. Boolean algebra is the algebra, among other things, of propositions. Propositions will be denoted here by small boldface let-

ters, as **a**, **b**, **c**. The meaning of a proposition in Boolean algebra corresponds to the value of a quantity in ordinary algebra. For example, just as, in ordinary algebra, a certain quantity may have a constant value throughout a given calculation or a variable one, so, in Boolean algebra, a proposition may have a fixed meaning throughout a given discourse or its meaning may vary according to the context within the discourse. Thus "Socrates is a man" is a familiar proposition of constant meaning in logical discourse, whereas the proposition, "I agree with all that the previous speaker has said," has a meaning variable according to the occasion. For another example of the same correspondence, just as an ordinary algebraic equation, such as

$$(a + b)c = ac + bc,$$

states that two quantities, although different in form, are nevertheless the same in value, so a Boolean equation states that two propositions of different form are the same in meaning.

Of the signs used for operations peculiar to Boolean algebra, we shall need only three, $\sim$, $\cdot$ and $\vee$, which denote respectively *not*, *and* and *or*.[6] Thus the proposition *not* **a**, called the *contradictory* of **a**, is denoted by $\sim\mathbf{a}$. The relation between **a** and $\sim\mathbf{a}$ is a mutual one, either being the other's contradictory. To deny $\sim\mathbf{a}$ is therefore to affirm **a**, so that

$$\sim\sim\mathbf{a} = \mathbf{a}.$$

The proposition **a** *and* **b**, called the *conjunction* of **a** and **b**, is denoted by $\mathbf{a}\cdot\mathbf{b}$. The order of propositions in the conjunction is the order in which they are stated. In ordinary speech and writing, if propositions describe events, it is customary to state them in the chronological order in which the events take place. So the nursery jingle runs, "Tuesday we iron and Wednesday we mend." It would have the same meaning, however, if it ran, "Wednesday we mend and Tuesday we iron." In this example, therefore, and also in general,

$$\mathbf{b}\cdot\mathbf{a} = \mathbf{a}\cdot\mathbf{b}.$$

Similarly the expression $\mathbf{a}\cdot\mathbf{a}$ means only that the proposition **a** is stated twice and not that an event described by **a** has occurred twice. Rhetorically it is more emphatic than **a**, but logically it is the same. Thus

$$\mathbf{a}\cdot\mathbf{a} = \mathbf{a}.$$

Parentheses are used in Boolean as in ordinary algebra to indicate that the expression they enclose is to be treated as a single entity in respect to an operation with an expression outside. They designate an order of operations, in that any operations indicated by signs in the enclosed expression are to be performed before those indicated by signs outside. The parentheses are unnecessary if the order of operations is immaterial. Thus $(\mathbf{a}\cdot\mathbf{b})\cdot\mathbf{c}$ denotes the proposition obtained by first conjoining **a** with **b** and then conjoining $\mathbf{a}\cdot\mathbf{b}$ with **c**, whereas $\mathbf{a}\cdot(\mathbf{b}\cdot\mathbf{c})$ denotes the proposition obtained by first conjoining **b** with **c** and then conjoining **a** with $\mathbf{b}\cdot\mathbf{c}$, but the propositions obtained in these two sequences of operations have the same meaning and the parentheses may therefore be omitted. Accordingly,

$$(\mathbf{a}\cdot\mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot(\mathbf{b}\cdot\mathbf{c}) = \mathbf{a}\cdot\mathbf{b}\cdot\mathbf{c}.$$

The proposition **a** *or* **b**, called the *disjunction* of **a** and **b**, is denoted by $\mathbf{a} \vee \mathbf{b}$. It is to be understood that *or* is used here in the sense intended by the notice, "Anyone hunting or fishing on this land will be prosecuted," which is meant to include persons who both hunt and fish along with those who engage in only one of these activities. This is to be distinguished from the sense intended by the item, "coffee or tea," on a bill of fare, which is meant to offer the patron either beverage but not both. Thus $\vee$ has the meaning which the form *and/or* is sometimes used to express.

Let us now consider expressions involving more than one of the signs, $\sim$, $\cdot$ and $\vee$. In this consideration it should be kept in mind that $\sim\mathbf{a}$ is not some particular proposition meant to contradict **a** item by item. For example, if **a** is the proposition, "The dog is

small, smooth-coated, bob-tailed and white all over except for black ears," $\sim\mathbf{a}$ is not the proposition, "The dog is large, wire-haired, long-tailed and black all over except for white ears." To assert $\sim\mathbf{a}$ means nothing more than to say that **a** is false at least in some part. If **a** is a conjunction of several propositions, to assert $\sim\mathbf{a}$ is not to say that they are all false but only to say that at least one of them is false. Thus we see that

$$\sim(\mathbf{a}\cdot\mathbf{b}) = \sim\mathbf{a} \vee \sim\mathbf{b}.$$

From this equation and the equality of $\sim\sim\mathbf{a}$ with **a**, there is derived a remarkable feature of Boolean algebra, which has no counterpart in ordinary algebra. This characteristic is a duality according to which the exchange of the signs, $\cdot$ and $\vee$, in any equation of propositions transforms the equation into another one equally valid.[7] For example, exchanging the signs in this equation itself, we obtain

$$\sim(\mathbf{a} \vee \mathbf{b}) = \sim\mathbf{a}\cdot\sim\mathbf{b},$$

which is proved as follows:

$$\mathbf{a} \vee \mathbf{b} = \sim\sim\mathbf{a} \vee \sim\sim\mathbf{b} = \sim(\sim\mathbf{a}\cdot\sim\mathbf{b}).$$

Hence

$$\sim(\mathbf{a} \vee \mathbf{b}) = \sim\sim(\sim\mathbf{a}\cdot\sim\mathbf{b}) = \sim\mathbf{a}\cdot\sim\mathbf{b}.$$

From the duality in this instance and the mutual relation of **a** and $\sim\mathbf{a}$, the duality in other instances follows by symmetry. We have, accordingly, from the equations just preceding,

$$\mathbf{b} \vee \mathbf{a} = \mathbf{a} \vee \mathbf{b},$$

$$\mathbf{a} \vee \mathbf{a} = \mathbf{a}$$

and

$$(\mathbf{a} \vee \mathbf{b}) \vee \mathbf{c} = \mathbf{a} \vee (\mathbf{b} \vee \mathbf{c}) = \mathbf{a} \vee \mathbf{b} \vee \mathbf{c}.$$

The propositions $(\mathbf{a} \vee \mathbf{b})\cdot\mathbf{c}$ and $\mathbf{a} \vee (\mathbf{b}\cdot\mathbf{c})$ are not equal. For, if **a** is true and **c** false, the first of them is false but the second is

true. Therefore the form $\mathbf{a} \vee \mathbf{b}\cdot\mathbf{c}$ is ambiguous. In verbal expressions the ambiguity is usually prevented by the meaning of the words. Thus, in a weather forecast, "rain or snow and high winds," would be understood to mean "(rain or snow) and high winds," whereas "snow or rising temperature and rain" would mean "snow or (rising temperature and rain)." In symbolic expressions, on the other hand, the meaning is not given and parentheses are therefore necessary.

When we assert $(\mathbf{a} \vee \mathbf{b})\cdot\mathbf{c}$, we mean that at least one of the propositions, **a** and **b**, is true, but **c** is true in any case. This is the same as to say that at least one of the propositions, $\mathbf{a}\cdot\mathbf{c}$ and $\mathbf{b}\cdot\mathbf{c}$, is true and thus

$$(\mathbf{a} \vee \mathbf{b})\cdot\mathbf{c} = (\mathbf{a}\cdot\mathbf{c}) \vee (\mathbf{b}\cdot\mathbf{c}).$$

The dual of this equation is

$$(\mathbf{a}\cdot\mathbf{b}) \vee \mathbf{c} = (\mathbf{a} \vee \mathbf{c})\cdot(\mathbf{b} \vee \mathbf{c}).$$

If, in either of these equations, we let **c** be equal to **b** and substitute **b** for its equivalent, $\mathbf{b}\cdot\mathbf{b}$ in the first equation or $\mathbf{b} \vee \mathbf{b}$ in the second, we find that

$$(\mathbf{a} \vee \mathbf{b})\cdot\mathbf{b} = (\mathbf{a}\cdot\mathbf{b}) \vee \mathbf{b}.$$

In this equation, the exchange of the signs, $\cdot$ and $\vee$, has only the effect of transposing the members; the equation is dual to itself. Each of the propositions, $(\mathbf{a} \vee \mathbf{b})\cdot\mathbf{b}$ and $(\mathbf{a}\cdot\mathbf{b}) \vee \mathbf{b}$, is, indeed, equal simply to **b**. Thus to say, "He is a fool or a knave and he is a knave," or "He is a fool and a knave or he is a knave," sounds perhaps more uncharitable than to say simply, "He is a knave," but the meaning is the same.

In ordinary algebra, if the value of one quantity depends on the values of one or more other quantities, the first is called a function of the others. Similarly, in Boolean algebra, we may call a proposition a function of one or more other propositions if its meaning depends on theirs. For example, $\mathbf{a} \vee \mathbf{b}$ is a Boolean function of the propositions **a** and **b** as $a + b$ is an ordinary function of the quantities a and b.

It may be remarked that the operations of Boolean algebra generate functions of infinitely less variety than is found among the functions of ordinary algebra. In ordinary algebra, because $a \times a = a^2$, $a \times a^2 = a^3$, . . . and $a + a = 2a$, $a + 2a = 3a$, . . . , there is no end to the functions of a single variable which can be generated by repeated multiplications and additions. By contrast, in Boolean algebra, $\mathbf{a} \cdot \mathbf{a}$ and $\mathbf{a} \vee \mathbf{a}$ are both equal simply to **a**, and thus the signs, $\cdot$ and $\vee$, when used with a single proposition, generate no functions.

The only Boolean functions of a single proposition are itself and its contradictory. In form there are more; thus $\mathbf{a} \vee \sim\mathbf{a}$ has the form of a function of **a**, but it is a function only in the trivial sense in which $x - x$ and x/x are functions of x. In Boolean algebra, $\mathbf{a} \vee \sim\mathbf{a}$ plays the part of a constant proposition, because it is a truism and remains a truism through all changes in the meaning of **a**. To assert a truism in conjunction with a proposition is no more than to assert the proposition alone. Thus

$$(\mathbf{a} \vee \sim\mathbf{a}) \cdot \mathbf{b} = \mathbf{b}$$

for every meaning of **a** or **b**. On the other hand, to assert a truism in disjunction with a proposition is only to assert the truism; $\mathbf{a} \vee \sim\mathbf{a} \vee \mathbf{b}$, being true for every meaning of **a** or **b**, is itself a truism, so that

$$\mathbf{a} \vee \sim\mathbf{a} \vee \mathbf{b} = \mathbf{a} \vee \sim\mathbf{a}.$$

Each of these equations has its dual and thus

$$(\mathbf{a} \cdot \sim\mathbf{a}) \vee \mathbf{b} = \mathbf{b}$$

and

$$\mathbf{a} \cdot \sim\mathbf{a} \cdot \mathbf{b} = \mathbf{a} \cdot \sim\mathbf{a}.$$

The proposition $\mathbf{a} \cdot \sim\mathbf{a}$ is an absurdity for every meaning of **a** and is thus another constant proposition. These two constant propositions, the truism and the absurdity, are mutually contradictory.

It will be convenient for future reference to have the following collection of the equations of this chapter.

$$\sim\sim\mathbf{a} = \mathbf{a}, \qquad (2.1)$$

$\mathbf{a}\cdot\mathbf{a} = \mathbf{a}$, (2.2 I)	$\mathbf{a} \vee \mathbf{a} = \mathbf{a}$, (2.2 II)
$\mathbf{b}\cdot\mathbf{a} = \mathbf{a}\cdot\mathbf{b}$, (2.3 I)	$\mathbf{b} \vee \mathbf{a} = \mathbf{a} \vee \mathbf{b}$, (2.3 II)
$\sim(\mathbf{a}\cdot\mathbf{b}) = \sim\mathbf{a} \vee \sim\mathbf{b}$, (2.4 I)	$\sim(\mathbf{a} \vee \mathbf{b}) = \sim\mathbf{a}\cdot\sim\mathbf{b}$, (2.4 II)
$(\mathbf{a}\cdot\mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot(\mathbf{b}\cdot\mathbf{c}) = \mathbf{a}\cdot\mathbf{b}\cdot\mathbf{c}$, (2.5 I)	$(\mathbf{a} \vee \mathbf{b}) \vee \mathbf{c} = \mathbf{a} \vee (\mathbf{b} \vee \mathbf{c}) = \mathbf{a} \vee \mathbf{b} \vee \mathbf{c}$, (2.5 II)
$(\mathbf{a} \vee \mathbf{b})\cdot\mathbf{c} = (\mathbf{a}\cdot\mathbf{c}) \vee (\mathbf{b}\cdot\mathbf{c})$, (2.6 I)	$(\mathbf{a}\cdot\mathbf{b}) \vee \mathbf{c} = (\mathbf{a} \vee \mathbf{c})\cdot(\mathbf{b} \vee \mathbf{c})$, (2.6 II)
$(\mathbf{a} \vee \mathbf{b})\cdot\mathbf{b} = \mathbf{b}$, (2.7 I)	$(\mathbf{a}\cdot\mathbf{b}) \vee \mathbf{b} = \mathbf{b}$, (2.7 II)
$(\mathbf{a} \vee \sim\mathbf{a})\cdot\mathbf{b} = \mathbf{b}$, (2.8 I)	$(\mathbf{a}\cdot\sim\mathbf{a}) \vee \mathbf{b} = \mathbf{b}$, (2.8 II)
$\mathbf{a} \vee \sim\mathbf{a} \vee \mathbf{b} = \mathbf{a} \vee \sim\mathbf{a}$, (2.9 I)	$\mathbf{a}\cdot\sim\mathbf{a}\cdot\mathbf{b} = \mathbf{a}\cdot\sim\mathbf{a}$. (2.9 II)

Each of these equations after the first is dual to the equation on the same line in the other column, from which it can be obtained by the exchange of the signs, $\cdot$ and $\vee$. In the preceding discussion, the equations on the left were taken as axioms and those on the right were derived from them and the first equation. If, instead, the equations on the right had been taken as axioms, those on the left would have been their consequences. Indeed any set which includes the first equation and one from each pair on the same line will serve as axioms for the derivation of the others.

More equations can be derived from these by mathematical induction. For example, it can be shown, by an induction from Eq. (2.4 I), that

$$\sim(\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m) = \sim\mathbf{a}_1 \vee \sim\mathbf{a}_2 \vee\ldots\vee \sim\mathbf{a}_m, \qquad (2.10\text{ I})$$

where $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$ are any propositions.

We first assume provisionally, for the sake of the induction, that this equation holds when m is some number k and thence

prove that it holds also when m is $k + 1$ and consequently when it is any number greater than k.

Replacing $\mathbf{a}$ in Eq. (2.4 I) by $\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_k$ and $\mathbf{b}$ by $\mathbf{a}_{k+1}$, we have

$$\sim[(\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_k) \cdot \mathbf{a}_{k+1}] = \sim(\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_k) \vee \sim\mathbf{a}_{k+1}.$$

By the provisional assumption just made,

$$\sim(\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_k) = \sim\mathbf{a}_1 \vee \sim\mathbf{a}_2 \vee \ldots \vee \sim\mathbf{a}_k,$$

and thus

$$\sim[(\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_k) \cdot \mathbf{a}_{k+1}] = (\sim\mathbf{a}_1 \vee \sim\mathbf{a}_2 \vee \ldots \vee \sim\mathbf{a}_k) \vee \sim\mathbf{a}_{k+1}.$$

Therefore, by Eqs. (2.5 I) and (2.5 II)

$$\sim(\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_k \cdot \mathbf{a}_{k+1}) = \sim\mathbf{a}_1 \vee \sim\mathbf{a}_2 \vee \ldots \vee \sim\mathbf{a}_k \vee \sim\mathbf{a}_{k+1}.$$

Thus Eq. (2.10 I) is proved when m is $k + 1$ if it is true when m is k. By Eq. (2.4 I), it is true when m is 2. Hence it is proved when m is 3 and thence when m is 4 and when it is any number, however great.

By exchanging the signs, $\cdot$ and $\vee$, in Eq. (2.10 I), we obtain its dual, also valid:

$$\sim(\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m) = \sim\mathbf{a}_1 \cdot \sim\mathbf{a}_2 \cdot \ldots \cdot \sim\mathbf{a}_m, \quad (2.10\text{ II})$$

an equation which can also be derived by mathematical induction from Eq. (2.4 II).

A mathematical induction from Eq. (2.6 I) gives:

$$(\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m) \cdot \mathbf{b} = (\mathbf{a}_1 \cdot \mathbf{b}) \vee (\mathbf{a}_2 \cdot \mathbf{b}) \vee \ldots \vee (\mathbf{a}_m \cdot \mathbf{b}). \quad (2.11\text{ I})$$

By an exchange of signs in this equation or an induction from Eq. (2.6 II), we obtain

$$(\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_m) \vee \mathbf{b} = (\mathbf{a}_1 \vee \mathbf{b}) \cdot (\mathbf{a}_2 \vee \mathbf{b}) \cdot \ldots \cdot (\mathbf{a}_m \vee \mathbf{b}). \quad (2.11\text{ II})$$

3. The Conjunctive Inference

Every conjecture is based on some hypothesis, which may consist wholly of actual evidence or may include assumptions made for the argument's sake. Let **h** denote an hypothesis and **i** a proposition reasonably entitled to partial assent as an inference from it. The probability is a measure of this assent, determined, more or less precisely, by the two propositions, **i** and **h**. It is therefore a numerical function of propositions, in contrast with the functions considered in the preceding chapter, which, being themselves propositions, may be called propositional functions of propositions. (Readers familiar with vector analysis may be reminded of the distinction between scalar and vector functions of vectors.)[8]

Let us denote the probability of the inference **i** on the hypothesis **h** by the symbol $\mathbf{i} \mid \mathbf{h}$, which will be enclosed in parentheses when it is a term or factor in a more complicated expression.[9] The choice of a scale on which probabilities are to be reckoned is still undecided at this stage of our consideration. If $\mathbf{i} \mid \mathbf{h}$ is a measure of the assent to which the inference **i** is reasonably entitled on the hypothesis **h**, it meets all the requirements of a probability which our discussion thus far has imposed. But, if $\mathbf{i} \mid \mathbf{h}$ is such a measure, then so also is an arbitrary function of $\mathbf{i} \mid \mathbf{h}$, such as $100\ (\mathbf{i} \mid \mathbf{h})$, $(\mathbf{i} \mid \mathbf{h})^2$ or $\ln (\mathbf{i} \mid \mathbf{h})$. The choice among the different possible scales of probability is made by conventions which will be considered later.

The probability on the hypothesis **h** of the inference formed by conjoining the two inferences **i** and **j** is represented, in the notation just given, by $\mathbf{i} \cdot \mathbf{j} \mid \mathbf{h}$. By the axiom (1.ii), this probability is a function of the two probabilities: $\mathbf{i} \mid \mathbf{h}$, the probability of the first inference on the original hypothesis, and $\mathbf{j} \mid \mathbf{h} \cdot \mathbf{i}$, the probability of the second inference on the hypothesis formed by conjoining the original hypothesis with the first inference. Calling this function F, we have:

$$\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h} = F[(\mathbf{i} \mid \mathbf{h}), (\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i})]. \tag{3.1}$$

Since the probabilities are all numbers, F is a numerical function of two numerical variables.

The form of the function F is in part arbitrary, but it can not be entirely so, because the equation must be consistent with Boolean algebra. Let us see what restriction is placed on the form of F by the Boolean equation

$$(\mathbf{a}\cdot\mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot(\mathbf{b}\cdot\mathbf{c}) = \mathbf{a}\cdot\mathbf{b}\cdot\mathbf{c}.$$

If we let

$$\mathbf{h} = \mathbf{a},\ \mathbf{i} = \mathbf{b},\ \mathbf{j} = \mathbf{c}\cdot\mathbf{d},$$

so that

$$\mathbf{i}\cdot\mathbf{j} = \mathbf{b}\cdot(\mathbf{c}\cdot\mathbf{d}) = \mathbf{b}\cdot\mathbf{c}\cdot\mathbf{d},$$

Eq. (3.1) becomes

$$\mathbf{b}\cdot\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a} = F[(\mathbf{b} \mid \mathbf{a}), (\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a}\cdot\mathbf{b})] = F[x, (\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a}\cdot\mathbf{b})],$$

where, for brevity, x has been written for $\mathbf{b} \mid \mathbf{a}$. Also, if we now let

$$\mathbf{h} = \mathbf{a}\cdot\mathbf{b},\ \mathbf{i} = \mathbf{c},\ \mathbf{j} = \mathbf{d},$$

so that

$$\mathbf{h}\cdot\mathbf{i} = (\mathbf{a}\cdot\mathbf{b})\cdot\mathbf{c} = \mathbf{a}\cdot\mathbf{b}\cdot\mathbf{c},$$

Eq. (3.1) becomes

$$\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a}\cdot\mathbf{b} = F[(\mathbf{c} \mid \mathbf{a}\cdot\mathbf{b}), (\mathbf{d} \mid \mathbf{a}\cdot\mathbf{b}\cdot\mathbf{c})] = F(y, z),$$

where y has been written for $\mathbf{c} \mid \mathbf{a}\cdot\mathbf{b}$ and z for $\mathbf{d} \mid \mathbf{a}\cdot\mathbf{b}\cdot\mathbf{c}$. Hence, by substitution in the expression just obtained for $\mathbf{b}\cdot\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a}$, we find,

$$\mathbf{b}\cdot\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a} = F[x, F(y, z)]. \tag{3.2}$$

Similarly, if, in Eq. (3.1), we let

$$\mathbf{h} = \mathbf{a},\ \mathbf{i} = \mathbf{b}\cdot\mathbf{c},\ \mathbf{j} = \mathbf{d},$$

we find

$$\mathbf{b}\cdot\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a} = F[(\mathbf{b}\cdot\mathbf{c} \mid \mathbf{a}), z],$$

and, if we now let

$$\mathbf{h} = \mathbf{a},\ \mathbf{i} = \mathbf{b},\ \mathbf{j} = \mathbf{c},$$

we have

$$\mathbf{b}\cdot\mathbf{c} \mid \mathbf{a} = F(x, y),$$

so that

$$\mathbf{b}\cdot\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a} = F[F(x, y), z].$$

Equating this expression for $\mathbf{b}\cdot\mathbf{c}\cdot\mathbf{d} \mid \mathbf{a}$ with that given by Eq. (3.2), we have

$$F[x, F(y, z)] = F[F(x, y), z], \tag{3.3}$$

as a functional equation to be satisfied by the function F.[10]

Let F be assumed differentiable and let $\partial F(u, v)/\partial u$ be denoted by $F_1(u, v)$ and $\partial F(u, v)/\partial v$ by $F_2(u, v)$. Then, by differentiating this equation with respect to x and y, we obtain the two equations,

$$F_1[x, F(y, z)] = F_1[F(x, y), z]F_1(x, y),$$

$$F_2[x, F(y, z)]F_1(y, z) = F_1[F(x, y), z]F_2(x, y).$$

Eliminating $F_1[F(x, y), z]$ between these equations gives a result which may be written in either of the two forms:

$$G[x, F(y, z)]F_1(y, z) = G(x, y), \tag{3.4}$$

$$G[x, F(y, z)]F_2(y, z) = G(x, y)G(y, z), \tag{3.5}$$

where $G(u, v)$ denotes $F_2(u, v)/F_1(u, v)$.

Differentiating the first of these equations with respect to z and the second with respect to y, we obtain equal expressions on the left and so find

$$\partial[G(x, y)G(y, z)]/\partial y = 0.$$

Thus G must be such a function as not to involve y in the product $G(x, y)G(y, z)$. The most general function which satisfies this restriction is given by

$$G(u, v) = aH(u)/H(v),$$

where a is an arbitrary constant and H is an arbitrary function of a single variable.

Substituting this expression for G in Eqs. (3.4) and (3.5), we obtain

$$F_1(y, z) = H[F(y, z)]/H(y),$$

$$F_2(y, z) = aH[F(y, z)]/H(z).$$

Therefore, since $dF(y, z) = F_1(y, z)\, dy + F_2(y, z)\, dz$, we find

$$\frac{dF(y, z)}{H[F(y, z)]} = \frac{dy}{H(y)} + a\frac{dz}{H(z)}.$$

Integrating, we obtain

$$CP[F(y, z)] = P(y)[P(z)]^a, \tag{3.6}$$

where C is a constant of integration and P is a function of a single variable, defined by the equation,

$$\ln P(u) = \int \frac{du}{H(u)}.$$

Because H is an arbitrary function, so also is P.

Equation (3.6) holds for arbitrary values of y and z and hence for arbitrary variables of which P and F may be functions. If we take the function P of both members of Eq. (3.3), we obtain an equation from which F may be eliminated by successive substitutions of $P(F)$ as given by Eq. (3.6). The result is to show that $a = 1$. Thus Eq. (3.6) becomes

$$CP[F(y, z)] = P(y)P(z).$$

If, in this equation, we let y be $\mathbf{i} \mid \mathbf{h}$ and z be $\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i}$, then, by Eq. (3.1), $F(y, z) = \mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}$. Thus

$$CP(\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}) = P(\mathbf{i} \mid \mathbf{h})P(\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i}).$$

The function P, being arbitrary, may be given any convenient form. Indeed, if we so choose, we may leave its form undetermined for, as was remarked earlier in this chapter, if $\mathbf{i} \mid \mathbf{h}$ measures probability, so also does an arbitrary function of $\mathbf{i} \mid \mathbf{h}$. We could give the name of probability to $P(\mathbf{i} \mid \mathbf{h})$ rather than to $\mathbf{i} \mid \mathbf{h}$ and never be concerned with the relation between the two quantities, because we should never have occasion to use $\mathbf{i} \mid \mathbf{h}$ except in the function $P(\mathbf{i} \mid \mathbf{h})$. In effect we should merely be adopting a different symbol of probability. Instead, let us retain the symbol $\mathbf{i} \mid \mathbf{h}$ and take advantage of the arbitrariness of the function P to let $P(u)$ be identical with u, so that the equation may be written

$$C(\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}) = (\mathbf{i} \mid \mathbf{h})(\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i}).$$

If, in this equation, we let $\mathbf{j} = \mathbf{i}$ and note that $\mathbf{i}\cdot\mathbf{i} = \mathbf{i}$ by Eq. (2.2 I), we obtain, after dividing by $(\mathbf{i} \mid \mathbf{h})$,

$$C = \mathbf{i} \mid \mathbf{h}\cdot\mathbf{i}.$$

Thus, when the hypothesis includes the inference in a conjunction, the probability has the constant value C, whatever the propositions may be. This is what we should expect, because an inference is certain on any hypothesis in which it is conjoined and we do not recognize degrees of certainty.

The value to be assigned to C is purely a matter of convenience, and different values may be assigned in different discourses. When we use the phrase, "three chances in ten," we are, in effect, adopting a scale of probability on which certainty is represented by 10 and we are saying that some other probability has the value 3 on this scale. Similarly, if we say that an inference is "95 per cent certain," we are saying that its probability is 95 on a scale on which certainty has the probability 100. Usually it is convenient to represent certainty by 1 and, with this convention, the equation for the probability of the conjunctive inference is

$$\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h} = (\mathbf{i} \mid \mathbf{h})(\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i}). \tag{3.7}$$

This equation expresses the familiar rule for the probability of a conjunctive inference or, as it is more often stated, the probability of a compound event. It is indeed the only equation for this probability which is consistent with the ordinary scale. It is worth remarking, however, that other scales beside the ordinary one are consistent with this equation. For, raising its members to a power r, we have

$$(\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h})^r = (\mathbf{i} \mid \mathbf{h})^r(\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i})^r, \tag{3.8}$$

whence it is evident that the r th powers of the ordinary probabilities satisfy the same equation as the ordinary probabilities themselves. It follows that the rule for the probability of the conjunctive inference would remain the same in any change by which arbitrary powers of the ordinary probabilities were used, instead of them, as probabilities on a new scale.

Equation (3.7), when $\mathbf{i}$ is the truism, $\mathbf{a} \vee \sim\mathbf{a}$, becomes

$$(\mathbf{a} \vee \sim\mathbf{a})\cdot\mathbf{j} \mid \mathbf{h} = (\mathbf{a} \vee \sim\mathbf{a} \mid \mathbf{h})[\mathbf{j} \mid \mathbf{h}\cdot(\mathbf{a} \vee \sim\mathbf{a})].$$

By Eq. (2.8 I), $(\mathbf{a} \vee \sim\mathbf{a})\cdot\mathbf{j} = \mathbf{j}$ and similarly $\mathbf{h}\cdot(\mathbf{a} \vee \sim\mathbf{a}) = \mathbf{h}$. Hence each of the probabilities, $(\mathbf{a} \vee \sim\mathbf{a})\cdot\mathbf{j} \mid \mathbf{h}$ and $\mathbf{j} \mid \mathbf{h}\cdot(\mathbf{a} \vee \sim\mathbf{a})$, is equal simply to $\mathbf{j} \mid \mathbf{h}$ and

$$\mathbf{a} \vee \sim\mathbf{a} \mid \mathbf{h} = 1.$$

The truism, as we should suppose, is thus certain on every hypothesis.

It is to be understood that the absurdity, $\mathbf{a}\cdot\sim\mathbf{a}$, is excluded as an hypothesis but, at the same time, it should be stressed that not every false hypothesis is thus excluded. A proposition is false if it contradicts a fact but absurd only if it contradicts itself. It is permissible logically and often worth while to consider the probability of an inference on an hypothesis which is contrary to fact in one respect or another.

An hypothesis $\mathbf{h}$, on which an inference $\mathbf{i}$ is certain, is said to *imply* the inference. Every hypothesis, for example, thus implies the truism. There are some discourses in which a proposition

h is common to the hypotheses of all the probabilities considered, while other propositions, **a**, **b**, . . . , are conjoined with **h** in some of the hypotheses. In such a discourse it is sometimes convenient, and need not be confusing, to omit reference to **h** and call an inference "implied by **a**" if it is implied by **a·h**. In this sense, an inference which is certain on the hypothesis **h** alone, and therefore certain throughout the discourse, can be said to be implied by each of the propositions, **a**, **b**, . . . , as the truism is implied by every proposition in any discourse.

Exchanging **i** and **j** in Eq. (3.7) and observing that $\mathbf{j \cdot i} = \mathbf{i \cdot j}$ by Eq. (2.3 I), we see that

$$(\mathbf{i} \mid \mathbf{h})(\mathbf{j} \mid \mathbf{h \cdot i}) = (\mathbf{j} \mid \mathbf{h})(\mathbf{i} \mid \mathbf{h \cdot j}),$$

whence

$$\frac{\mathbf{j} \mid \mathbf{h \cdot i}}{\mathbf{j} \mid \mathbf{h}} = \frac{\mathbf{i} \mid \mathbf{h \cdot j}}{\mathbf{i} \mid \mathbf{h}}.$$

If $\mathbf{j} \mid \mathbf{h \cdot i} = \mathbf{j} \mid \mathbf{h}$, **i** is said to be *irrelevant* to **j** on the hypothesis **h**. The equation just obtained shows that also **j** is then irrelevant to **i** on the same hypothesis. The relation is therefore one of *mutual irrelevance* between the propositions **i** and **j** on the hypothesis **h**, and it is conveniently defined by the condition,

$$\mathbf{i \cdot j} \mid \mathbf{h} = (\mathbf{i} \mid \mathbf{h})(\mathbf{j} \mid \mathbf{h}). \tag{3.9}$$

If **h** alone implies **j**, so also does **h·i**. Then $\mathbf{j} \mid \mathbf{h}$ and $\mathbf{j} \mid \mathbf{h \cdot i}$ are both unity and therefore equal, and **i** and **j** are mutually irrelevant. Thus every proposition implied by a given hypothesis is irrelevant on that hypothesis to every other proposition.

4. The Contradictory Inference

By the axiom (1.i), the probability of the inference **i** on the hypothesis **h** determines that of the contradictory inference, $\sim\mathbf{i}$, on the same hypothesis. Thus

$$\sim \mathbf{i} \mid \mathbf{h} = f(\mathbf{i} \mid \mathbf{h}), \tag{4.1}$$

where f is a numerical function of a single variable, which must be consistent in form both with Boolean algebra and the rule for the probability of the conjunctive inference, as given in Eq. (3.7).

To see what are the requirements of this consistency, first let $\mathbf{i} = \sim\mathbf{j}$ in the equation. Thus we find

$$\sim\sim\mathbf{j} \mid \mathbf{h} = f(\sim\mathbf{j} \mid \mathbf{h}) = f[f(\mathbf{j} \mid \mathbf{h})].$$

But $\sim\sim\mathbf{j} = \mathbf{j}$ by Eq. (2.1) and thus

$$\mathbf{j} \mid \mathbf{h} = f[f(\mathbf{j} \mid \mathbf{h})].$$

Therefore f must be such a function that

$$f[f(x)] = x. \tag{4.2}$$

This equation, by itself, imposes only a rather weak restriction on the form of f. A more stringent condition is found if we replace $\mathbf{i}$ in Eq. (4.1) by $\mathbf{i} \vee \mathbf{j}$ and thus obtain, by the use of Eq. (2.4 II),

$$f(\mathbf{i} \vee \mathbf{j} \mid \mathbf{h}) = \sim(\mathbf{i} \vee \mathbf{j}) \mid \mathbf{h} = \sim\mathbf{i} \cdot \sim\mathbf{j} \mid \mathbf{h}.$$

By Eqs. (3.7) and (4.1),

$$\sim\mathbf{i} \cdot \sim\mathbf{j} \mid \mathbf{h} = (\sim\mathbf{i} \mid \mathbf{h})(\sim\mathbf{j} \mid \mathbf{h} \cdot \sim\mathbf{i}) = f(\mathbf{i} \mid \mathbf{h}) f(\mathbf{j} \mid \mathbf{h} \cdot \sim\mathbf{i}).$$

Thus

$$f(\mathbf{j} \mid \mathbf{h} \cdot \sim\mathbf{i}) = \frac{f(\mathbf{i} \vee \mathbf{j} \mid \mathbf{h})}{f(\mathbf{i} \mid \mathbf{h})}.$$

Taking the function f of both members of this equation and using Eq. (4.2), on the left, we have

$$\mathbf{j} \mid \mathbf{h} \cdot \sim\mathbf{i} = f\left[\frac{f(\mathbf{i} \vee \mathbf{j} \mid \mathbf{h})}{f(\mathbf{i} \mid \mathbf{h})}\right].$$

Making use again of Eq. (3.7), we find that

$$\mathbf{j} \mid \mathbf{h} \cdot \sim\mathbf{i} = \frac{\sim\mathbf{i} \cdot \mathbf{j} \mid \mathbf{h}}{\sim\mathbf{i} \mid \mathbf{h}} = \frac{\sim\mathbf{i} \cdot \mathbf{j} \mid \mathbf{h}}{f(\mathbf{i} \mid \mathbf{h})},$$

whence, by the preceding equation,

$$\sim\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h} = f(\mathbf{i} \mid \mathbf{h}) f\left[\frac{f(\mathbf{i} \vee \mathbf{j} \mid \mathbf{h})}{f(\mathbf{i} \mid \mathbf{h})}\right].$$

By Eqs. (2.3 I), (3.7) and (4.1),

$$\sim\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h} = \mathbf{j}\cdot\sim\mathbf{i} \mid \mathbf{h} = (\mathbf{j} \mid \mathbf{h})(\sim\mathbf{i} \mid \mathbf{h}\cdot\mathbf{j}) = (\mathbf{j} \mid \mathbf{h}) f(\mathbf{i} \mid \mathbf{h}\cdot\mathbf{j})$$

$$= (\mathbf{j} \mid \mathbf{h}) f\left(\frac{\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}}{\mathbf{j} \mid \mathbf{h}}\right).$$

With this result the preceding equation becomes

$$(\mathbf{j} \mid \mathbf{h}) f\left(\frac{\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}}{\mathbf{j} \mid \mathbf{h}}\right) = f(\mathbf{i} \mid \mathbf{h}) f\left[\frac{f(\mathbf{i} \vee \mathbf{j} \mid \mathbf{h})}{f(\mathbf{i} \mid \mathbf{h})}\right]. \tag{4.3}$$

This equation holds for arbitrary meanings of **i** and **j**. Let

$$\mathbf{i} = \mathbf{a}\cdot\mathbf{b}, \quad \mathbf{j} = \mathbf{a} \vee \mathbf{b},$$

so that

$$\mathbf{i}\cdot\mathbf{j} = (\mathbf{a}\cdot\mathbf{b})\cdot(\mathbf{a} \vee \mathbf{b}) = \mathbf{a}\cdot[\mathbf{b}\cdot(\mathbf{a} \vee \mathbf{b})] \quad \text{by Eq. (2.5 I)}$$

$$= \mathbf{a}\cdot\mathbf{b} \text{ by Eqs. (2.3 I) and (2.7 I)} = \mathbf{i},$$

and, by a similar argument resting on Eqs. (2.5 II), (2.3 I) and (2.7 II),

$$\mathbf{i} \vee \mathbf{j} = \mathbf{j}.$$

Thus Eq. (4.3) becomes

$$(\mathbf{j} \mid \mathbf{h}) f\left(\frac{\mathbf{i} \mid \mathbf{h}}{\mathbf{j} \mid \mathbf{h}}\right) = f(\mathbf{i} \mid \mathbf{h}) f\left[\frac{f(\mathbf{j} \mid \mathbf{h})}{f(\mathbf{i} \mid \mathbf{h})}\right].$$

This equation is given in a more concise and symmetrical form if we denote $\mathbf{i} \mid \mathbf{h}$ by $f(y)$, so that $f(\mathbf{i} \mid \mathbf{h}) = y$, and $\mathbf{j} \mid \mathbf{h}$ by z. In this way we obtain the equation,

$$zf\left[\frac{f(y)}{z}\right] = yf\left[\frac{f(z)}{y}\right]. \tag{4.4}$$

This equation and the three derived from it by differentiation with respect to y, to z and to y and z can be written

$$\begin{aligned} zf(u) &= yf(v), \\ f'(u)f'(y) &= f(v) - vf'(v), \\ f(u) - uf'(u) &= f'(v)f'(z), \\ uf''(u)f'(y)/z &= vf''(v)f'(z)/y, \end{aligned}$$

where u denotes $f(y)/z$, v denotes $f(z)/y$, f' the first derivative of f and f'' the second derivative.

Multiplying together the corresponding members of the first and last of these equations, we eliminate y and z at the same time, obtaining

$$uf''(u)f(u)f'(y) = vf''(v)f(v)f'(z).$$

With this equation and the second and third of the preceding group, it is possible to eliminate $f'(y)$ and $f'(z)$. The resulting equation is

$$\frac{uf''(u)\,f(u)}{[uf'(u) - f(u)]\,f'(u)} = \frac{vf''(v)\,f(v)}{[vf'(v) - f(v)]\,f'(v)}.$$

Each member of this equation is the same function of a different variable and the two variables, u and v, are mutually independent. This function of an arbitrary variable x must therefore be equal to a constant. Calling this constant c, we have

$$xf''(x)f(x) = c[xf'(x) - f(x)]f'(x).$$

This equation may be put in the form

$$df'/f' = c(df/f - dx/x),$$

whence, by integration, we find that

$$f' = A(f/x)^c,$$

where A is a constant. The variables being separable, another integration gives

$$f^r = Ax^r + B,$$

where r has been written for $1 - c$, and B is another constant. It is now found by substitution that this result satisfies Eq. (4.4) for arbitrary values of y and z only if $B = A^2$. Equation (4.2) is also to be satisfied and for this it is necessary that $A = -1$. No restriction is imposed on r, which thus remains arbitrary. We have then finally

$$x^r + [f(x)]^r = 1$$

or

$$(\mathbf{i} \mid \mathbf{h})^r + (\sim\mathbf{i} \mid \mathbf{h})^r = 1.$$

We might, if we wished, leave the value of r unspecified by using $(\mathbf{i} \mid \mathbf{h})^r$ as the symbol of probability here and in Eq. (3.8). With a free choice in the matter, it is more convenient to take r as unity. By this convention,

$$(\mathbf{i} \mid \mathbf{h}) + (\sim\mathbf{i} \mid \mathbf{h}) = 1. \tag{4.5}$$

If, in this equation, we replace $\mathbf{h}$ by $\mathbf{h}\cdot\mathbf{i}$ and recall that $\mathbf{i} \mid \mathbf{h}\cdot\mathbf{i} = 1$, we see that $\sim\mathbf{i} \mid \mathbf{h}\cdot\mathbf{i} = 0$. Thus impossibility has the fixed probability zero as certainty has the fixed probability unity.

A theorem frequently useful is obtained as follows. By Eq. (3.7),

$$(\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}) + (\mathbf{i}\cdot\sim\mathbf{j} \mid \mathbf{h}) = (\mathbf{i} \mid \mathbf{h})[(\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i}) + (\sim\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i})],$$

whence, by Eq. (4.5),

$$(\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}) + (\mathbf{i}\cdot\sim\mathbf{j} \mid \mathbf{h}) = \mathbf{i} \mid \mathbf{h}. \tag{4.6}$$

An immediate consequence of this theorem, obtained by making $\mathbf{j}$ equal to $\mathbf{i}$ and noting that $\mathbf{i}\cdot\mathbf{i} = \mathbf{i}$, is

$$\mathbf{i}\cdot\sim\mathbf{i} \mid \mathbf{h} = 0.$$

Thus the absurdity, $\mathbf{i}\cdot\sim\mathbf{i}$, has zero probability on every hypothesis, as we should expect. There would be an inconsistency here if the absurdity itself were admitted as an hypothesis, for then it

would appear to be certain as an inference and to have unit probability. There is, of course, nothing astonishing about this, because an inconsistency is just what we should expect as the logical consequence of a self-contradictory hypothesis.

Only the absurdity is impossible on every hypothesis, but every proposition except the truism is impossible on some hypotheses. If each of the two propositions, **i** and **j**, is possible without the other on the hypothesis **h**, but their conjunction, **i·j**, is impossible, it follows from Eq. (3.7) directly that

$$\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i} = 0$$

and, by the exchange of **i** and **j**, that

$$\mathbf{i} \mid \mathbf{h}\cdot\mathbf{j} = 0.$$

The propositions **i** and **j** are said in this case to be *mutually exclusive* on the hypothesis **h**, because the conjunction of either of them with **h** in the hypothesis makes the other impossible.

If **i** is impossible on the hypothesis **h** alone, **h·i** is self-contradictory and therefore inadmissible as an hypothesis. In this case, therefore, no meaning can be attached to $\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i}$. But $\mathbf{i} \mid \mathbf{h}\cdot\mathbf{j}$ has still a meaning and the value zero, unless **j** is also impossible on the hypothesis **h** alone, and, in any case, $\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h} = 0$. If both $\mathbf{i} \mid \mathbf{h}$ and $\mathbf{j} \mid \mathbf{h}$ are zero, then both $\mathbf{j} \mid \mathbf{h}\cdot\mathbf{i}$ and $\mathbf{i} \mid \mathbf{h}\cdot\mathbf{j}$ are meaningless, but, *a fortiori*, $\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h} = 0$. It is convenient to comprise all these cases under a common term and call any two propositions mutually exclusive on a given hypothesis if their conjunction is impossible on that hypothesis, whether they are singly so or not. In this sense, any proposition which is impossible on an hypothesis is mutually exclusive on that hypothesis with every proposition, including even itself, and the absurdity is mutually exclusive with every proposition on every possible hypothesis.

It is worth remarking that, if two propositions are mutually irrelevant on a given hypothesis, then each is irrelevant to the contradictory of the other and the contradictories of both are mutually irrelevant. To see this, let **i** and **j** be propositions

mutually irrelevant on the hypothesis **h**, so that $\mathbf{i} \mid \mathbf{h}\cdot\mathbf{j} = \mathbf{i} \mid \mathbf{h}$. Then, by Eq. (4.5), $\sim\mathbf{i} \mid \mathbf{h}\cdot\mathbf{j} = \sim\mathbf{i} \mid \mathbf{h}$ and **j** is thus irrelevant to $\sim\mathbf{i}$. Exchanging the propositions proves that **i** is irrelevant to $\sim\mathbf{j}$ and repeating the argument proves the mutual irrelevance of $\sim\mathbf{i}$ and $\sim\mathbf{j}$. Every instance of irrelevance is thus a relation between pairs of propositions, such as **i**, $\sim\mathbf{i}$ and **j**, $\sim\mathbf{j}$, each proposition of either pair being irrelevant to each of the other pair.

5. *The Disjunctive Inference*

The two axioms which, in the two chapters preceding this one, have been found sufficient for the probabilities of the conjunctive and contradictory inferences, suffice also for the probability of the disjunctive inference. That only two axioms are required is a consequence of the fact that, among the three operations: contradiction, conjunction and disjunction, there are only two independent ones: contradiction and either of the others but not both. For the Boolean equations, $\sim(\mathbf{i} \vee \mathbf{j}) = \sim\mathbf{i}\cdot\sim\mathbf{j}$ and $\sim\sim(\mathbf{i} \vee \mathbf{j}) = \mathbf{i} \vee \mathbf{j}$, can be combined to give

$$\mathbf{i} \vee \mathbf{j} = \sim(\sim\mathbf{i}\cdot\sim\mathbf{j}),$$

an equation which defines disjunction in terms of contradiction and conjunction. Alternatively, conjunction can be defined in terms of contradiction and disjunction.

By Eq. (4.5), therefore,

$$\mathbf{i} \vee \mathbf{j} \mid \mathbf{h} = 1 - (\sim\mathbf{i}\cdot\sim\mathbf{j} \mid \mathbf{h})$$

and, by Eq. (4.6),

$$\sim\mathbf{i}\cdot\sim\mathbf{j} \mid \mathbf{h} = (\sim\mathbf{i} \mid \mathbf{h}) - (\sim\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}) = 1 - (\mathbf{i} \mid \mathbf{h}) - (\sim\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}).$$

Thus

$$\mathbf{i} \vee \mathbf{j} \mid \mathbf{h} = (\mathbf{i} \mid \mathbf{h}) + (\sim\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}).$$

By Eqs. (2.3 I) and (4.6),

$$\sim\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h} = \mathbf{j}\cdot\sim\mathbf{i} \mid \mathbf{h} = (\mathbf{j} \mid \mathbf{h}) - (\mathbf{j}\cdot\mathbf{i} \mid \mathbf{h}) = (\mathbf{j} \mid \mathbf{h}) - (\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}).$$

Therefore

$$\mathbf{i} \vee \mathbf{j} \mid \mathbf{h} = (\mathbf{i} \mid \mathbf{h}) + (\mathbf{j} \mid \mathbf{h}) - (\mathbf{i}\cdot\mathbf{j} \mid \mathbf{h}). \tag{5.1}$$

It is worth noticing that the exchange of the signs, $\vee$ and $\cdot$, in this equation has only the effect of transposing terms and so leaves the equation unchanged in meaning and therefore still valid.

This equation, rewritten with a change of notation whereby $\mathbf{i}$ and $\mathbf{j}$ are replaced by $\mathbf{a}_1$ and $\mathbf{a}_2$, becomes

$$\mathbf{a}_1 \vee \mathbf{a}_2 \mid \mathbf{h} = (\mathbf{a}_1 \mid \mathbf{h}) + (\mathbf{a}_2 \mid \mathbf{h}) - (\mathbf{a}_1\cdot\mathbf{a}_2 \mid \mathbf{h}). \tag{5.2}$$

In this form, it is a special case of the general equation, now to be proved, for the probability of the disjunction of m propositions. This is

$$(\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m \mid \mathbf{h}) = \sum_{i=1}^{m} (\mathbf{a}_i \mid \mathbf{h}) - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) + \sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \sum_{k=j+1}^{m} (\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{a}_k \mid \mathbf{h}) - \ldots \pm (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}). \tag{5.3}$$

The limits of the summations in this equation are such that none of the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, is conjoined with itself in any inference and also that no two inferences in any summation are conjunctions of the same propositions in different order. In the three-fold summation, for example, there is no such term as $\mathbf{a}_1\cdot\mathbf{a}_1\cdot\mathbf{a}_2 \mid \mathbf{h}$, and the only conjunction of $\mathbf{a}_1$, $\mathbf{a}_2$, and $\mathbf{a}_3$ is in the term $\mathbf{a}_1\cdot\mathbf{a}_2\cdot\mathbf{a}_3 \mid \mathbf{h}$, because the limits exclude probabilities such as $\mathbf{a}_2\cdot\mathbf{a}_1\cdot\mathbf{a}_3 \mid \mathbf{h}$, obtained from this one by permuting the propositions. For the m-fold summation, therefore, there is only one possible order of the m propositions and the summation is reduced to a single term. Its sign is positive if m is odd and negative if m is even.

The proof of the equation is by mathematical induction and consists in showing that it holds for the disjunction of $m + 1$ propositions if it holds for the disjunction of m. If we let $\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m$ be $\mathbf{i}$ in Eq. (5.1) and $\mathbf{a}_{m+1}$ be $\mathbf{j}$, we have

$$\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_{m+1} \mid \mathbf{h} = (\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m \mid \mathbf{h}) + (\mathbf{a}_{m+1} \mid \mathbf{h}) - [(\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m) \cdot \mathbf{a}_{m+1} \mid \mathbf{h}].$$

By letting $\mathbf{b}$ be $\mathbf{a}_{m+1}$ in Eq. (2.11 I), we see that

$$(\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m) \cdot \mathbf{a}_{m+1} = (\mathbf{a}_1 \cdot \mathbf{a}_{m+1}) \vee (\mathbf{a}_2 \cdot \mathbf{a}_{m+1}) \vee \ldots \vee (\mathbf{a}_m \cdot \mathbf{a}_{m+1})$$

and hence

$$\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_{m+1} \mid \mathbf{h} = (\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m \mid \mathbf{h}) + (\mathbf{a}_{m+1} \mid \mathbf{h}) - [(\mathbf{a}_1 \cdot \mathbf{a}_{m+1}) \vee (\mathbf{a}_2 \cdot \mathbf{a}_{m+1}) \vee \ldots \vee (\mathbf{a}_m \cdot \mathbf{a}_{m+1}) \mid \mathbf{h}]. \quad (5.4)$$

Of the three probabilities now on the right, both the first and the third are those of disjunctions of m propositions, for which we assume, for the sake of the mathematical induction, that Eq. (5.3) is valid. For the first of these probabilities, Eq. (5.3) gives an expression which can be substituted without change in Eq. (5.4). The expression to be substituted for the other is obtained by replacing $\mathbf{a}_1$ in Eq. (5.3) by $\mathbf{a}_1 \cdot \mathbf{a}_{m+1}$, $\mathbf{a}_2$ by $\mathbf{a}_2 \cdot \mathbf{a}_{m+1}$, $\ldots$ $\mathbf{a}_m$ by $\mathbf{a}_m \cdot \mathbf{a}_{m+1}$. This expression, with the simplification allowed by the equality of $\mathbf{a}_{m+1}$ and $\mathbf{a}_{m+1} \cdot \mathbf{a}_{m+1}$, is given by the equation,

$$(\mathbf{a}_1 \cdot \mathbf{a}_{m+1}) \vee (\mathbf{a}_2 \cdot \mathbf{a}_{m+1}) \vee \ldots \vee (\mathbf{a}_m \cdot \mathbf{a}_{m+1}) \mid \mathbf{h} = \sum_{i=1}^{m} (\mathbf{a}_i \cdot \mathbf{a}_{m+1} \mid \mathbf{h}) - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (\mathbf{a}_i \cdot \mathbf{a}_j \cdot \mathbf{a}_{m+1} \mid \mathbf{h}) + \ldots \pm (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_{m+1} \mid \mathbf{h}). \quad (5.5)$$

By making the substitutions just described in Eq. (5.4) and grouping the terms conveniently, we obtain

$$\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_{m+1} \mid \mathbf{h} = \left[\sum_{i=1}^{m} (\mathbf{a}_i \mid \mathbf{h}) + (\mathbf{a}_{m+1} \mid \mathbf{h})\right]$$
$$- \left[\sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{h}) + \sum_{i=1}^{m} (\mathbf{a}_i \cdot \mathbf{a}_{m+1} \mid \mathbf{h})\right]$$
$$+ \left[\sum_{i=1}^{m-2} \sum_{j=i+1}^{m-1} \sum_{k=j+1}^{m} (\mathbf{a}_i \cdot \mathbf{a}_j \cdot \mathbf{a}_k \mid \mathbf{h}) + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} (\mathbf{a}_i \cdot \mathbf{a}_j \cdot \mathbf{a}_{m+1} \mid \mathbf{h})\right]$$
$$- \ldots \pm (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_{m+1} \mid \mathbf{h}).$$

The first bracket on the right includes the first summation taken from Eq. (5.3) with the term $\mathbf{a}_{m+1} \mid \mathbf{h}$ of Eq. (5.4). Each succeeding bracket includes a summation taken from Eq. (5.3) with the summation of next lower order taken from Eq. (5.5).

It is obvious on sight that, in the first bracket,

$$\sum_{i=1}^{m} (\mathbf{a}_i \mid \mathbf{h}) + (\mathbf{a}_{m+1} \mid \mathbf{h}) = \sum_{i=1}^{m+1} (\mathbf{a}_i \mid \mathbf{h}),$$

and it is evident on consideration that, in each succeeding bracket, the change of m to $m + 1$ in the upper limits of the first summation makes it include the second. Thus the equation may be written,

$$\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_{m+1} \mid \mathbf{h}$$
$$= \sum_{i=1}^{m+1} (\mathbf{a}_i \mid \mathbf{h}) - \sum_{i=1}^{m} \sum_{j=i+1}^{m+1} (\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{h})$$
$$+ \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \sum_{k=j+1}^{m+1} (\mathbf{a}_i \cdot \mathbf{a}_j \cdot \mathbf{a}_k \mid \mathbf{h}) - \ldots$$
$$\pm (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_{m+1} \mid \mathbf{h}).$$

This is the same as Eq. (5.3), except that the number of propositions appearing in the inferences, which was m in that equation, is $m + 1$ in this one. Therefore Eq. (5.3), being valid when m is 2, as in Eq. (5.2), is now proved for all values of m.

The rather elaborate way in which the limits of summation were indicated in the preceding equations was needed to avoid

ambiguity in the argument. In most discussion, however, no confusion is made by writing Eq. (5.3) with a simpler indication of the limits, as follows:

$$\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m \mid \mathbf{h} = \sum_i(\mathbf{a}_i \mid \mathbf{h}) - \sum_i\sum_{j>i}(\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) + \sum_i\sum_{j>i}\sum_{k>j}(\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{a}_k \mid \mathbf{h}) - \ldots \pm (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}). \quad (5.6)$$

A review of the induction of Eq. (5.3) or (5.6) from Eq. (5.1) will show that every equation used in the argument remains valid after the exchange of the signs, $\cdot$ and $\vee$. We may therefore make this exchange in Eq. (5.6) and thereby obtain, as a valid equation,

$$\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h} = \sum_i(\mathbf{a}_i \mid \mathbf{h}) - \sum_i\sum_{j>i}(\mathbf{a}_i \vee \mathbf{a}_j \mid \mathbf{h}) + \sum_i\sum_{j>i}\sum_{k>j}(\mathbf{a}_i \vee \mathbf{a}_j \vee \mathbf{a}_k \mid \mathbf{h}) - \ldots \pm (\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m \mid \mathbf{h}). \quad (5.7)$$

If the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, are all mutually exclusive on the hypothesis $\mathbf{h}$, so that every conjunction of two or more of them is impossible, Eq. (5.6) becomes simply

$$\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m \mid \mathbf{h} = \sum_i(\mathbf{a}_i \mid \mathbf{h}). \quad (5.8)$$

It is often the case that an argument has to do with a set of propositions, none of which, it may be, is certain, but which, on the given hypothesis, can not all be false. Such a set is called *exhaustive* on the hypothesis. Let W propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_W$, comprise such a set. Then (whether or not the propositions are mutually exclusive)

$$\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_W \mid \mathbf{h} = 1 \quad (5.9)$$

and, if they are mutually exclusive,

$$\sum_{i=1}^{W}(\mathbf{a}_i \mid \mathbf{h}) = 1. \quad (5.10)$$

If, finally, these propositions are all equally probable on the hypothesis **h**, it follows from this equation that each has the probability $1/W$. Hence, by Eq. (5.8), the disjunction of any w propositions of the set has the probability w/W.

6. A Remark on Measurement

It has been the thesis of the preceding chapters that probable inference of every kind, the casual and commonplace no less than the formalized and technical, is governed by the same rules, and that these rules are all derived from two principles, both of them agreeable to common sense and simple enough to be accepted as axioms.

It does not follow that all probabilities can be estimated with the same precision. Some probabilities are well defined, others are ill defined and still others are scarcely defined at all except that they are limited, as all probabilities are, by the extremes of certainty and impossibility. In this respect, however, probability is not essentially different from other quantities, for example length. A steel cylinder, carefully faced and polished, has a better defined length than a plank. The length of a rope frayed at the ends is ill defined and that of a trail of smoke is very ill defined indeed. The differences, however, are differences of degree, not of kind, and we speak of a trail of smoke two or three miles long as naturally as we speak of a yardstick. There are always, as the Bishop in Robert Browning's poem[11] said in another connection, "clouds of fuzz where matters end," even if the fuzz is only the attenuation of interatomic forces. The difference between one of these lengths and another is only that some clouds are fuzzier than others. There is no length defined with complete precision, nor is length the only quantity of which this can be said. Reflection suggests, indeed, that the only perfectly precise measurement is counting and that the only quantities defined perfectly are those defined in terms of whole numbers.[12]

In the case of physical measurements, it is sometimes impractical to discriminate between indeterminacies due to vagueness of definition on the one hand and mistakes caused by lack of skill or care on the other. Both are therefore often lumped under the head of experimental error. There is, however, a significant distinction in principle between them. Consider, for example, the counting of children on an enclosed playground. This is an example of a measurement very much subject to error, because children will not stand still long enough to be counted. The number itself, however, is a perfectly defined quantity; if there are 40 children on the playground, anyone who counts 37 or 42 has made a mistake. By contrast, the length of a trail of smoke has an intrinsic indeterminacy, which can not be eliminated by any skill or care in its measurement. It has no one true value from which every deviation is a mistake.

As a rule, probable inference is more like measuring smoke than counting children, in that the probabilities themselves are not well defined. There are some instances, however, in which the definition is precise and in any such case there are unique values of the probabilities, from which deviations can occur only as the result of mistakes in logic or arithmetic. An obvious example is the case in which the hypothesis logically implies or contradicts the inference, so that the probability is that of certainty or impossibility and can be reckoned otherwise only by false reasoning.

With two or more inferences, it is sometimes possible to make a judgment variously called one of *non-sufficient* or *insufficient reason* or *indifference*.[13] This is a judgment of equal probability, which can be made among several inferences when everything asserted in the hypothesis in proof or disproof of any one of them is equally asserted in proof or disproof of every other. Like the judgments of certainty and impossibility, it is independent of the scale of measurement, because inferences equally probable on one scale are so on all scales.

A combination of these three judgments, when it is possible,

affords a precise definition of probabilities. We have seen, in the chapter before this one, if the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_W$, form an exhaustive set and are mutually exclusive and equally probable on the hypothesis $\mathbf{h}$, that an inference expressible as the disjunction of w of them has the probability w/W on this hypothesis. That the propositions form an exhaustive set is a judgment of certainty, according to which $\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_W \mid \mathbf{h} = 1$. That they are mutually exclusive is a judgment of impossibility, according to which $\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{h} = 0$ for all different values of i and j. Finally that they are equally probable is a judgment of indifference, according to which $\mathbf{a}_1 \mid \mathbf{h} = \mathbf{a}_2 \mid \mathbf{h} = \ldots = \mathbf{a}_W \mid \mathbf{h}$.

Some writers on probability have supposed that two inferences are equally probable and each has therefore the probability $\frac{1}{2}$ when nothing is known about them except that each is the other's contradictory.[14] According to this opinion, for example, a snark is just as likely as not to be a boojum on the hypothesis which says nothing about either snarks or boojums except that every snark either is or is not a boojum.[15] In more formal terms, it is supposed that $\mathbf{a} \mid \mathbf{a} \vee \sim\mathbf{a} = \frac{1}{2}$ for arbitrary meanings of $\mathbf{a}$.

In disproof of this supposition, let us consider the probability of the conjunction $\mathbf{a} \cdot \mathbf{b}$ on each of the two hypotheses, $\mathbf{a} \vee \sim\mathbf{a}$ and $\mathbf{b} \vee \sim\mathbf{b}$. We have

$$\mathbf{a} \cdot \mathbf{b} \mid \mathbf{a} \vee \sim\mathbf{a} = (\mathbf{a} \mid \mathbf{a} \vee \sim\mathbf{a})[\mathbf{b} \mid (\mathbf{a} \vee \sim\mathbf{a}) \cdot \mathbf{a}].$$

By Eq. (2.8 I), $(\mathbf{a} \vee \sim\mathbf{a}) \cdot \mathbf{a} = \mathbf{a}$ and therefore

$$\mathbf{a} \cdot \mathbf{b} \mid \mathbf{a} \vee \sim\mathbf{a} = (\mathbf{a} \mid \mathbf{a} \vee \sim\mathbf{a})(\mathbf{b} \mid \mathbf{a}).$$

Similarly

$$\mathbf{a} \cdot \mathbf{b} \mid \mathbf{b} \vee \sim\mathbf{b} = (\mathbf{b} \mid \mathbf{b} \vee \sim\mathbf{b})(\mathbf{a} \mid \mathbf{b}).$$

But, also by Eq. (2.8 I), $\mathbf{a} \vee \sim\mathbf{a}$ and $\mathbf{b} \vee \sim\mathbf{b}$ are each equal to $(\mathbf{a} \vee \sim\mathbf{a}) \cdot (\mathbf{b} \vee \sim\mathbf{b})$ and each is therefore equal to the other. Thus

$$\mathbf{a} \cdot \mathbf{b} \mid \mathbf{b} \vee \sim\mathbf{b} = \mathbf{a} \cdot \mathbf{b} \mid \mathbf{a} \vee \sim\mathbf{a}$$

and hence

$$(\mathbf{a} \mid \mathbf{a} \vee \sim\mathbf{a})(\mathbf{b} \mid \mathbf{a}) = (\mathbf{b} \mid \mathbf{b} \vee \sim\mathbf{b})(\mathbf{a} \mid \mathbf{b}).$$

If then $\mathbf{a} \mid \mathbf{a} \vee \sim\mathbf{a}$ and $\mathbf{b} \mid \mathbf{b} \vee \sim\mathbf{b}$ were each equal to $\frac{1}{2}$, it would follow that $\mathbf{b} \mid \mathbf{a} = \mathbf{a} \mid \mathbf{b}$ for arbitrary meanings of **a** and **b**. This would be a monstrous conclusion, because $\mathbf{b} \mid \mathbf{a}$ and $\mathbf{a} \mid \mathbf{b}$ can have any ratio from zero to infinity. Instead of supposing that $\mathbf{a} \mid \mathbf{a} \vee \sim\mathbf{a} = \frac{1}{2}$, we may more reasonably conclude, when the hypothesis is the truism, that all probabilities are entirely undefined except those of the truism itself and its contradictory, the absurdity. This conclusion agrees with common sense and might perhaps have been reached without the formal argument, because the knowledge of a probability, though it is knowledge of a particular and limited kind, is still knowledge, and it would be surprising if it could be derived from the truism, which is the expression of complete ignorance, asserting nothing.

Not only must the hypothesis of a probability assert something, if the probability is to be defined within any limits narrower than the extremes of certainty and impossibility, but also what it asserts must have some relevance to the inference. For example, the probability of the inference, "There will be scattered thundershowers tonight in the lower Shenandoah Valley," is entirely undefined on the hypothesis, "Dingoes are used as half tamed hunting dogs by the Australian aborigines," although the hypothesis is by no means without meaning and gives a fairly precise definition and a value near certainty to the inference, "The Australian aborigines are not vegetarians."

The instances in which probabilities are precisely defined are thus circumscribed on two sides. On the one hand, the hypothesis must provide some information relevant to the inferences, for otherwise their probabilities are not defined at all. On the other hand, this information must contain nothing which favors one of the inferences more than another, for then the judgment of indifference on which precise definition rests is impossible. The cases are exceptional in which our actual knowledge provides an

hypothesis satisfying these conditions. Although we are apt to say, especially when we are perplexed, that one guess is as good as another, the circumstances are rare in which this is really true. They are present in games of chance, but there they are prescribed by the rules of the game or result from the design of its equipment.[16] It is to insure indifference that cards are shuffled and cut and dice are shaken. For the same reason, the cards of a pack are made identical except for the designs on their faces and dice are made symmetrical in shape and homogeneous in composition. In certain statistical studies also, where indifference is attained or at least closely approximated, it is attained by intention and sometimes only by elaborate precautions.

It is mainly, if not indeed only, in cases like these that probabilities can be precisely estimated. Most of the time we are limited instead to approximations or judgments of more or less. Someone will say, for example, in discussing the prospects of a candidate for political office, "There are at least three other candidates more likely than he is to be nominated and, even if he wins the nomination, he will have no better than an even chance of election." Thence it is argued that his chances are very poor.

To see the formal structure of this argument, let $\mathbf{a}_i$ be the inference that the ith candidate will be nominated and $\mathbf{b}_i$ the inference that he will be elected, and let $\mathbf{h}$ be the unstated initial hypothesis. Then the quoted remark asserts that

$$\mathbf{a}_1 \mid \mathbf{h} < \mathbf{a}_i \mid \mathbf{h} \tag{6.1}$$

and

$$\mathbf{b}_1 \mid \mathbf{a}_1 \cdot \mathbf{h} \leq \sim \mathbf{b}_1 \mid \mathbf{a}_1 \cdot \mathbf{h}, \tag{6.2}$$

where the subscript 1 refers to the candidate under discussion and i has each of the values, 2, 3, 4, in reference to the three candidates mentioned in comparison.

The propositions, $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ and $\mathbf{a}_4$, are mutually exclusive but they do not form an exhaustive set, because the words, "at least", imply that there are still more candidates. Therefore

$$(\mathbf{a}_1 \mid \mathbf{h}) + (\mathbf{a}_2 \mid \mathbf{h}) + (\mathbf{a}_3 \mid \mathbf{h}) + (\mathbf{a}_4 \mid \mathbf{h})$$

$$= \mathbf{a}_1 \vee \mathbf{a}_2 \vee \mathbf{a}_3 \vee \mathbf{a}_4 \mid \mathbf{h} < 1,$$

whence it follows, by the inequality (6.1), that

$$\mathbf{a}_1 \mid \mathbf{h} < \tfrac{1}{4}.$$

Also, $\mathbf{b}_1 \mid \mathbf{a}_1 \cdot \mathbf{h} = 1 - (\sim \mathbf{b}_1 \mid \mathbf{a}_1 \cdot \mathbf{h})$ and thus, by the inequality (6.2),

$$\mathbf{b}_1 \mid \mathbf{a}_1 \cdot \mathbf{h} \leq \tfrac{1}{2}.$$

Finally, $\mathbf{a}_1 \cdot \mathbf{b}_1 \mid \mathbf{h} = (\mathbf{a}_1 \mid \mathbf{h})(\mathbf{b}_1 \mid \mathbf{a}_1 \cdot \mathbf{h})$, and thus we find that

$$\mathbf{a}_1 \cdot \mathbf{b}_1 \mid \mathbf{h} < \tfrac{1}{8},$$

so that the odds against this candidate are more than 7 to 1.

It is seldom worth the time it takes to trace in such detail as this the steps of probable inference any more than it is ordinarily worth while to reduce deductive reasoning to syllogisms. This one example is offered to support the argument that, however much we are obliged to forego numerical precision in probable inference, we do not, in reasonable discourse, dispense with the rules of probability, although we may use them so familiarly as to be unaware of them. When we employ probable inference as a guide to reasonable decisions, it is by these rules that we judge that one alternative is more probable than another or that some inference is so nearly certain that we can take it for granted or some contingency so nearly impossible that we can leave it out of our calculation.

II

Entropy

7. Entropy as Diversity and Uncertainty and the Measure of Information

It is often convenient to consider as a group rather than as single propositions the inferences which, on some given evidence, form an exhaustive set. A number of remarks are commonplace in such consideration, sometimes one, sometimes another, as the circumstances vary. In some cases, for example, it may be appropriate to say, "There are many possibilities, one as likely as another and no two of them the same." By contrast, it may be said under other circumstances, "There are not many different possibilities and, of these, only a few are at all probable." Comments such as these show, in a rough way, differences which are made quantitative by the concept of *entropy*.[17]

The meaning of entropy is not the same in all respects as that of anything which has a familiar name in common use, and it is therefore impossible to give a simple verbal description of it, which is, at the same time, an accurate definition. It is evident, however, that what is aimed at in remarks such as those just quoted is an estimate of something like the diversity among the inferences and also something like the uncertainty, on the given hypothesis, of the whole set.

Now, if entropy is to measure the diversity among the inferences, it must depend on their number, increasing as the number is increased, when other things are kept as far as possible the same, for a single inference without an alternative obviously has

no diversity. But, if entropy is to measure also their uncertainty, it can not depend on their number alone but must involve their probabilities as well, for impossible propositions, no matter how numerous, add nothing to the uncertainty, and propositions nearly impossible add little. Finally, if entropy is to measure either diversity or uncertainty, it must depend on the extent to which the inferences are mutually compatible, diminishing as their compatibility is increased. For compatibility, carried to the limit, becomes identity, and, if two propositions are identical, the set which includes both of them is no more diverse or uncertain than that which includes only one.

With this understanding of the meaning of entropy, let us consider first its dependence on the number of inferences. We postpone consideration of differences in their probabilities by assuming them all equally probable. In order similarly to avoid considering the effect of their mutual compatibility, we choose the extreme case in which they are completely incompatible and assume them all mutually exclusive. Thus we consider, as the simplest example, the entropy of an exhaustive set of equally probable and mutually exclusive propositions.

As a familiar hypothesis for such an example, let us suppose that a card is drawn from a well shuffled pack. Then the propositions, "The card drawn is the six of diamonds" and "The card drawn is the ten of clubs," are two from an exhaustive set of fifty-two mutually exclusive and equally probable propositions. By the description of entropy just given, it will be determined in this example by the number 52.

There is an implication here which should be made explicit, that entropy measures uncertainty and diversity in a distinct and quite restricted sense, according to which differences in meaning among the propositions of a set are significant only insofar as they affect the probabilities of the propositions. In another sense, the uncertainty in the present example would be altered by a wager placed on the drawing of the card, but the entropy is the same whether there is a fortune at stake or a trifle or nothing at

all. Again, there is a sense in which the diversity would depend on the pictorial contrast among the cards and would be greater if the queen of hearts had red hair and the queen of diamonds golden hair than if they were both blondes of the same hue, but the entropy is unaffected by such differences as long as means remain by which each card can be distinguished from the others. Readers familiar with entropy in thermodynamics, where it was first given a clear meaning and a name, will recall, in further illustration of the same principle, that the entropy of mixing ideal gases depends only on the existence of a detectible difference between the molecules of the several gases and not on the nature and magnitude of the difference.[18]

Let us note now that the proposition, "The card drawn is the king of spades," is the conjunction of the two propositions, "The card drawn is a spade" and "The card drawn is a king." There are four equally probable propositions for naming the suit of the card and thirteen for naming the card in the suit. To specify one proposition among the four and one among the thirteen is the same as to specify one in the set of fifty-two. Thus the diversities of these two sets jointly make the diversity of the set of conjunctions. It proves convenient to define entropy in such a way as to measure the total diversity by the sum of the entropies which measure the partial diversities. If, therefore, we denote by $\eta(w)$ the entropy of an exhaustive set of w equally probable and mutually exclusive propositions, we have, in this example,

$$\eta(52) = \eta(4) + \eta(13)$$

and, in general,

$$\eta(xy) = \eta(x) + \eta(y). \tag{7.1}$$

Differentiation with respect to x and y gives

$$y \frac{\mathrm{d}\eta(xy)}{\mathrm{d}(xy)} = \frac{\mathrm{d}\eta(x)}{\mathrm{d}x}$$

and

$$x \frac{d\eta(xy)}{d(xy)} = \frac{d\eta(y)}{dy},$$

whence we obtain, by eliminating $d\eta(xy)/d(xy)$,

$$x\, d\eta(x)/dx = y\, d\eta(y)/dy.$$

Since x and y are independent variables, this equation requires that each of its members be equal to a constant. Calling this constant k, we have then

$$d\eta(w) = (k/w)\, dw,$$

whence we find by integration that

$$\eta(w) = k \ln w + C,$$

where C is a constant of integration. By substitution in Eq. (7.1), we find that $C = 0$. Thus

$$\eta(w) = k \ln w.$$

In thermodynamics, k is the well known Boltzmann constant and has a value determined by the unit of heat and the scale of temperature. In the theory of probability it is convenient to assign it unit value, so that

$$\eta(w) = \ln w. \tag{7.2}$$

Whatever value is assigned to k, when $w = 1$, $\eta = 0$; when there is only one possible inference, there is no diversity or uncertainty.

The special appropriateness of the logarithm rather than some other function in this expression can be made plainer by considering the game of twenty questions, in which one player or one side chooses a subject and the other player or side asks questions to find out what it is. The rules vary with the age and skill of the players, but a usual requirement is that all questions must be answerable by "yes" or "no." The skill of the questioner is shown by finding the subject with as few questions as possible. If one player opens the game by saying, "I am thinking of a famous person," the other, if it is a child just learning to play,

may ask, "Is it Christopher Columbus?" or "Is it Pocahontas?" A bright child soon learns, however, that the game can usually be ended earlier by beginning with general questions, such as "Is it a man?" or "Is it someone living now?" for which the probabilities of "yes" and "no" for the answer are somewhere near to being equal.

As an example of the simplest kind, let one player say, "I am thinking of a whole number between 1 and 32." If the other player chooses to go through the numbers one at a time, asking, "Is it 1? Is it 2?" and so on to "Is it 31?" it is possible that he will win on the first question. But he may have to ask thirty-one questions, whereas he is sure to win in five questions if he asks first, "Is it greater than 16?" and then, according to the answer, "Is it greater than 8?" or "Is it greater than 24?" and so continues, choosing each question so that its answer will halve the number of alternatives left by the preceding one. If his opponent chooses numbers with no systematic preference, no other strategy will end the game, on the average, with as small a number of questions.

The game in this example has the following description in terms of entropy. The propositions, "The number is 1, the number is 2, . . . the number is 32," are mutually exclusive and, it was assumed, equally probable, and they form an exhaustive set, of which the entropy, therefore, is ln 32. The answer to the first question leaves 16 possible alternatives, forming a set with the entropy, ln 16. At any question, if the number of alternatives is w, the answer reduces it to $\frac{1}{2}w$ and thus diminishes the entropy by ln 2. Hence n questions will diminish the entropy to zero from an initial value n ln 2. With 20 questions it becomes possible to find a chosen integer between 1 and 2^{20} or 1,048,576.

The usual reason for asking questions, other than rhetorical ones, is to obtain information, and the more information is needed, the more questions must be asked. We have just seen also that the greater is the entropy of a set of propositions, the more questions are required to find which one of them is true. Entropy

thus appears in yet another aspect, as the measure of information. The amount of information elicited by a question to which there are only two possible answers, which are equally probable, is measured by the entropy of a pair of mutually exclusive and equally probable propositions. In the theory of communication this is often a convenient unit. It is called one *bit*. In the strategy just described for twenty questions, each question elicits one bit of information, and the number of questions required to end the game is the number of bits in the initial entropy.[19]

8. *Entropy and Probability*

Considering entropy as a measure of information, let us now inquire how it may be expressed when the inferences of which it is a function are no longer required to be equally probable, though they are still assumed to be mutually exclusive and to form an exhaustive set.

In order to make use of the result obtained in the preceding chapter, let us take a case of equal probabilities as a point of departure. For instance, we may consider a raffle in which W equal chances are offered for sale. By Eq. (7.2), the entropy which measures the information required to identify the winning chance is equal to $\ln W$.

Let us suppose now that the chances are sold in blocks, so that, for example, a block of w_1 chances is sold to the Board of Trade for resale to its members. Let w_2 chances be distributed in the same way to members of the League of Women Voters, w_3 chances to the Boy Scouts, and so on until every chance is sold to a member of some one of m societies. It is to be assumed that the societies are mutually exclusive, so that no purchaser belongs to more than one of them.

Let all of these assumptions be expressed in the hypothesis $\mathbf{h}$ and let $\mathbf{a}_i$ denote the proposition that the winning chance is held by a member of the i th society. Then, on the hypothesis $\mathbf{h}$, the

propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, are mutually exclusive and form an exhaustive set. This is the set of inferences for the entropy of which we now seek an expression.

If the same number of chances were in every block, the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, would all be equally probable. Their entropy could be denoted by $\eta(m)$ and it would be equal to $\ln m$. The case would be formally identical with that considered in the preceding chapter, the number of societies in the present example corresponding to the number of suits of cards in the former one and the number of chances held in each society corresponding to the number of cards in each suit. The entropy, $\eta(m)$, would measure the information required to find in which society the winning chance is held, and the additional information required to find the winning chance among those held there would be measured by an entropy denoted by $\eta(w)$ and equal to $\ln w$, where w is the number of chances sold in each block. In this case, therefore, we should have the equation,

$$\eta(m) + \eta(w) = \ln m + \ln w = \ln (mw) = \ln W.$$

When there are different numbers of chances in the various blocks and the inferences, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, are therefore no longer equally probable, their entropy is no longer a function of m alone. Consequently it can not be denoted by $\eta(m)$ and it is, of course, not equal to $\ln m$. Let us denote it by $\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h})$ until, in a later chapter, we can explain and justify a simpler notation, and let us seek an expression for it by asking how much additional information we shall need to find the winning chance, if we suppose that we first obtain the information which this entropy measures.

If we find in the first inquiry that a member of the Board of Trade holds the winning chance, the required additional information will be measured by $\eta(w_1)$, whereas it will be measured by $\eta(w_2)$ if we find that the winning chance is held in the League of Women Voters. We can not know, in advance of the first inquiry, how much additional information will be needed after

the inquiry is made. We know only that there is a probability, $\mathbf{a}_1 \mid \mathbf{h}$, that it will be measurable by $\eta(w_1)$, a probability, $\mathbf{a}_2 \mid \mathbf{h}$, that it will be measurable by $\eta(w_2)$ and, in general, a probability, $\mathbf{a}_i \mid \mathbf{h}$, that it will be measurable by $\eta(w_i)$. Our best estimate, *a priori*, of the entropy which will measure this information is $\Sigma_i(\mathbf{a}_i \mid \mathbf{h})\eta(w_i)$, where the summation is over values of i from 1 to m. Therefore we may reasonably require $\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h})$ to satisfy the equation,

$$\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h}) + \textstyle\sum_i(\mathbf{a}_i \mid \mathbf{h})\eta(w_i) = \ln W. \qquad (8.1)$$

By Eq. (7.2), $\eta(w_i) = \ln w_i$, and, by the familiar rule for the measurement of probabilities discussed in Chapter 6, $\mathbf{a}_i \mid \mathbf{h} = w_i/W$, so that $w_i = (\mathbf{a}_i \mid \mathbf{h})W$. Hence

$$\begin{aligned}\textstyle\sum_i(\mathbf{a}_i \mid \mathbf{h})\eta(w_i) &= \textstyle\sum_i(\mathbf{a}_i \mid \mathbf{h})[\ln (\mathbf{a}_i \mid \mathbf{h}) + \ln W] \\ &= \textstyle\sum_i(\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}) + \ln W,\end{aligned}$$

because $\Sigma_i(\mathbf{a}_i \mid \mathbf{h}) = 1$. Substituting this expression above, we find

$$\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h}) = -\textstyle\sum_i(\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}). \qquad (8.2)$$

The constant, k, if it had not been given unit value in the preceding chapter, would appear as a factor on the right in this equation. Except for this omission, the equation gives the most general expression possible for the entropy of a set of mutually exclusive propositions.

Because the limits of probability are 0 and 1 and the logarithm of any number between these limits is negative, it follows that:

> The entropy of a set of mutually exclusive propositions can not be negative. (8.i)

If any proposition of the set is impossible, the term it contributes to the entropy is equal to the limit, as x approaches zero, of $x \ln x$. This limit is zero and thus we see that the inclusion, among a set of inferences, of any proposition impossible on the hypothesis does not change the entropy of the set. If any of a

set of mutually exclusive propositions is certain, all the others are impossible. The entropy is thus reduced to that of a single inference with no alternatives, which, as we have seen before, is zero.

Regarding entropy again as the measure of uncertainty, we should expect it to have its maximum value when the hypothesis favors no inference more than another and thus assigns the probability $1/m$ to each of them and the entropy $\ln m$ to the set. That this is true is seen by making infinitesimal variations, $\delta(\mathbf{a}_1 \mid \mathbf{h})$, $\delta(\mathbf{a}_2 \mid \mathbf{h})$, $\ldots$ $\delta(\mathbf{a}_m \mid \mathbf{h})$, in the probabilities in Eq. (8.2) to find the resulting variation in the entropy. Thus we obtain

$$\delta\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h}) = -\sum_i [\ln (\mathbf{a}_i \mid \mathbf{h}) + 1]\delta(\mathbf{a}_i \mid \mathbf{h}),$$

which becomes, when the inferences are all equally probable,

$$\delta\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h}) = (\ln m - 1)\sum_i \delta(\mathbf{a}_i \mid \mathbf{h}).$$

Because $\Sigma_i(\mathbf{a}_i \mid \mathbf{h}) = \mathbf{1}$, it follows that $\Sigma_i\, \delta(\mathbf{a}_i \mid \mathbf{h}) = 0$ and thus

$$\delta\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h}) = 0.$$

This vanishing of the variation of the entropy confirms our expectation and proves the theorem:

> The entropy of a set of mutually exclusive propositions is maximum when they are equally probable and is then equal to $\ln m$, where m is their number.[20] (8.ii)

If nothing else, then curiosity alone might urge us here to go farther and seek an expression for the entropy of inferences which form an exhaustive set but are not required to be mutually exclusive any more than equally probable.[21] For this purpose, the raffle we have been considering will still serve as an example, if it is allowed, in contradiction to what was assumed before, that some of those who hold chances belong to more than one of the societies. As before, we denote by w_i the number of chances held by members of the i th society and by $\mathbf{a}_i$ the proposition that one of these is the winning chance, but we no longer suppose that

$\mathbf{a}_i \cdot \mathbf{a}_j$ is impossible and we denote by w_{ij} the number of chances held by persons who belong to both the i th and j th societies.

Consider the term $\Sigma_i(\mathbf{a}_i \mid \mathbf{h})\eta(w_i)$ in Eq. (8.1). When it was assumed that $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$ were mutually exclusive propositions, this term measured the information which was not included in that measured by $\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h})$ and was anticipated as necessary for finding the winning chance. But on the new assumption it is too large for that purpose, because now there are chances held by persons who are members of two societies and this summation counts all of these chances twice. For example, a chance held by someone who is a member of both the Board of Trade and the League of Women Voters would be taken account of in both of the terms $(\mathbf{a}_1 \mid \mathbf{h})\eta(w_1)$ and $(\mathbf{a}_2 \mid \mathbf{h})\eta(w_2)$. Allowance for the overlapping membership of these two societies requires the subtraction of a corrective term, $(\mathbf{a}_1 \cdot \mathbf{a}_2 \mid \mathbf{h})\eta(w_{12})$. The correction for duplicate membership among all pairs of societies is $\Sigma_i\Sigma_{j>i}(\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{h})\eta(w_{ij})$, where it is to be understood, as in Chapter 5, that the upper limits of summation are $m - 1$ for i and m for j and the restriction of j to values greater than i insures that the correction is made only once for each pair of societies.

But now, if there are persons holding chances who belong to three societies, this correction will be excessive and will itself have to be corrected by subtracting from it

$$\sum_i \sum_{j>i} \sum_{k>j} (\mathbf{a}_i \cdot \mathbf{a}_j \cdot \mathbf{a}_k \mid \mathbf{h})\eta(w_{ijk}),$$

where w_{ijk} denotes the number of chances held by those who are members of the ith, jth and kth societies. The same reasoning, continued, calls for a series of corrections, which alternate in sign because each one corrects for the excess of the one preceding it. The series ends with the correction required by the chances held by those who are members of all m societies. The complete equation, replacing Eq. (8.1), is therefore

$$\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h}) + \sum_i (\mathbf{a}_i \mid \mathbf{h})\eta(w_i)$$
$$- \sum_i \sum_{j>i} (\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{h})\eta(w_{ij}) + \sum_i \sum_{j>i} \sum_{k>j} (\mathbf{a}_i \cdot \mathbf{a}_j \cdot \mathbf{a}_k \mid \mathbf{h})\eta(w_{ijk})$$
$$- \ldots \pm (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_m \mid \mathbf{h})\eta(w_{12\ldots m}) = \ln W.$$

From this, by means of the equations,

$$\eta(w_i) = \ln w_i, \qquad \eta(w_{ij}) = \ln w_{ij}, \ldots$$

$$\eta(w_{12\ldots m}) = \ln w_{12\ldots m},$$

and

$$(\mathbf{a}_i \mid \mathbf{h}) = w_i/W, \qquad \mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h} = w_{ij}/W, \ldots$$

$$\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h} = w_{12\ldots m}/W,$$

we obtain

$$\begin{aligned}\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h}) = &- \textstyle\sum_i(\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}) \\ &+ \textstyle\sum_i\sum_{j>i}(\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) \\ &- \textstyle\sum_i\sum_{j>i}\sum_{k>j}(\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{a}_k \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{a}_k \mid \mathbf{h}) + \ldots \\ &\pm (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}) \ln (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}) \\ &- [\textstyle\sum_i(\mathbf{a}_i \mid \mathbf{h}) - \sum_i\sum_{j>i}(\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) + \ldots \\ &\mp (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}) - 1] \ln W.\end{aligned}$$

By Eq. (5.6), the expression in brackets on the right is equal to $(\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m \mid \mathbf{h}) - 1$ and is thus zero, since the set of inferences is exhaustive. Thus we have finally, as the most general expression for entropy, the equation,

$$\begin{aligned}\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m \mid \mathbf{h}) = &- \textstyle\sum_i(\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}) \\ &+ \textstyle\sum_i\sum_{j>i}(\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) \\ &- \textstyle\sum_i\sum_{j>i}\sum_{k>j}(\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{a}_k \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{a}_k \mid \mathbf{h}) + \ldots \\ &\pm (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}) \ln (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}). \quad (8.3)\end{aligned}$$

It can be seen that this equation becomes identical with Eq. (8.2) when the inferences are mutually exclusive, because all the conjunctions are then impossible and therefore the terms which involve them vanish. If the inferences appearing in Eq. (8.3) are not only mutually exclusive but also equally probable, the equation becomes the same as Eq. (7.2), except that the number

of inferences is denoted by different letters in the two equations.

For the proof of theorems in this and later chapters, we shall find it convenient to have an expression for the entropy in which the terms involving one proposition of the set of inferences are separated from the rest of the terms. To emphasize the separation, let us denote the proposition thus singularly treated by $\mathbf{b}$ and the other propositions by $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, so that there are $m + 1$ propositions in the set. Equation (8.3), when modified to express the entropy of this set, becomes

$$\begin{aligned}\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m, \mathbf{b} \mid \mathbf{h}) \\ = &-\sum_i(\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}) + \sum_i\sum_{j>i}(\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{a}_j \mid \mathbf{h}) \\ &- \ldots \qquad \pm (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}) \ln (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m \mid \mathbf{h}) \\ &- [(\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{b} \mid \mathbf{h}) - \sum_i(\mathbf{a}_i\cdot\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{b} \mid \mathbf{h}) \\ &+ \sum_i\sum_{j>i}(\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}) - \ldots \\ &\pm (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m\cdot\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{a}_1\cdot\mathbf{a}_2\cdot\ldots\cdot\mathbf{a}_m\cdot\mathbf{b} \mid \mathbf{h})]. \qquad (8.4)\end{aligned}$$

By this equation we may now prove the theorem:

If one proposition of a set implies another proposition of the same set, it does not contribute to the entropy of the set. (8.iii)

Let $\mathbf{b}$ imply $\mathbf{a}_1$. Then $\mathbf{a}_1 \mid \mathbf{b}\cdot\mathbf{h} = 1$ and, since $\mathbf{a}_1\cdot\mathbf{b} \mid \mathbf{h} = (\mathbf{a}_1 \mid \mathbf{b}\cdot\mathbf{h})(\mathbf{b} \mid \mathbf{h})$, it follows that

$$\mathbf{a}_1\cdot\mathbf{b} \mid \mathbf{h} = \mathbf{b} \mid \mathbf{h}.$$

Similarly,

$$\mathbf{a}_1\cdot\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h} = \mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}, \ldots$$

Therefore, in Eq. (8.4),

$$\sum_i(\mathbf{a}_i\cdot\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{b} \mid \mathbf{h}) = (\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{b} \mid \mathbf{h}) + \sum_{i>1}(\mathbf{a}_i\cdot\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{b} \mid \mathbf{h}),$$

and

$$\sum_i\sum_{j>i}(\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}) = \sum_{j>1}(\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}) + \sum_{i>1}\sum_{j>i}(\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{a}_i\cdot\mathbf{a}_j\cdot\mathbf{b} \mid \mathbf{h}).$$

A change of subscripts makes the first summation on the right in this equation identical with the summation on the right in the preceding equation. Thus, when these and other expressions similarly obtained are substituted in Eq. (8.4), the quantity in brackets there becomes a series of pairs of terms equal in magnitude and opposite in sign. In this way, all the terms involving the proposition **b** vanish from the equation and the theorem is proved.

From this theorem there follows the one already proved in the case of mutually exclusive propositions, that:

If any proposition of a set is certain, the entropy of the set is zero. (8.iv)

This is because an inference which is certain on a given hypothesis is implied by every proposition which is possible on the hypothesis. Therefore, by the theorem just proved, no other proposition of the set contributes to the entropy of the set. The entropy is thus reduced to a single term, $(\mathbf{a}_1 \mid \mathbf{h}) \ln (\mathbf{a}_1 \mid \mathbf{h})$, where $\mathbf{a}_1$ is the inference which is certain, and this term is zero because $\ln 1 = 0$.

This theorem can also be proved directly, without making use of the preceding one, by returning to Eq. (8.4) and letting **b** be certain. Then $\mathbf{a}_i \cdot \mathbf{b} \mid \mathbf{h} = \mathbf{a}_i \mid \mathbf{h}$, $\mathbf{a}_i \cdot \mathbf{a}_j \cdot \mathbf{b} \mid \mathbf{h} = \mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{h}$, . . . and thus the terms in the brackets are all canceled by those outside, except $(\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{b} \mid \mathbf{h})$, which is zero.

Equations (7.2), (8.2) and (8.3) express the entropy in three different cases, of which the first is the most restricted and the third is the most general, but each gain in generality is accompanied by a loss in the formal simplicity of the expression, which reflects a corresponding loss in the intuitive simplicity of the concept. In Eq. (7.2), which is applicable only to the case in which the inferences are equally probable and mutually exclusive, the entropy, being given by $\ln w$, measures the diversity of the inferences in the simplest and most immediate sense of their mere number. Equation (8.2) is the generalization obtained by dis-

carding the requirement of equal probability while retaining that of mutual exclusion. The expression so obtained,

$$-\sum_i(\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}),$$

not only is formally less simple than ln w, but also can not be so immediately interpreted as the measure of diversity. It is instead more adequately described by the more complex notion of uncertainty. In Eq. (8.3) the requirement of mutual exclusion is also discarded and the result is a much more elaborate expression for the entropy. Moreover, when the inferences are not mutually exclusive, the certainty of one proposition no longer implies that all the others are impossible but allows, on the contrary, a great deal of uncertainty among them, although, by the theorem just proved, the entropy is zero when one proposition is certain, no matter how numerous and uncertain the others may be. Thus the uncertainty does not vanish with the entropy, and entropy is therefore no longer adequately described as the measure of uncertainty. The idea of entropy as a measure of information, however, continues to be useful, and formal simplicity is in large part regained by introducing the concept of a system of propositions.

9. *Systems of Propositions*

The term, *system of propositions*, will have here a meaning different from the usual one and in some respects almost opposite to it. Ordinarily we think of a system as beginning with a set of axioms, all of them certain by hypothesis, and including, along with these axioms, whatever propositions they imply. Since the axioms are certain, so are all the propositions of the system and hence also the conjunction of all of them. Such a system, in contradistinction to the kind we are about to consider, may be called a "system of consequents" or a "deductive system."

By contrast, we consider here what may be called a "system

of implicants" or an "inductive system." The propositions with which it begins are any which form an exhaustive set. None of them, in the general case, is certain and therefore they can not be called axioms. The complete system comprises these propositions, together with whatever propositions imply them, but it does not include the propositions which they imply. The whole system is exhaustive, because it begins with an exhaustive set, and the disjunction of all of its propositions is therefore certain, but their conjunction is never certain and, in general, none of them is more than probable. This is the only kind of system we have to consider. We can therefore reserve the name of system exclusively for it and dispense with further use of the terms, "system of implicants" and "inductive system."

Although it was convenient in the preceding discussion to describe a system as beginning with a particular set of propositions, it is possible to define it without reference to such a set, and there is some advantage in doing so. Let us therefore define a system of propositions by the two following principles:

The propositions of a system form an exhaustive set. (9.i)

Every proposition which implies a proposition of a system itself belongs to that system. (9.ii)

There is some ambiguity here in that a set of propositions may be exhaustive, or one proposition may imply another, on some hypotheses and not on others. In every self-consistent argument, however, there is an hypothesis common to the whole discourse and the more particular hypotheses employed in its various stages are all conjunctions of this with other propositions, in which alone they differ. A set of propositions exhaustive on the common hypothesis is exhaustive on every special one, and a proposition implied by another on the common hypothesis is similarly implied by the special ones as well. It is to be understood in any argument, when propositions are taken as forming a system, that it is in respect to the common hypothesis of the argument that they satisfy the two rules just given.

By the second of these rules, every system includes an unlimited number of propositions. For, if **a** is a proposition belonging to a given system and **f**, **g**, . . . are arbitrary propositions, then **a**·**f**, **a**·**g**, . . ., **a**·**f**·**g**, . . . if they are possible, all imply **a** and therefore all belong to the system. If they are impossible, the question whether or not they imply **a** is left open, because impossible propositions are not admissible in an hypothesis. Happily, however, the inclusion of impossible propositions in a system or their exclusion from it proves to be a matter of no consequence.

If a proposition is certain on the common hypothesis, it is implied by every possible proposition. Hence it follows that:

A system which includes any proposition which is certain includes all possible propositions. (9.iii)

Let us denote systems of propositions by capital boldface letters, **A**, **B**, **C**, . . . and let us consider the set of propositions which includes every one belonging to either **A** or **B** and none which belongs to neither of them. Since **A** and **B** are exhaustive sets, so *a fortiori* is this set. Also it includes every proposition which implies one belonging to it, since it includes every proposition which implies a proposition of either **A** or **B**. Therefore it is itself a system, satisfying, as it does, both of the requirements, (9.i) and (9.ii). It is appropriately called the disjunction of **A** and **B** and denoted by $\mathbf{A} \vee \mathbf{B}$. It is defined by the rule:

The system $\mathbf{A} \vee \mathbf{B}$ includes every proposition belonging to either **A** or **B** and no others. (9.iv)

From the notation it might be supposed, if **a** is a proposition belonging to **A** and **b** is one belonging to **B**, that $\mathbf{a} \vee \mathbf{b}$ would be a proposition of $\mathbf{A} \vee \mathbf{B}$. This, however, does not follow from the definition and is not generally true, for $\mathbf{a} \vee \mathbf{b}$ does not belong to either **A** or **B** except in special cases.

It follows from the rule by which $\mathbf{A} \vee \mathbf{B}$ was just defined that $\mathbf{A} \vee \mathbf{A}$ includes the same propositions as **A**, $\mathbf{B} \vee \mathbf{A}$ the same as $\mathbf{A} \vee \mathbf{B}$, and $(\mathbf{A} \vee \mathbf{B}) \vee \mathbf{C}$ the same as $\mathbf{A} \vee (\mathbf{B} \vee \mathbf{C})$. Thus we find

valid for systems of propositions the three equations, familiar in Boolean algebra:

$$\mathbf{A} \vee \mathbf{A} = \mathbf{A}, \qquad \mathbf{B} \vee \mathbf{A} = \mathbf{A} \vee \mathbf{B}$$

and

$$(\mathbf{A} \vee \mathbf{B}) \vee \mathbf{C} = \mathbf{A} \vee (\mathbf{B} \vee \mathbf{C}) = \mathbf{A} \vee \mathbf{B} \vee \mathbf{C}.$$

Next let us consider the set of propositions which includes every one belonging to both **A** and **B** and none which belongs to neither of them or only one. If **a** is any proposition of **A**, and **b** is any proposition of **B**, $\mathbf{a} \cdot \mathbf{b}$ belongs to this set, because it implies both **a** and **b** and therefore belongs to both **A** and **B**. Now **A** and **B**, being exhaustive sets, must each include one or more true propositions, although the hypothesis, as a rule, does not show which ones they are. Consequently there is at least one true conjunction of propositions of **A** and **B**, and the set which includes all the conjunctions includes this one also and is therefore itself exhaustive. This set has thus the first characteristic of a system, as stated in the rule (9.i).

Moreover, every proposition which implies one of this set thereby implies one which belongs to both **A** and **B**. Every such proposition therefore belongs to both **A** and **B** and hence to this set also. Thus this set has the second characteristic of a system, as given by the rule (9.ii), and, having both characteristics, is, like $\mathbf{A} \vee \mathbf{B}$, itself a system. It is appropriately denoted by $\mathbf{A} \cdot \mathbf{B}$, so that we have the conjunction of two systems defined by the rule:

The system $\mathbf{A} \cdot \mathbf{B}$ includes every proposition which belongs to both **A** and **B** and no others. (9.v)

From this it is evident on consideration that

$$\mathbf{A} \cdot \mathbf{A} = \mathbf{A}, \qquad \mathbf{B} \cdot \mathbf{A} = \mathbf{A} \cdot \mathbf{B},$$

and

$$(\mathbf{A} \cdot \mathbf{B}) \cdot \mathbf{C} = \mathbf{A} \cdot (\mathbf{B} \cdot \mathbf{C}) = \mathbf{A} \cdot \mathbf{B} \cdot \mathbf{C}.$$

The propositions which compose the system $(\mathbf{A} \vee \mathbf{B}) \cdot \mathbf{C}$ are those which belong to either **A** or **B** and to **C** and therefore to both

A and **C** or else to both **B** and **C**. But those which belong to **A** and **C** compose the system $\mathbf{A}\cdot\mathbf{C}$, those which belong to **B** and **C** compose the system $\mathbf{B}\cdot\mathbf{C}$, and therefore those which belong to **A** and **C** or to **B** and **C** compose the system $(\mathbf{A}\cdot\mathbf{C}) \vee (\mathbf{B}\cdot\mathbf{C})$. Thus

$$(\mathbf{A} \vee \mathbf{B})\cdot\mathbf{C} = (\mathbf{A}\cdot\mathbf{C}) \vee (\mathbf{B}\cdot\mathbf{C})$$

and, by similar reasoning,

$$(\mathbf{A}\cdot\mathbf{B}) \vee \mathbf{C} = (\mathbf{A} \vee \mathbf{C})\cdot(\mathbf{B} \vee \mathbf{C}).$$

By making **C** and **B** the same in either of these equations, we find that

$$(\mathbf{A}\cdot\mathbf{B}) \vee \mathbf{B} = (\mathbf{A} \vee \mathbf{B})\cdot\mathbf{B}.$$

Now $(\mathbf{A}\cdot\mathbf{B}) \vee \mathbf{B}$ comprises the propositions which belong to both **A** and **B** or to **B**, but all of those belonging to both **A** and **B** necessarily belong to **B**. Thus $(\mathbf{A}\cdot\mathbf{B}) \vee \mathbf{B}$ comprises all the propositions which belong to **B** and no others. Therefore

$$(\mathbf{A}\cdot\mathbf{B}) \vee \mathbf{B} = \mathbf{B}$$

and

$$(\mathbf{A} \vee \mathbf{B})\cdot\mathbf{B} = \mathbf{B}.$$

A comparison between the equations of this chapter and those of Chapter 2 will show that the definitions of this chapter are such as to make the rules of Boolean algebra hold for systems as for individual propositions. To this correspondence, however, there is a striking exception in that the sign $\sim$ has not appeared in this chapter.

It might be supposed possible to define a system $\sim\mathbf{A}$ corresponding to every system **A** and satisfying the pertinent equations of Boolean algebra, among others,

$$(\mathbf{A} \vee \sim\mathbf{A})\cdot\mathbf{B} = \mathbf{B}.$$

Because every proposition belonging to the conjunction of two systems belongs to both of them, this equation would make every proposition belonging to **B** belong also to $\mathbf{A} \vee \sim\mathbf{A}$. Since **B** is

arbitrary, all possible propositions would thus be included in $\mathbf{A} \vee \sim\mathbf{A}$ and whatever propositions were not included in **A** would be included in $\sim\mathbf{A}$, the truism along with the rest. But, by the theorem (9.iii), a system which includes the truism includes every proposition, and all such systems are therefore identical. Thus, if **A**, **B**, **C**, ... are systems which do not include the truism, it would follow that $\sim\mathbf{A} = \sim\mathbf{B} = \sim\mathbf{C} = \ldots$ But then, by the equation, $\sim\sim\mathbf{A} = \mathbf{A}$, it would follow that $\mathbf{A} = \mathbf{B} = \mathbf{C} = \ldots$ and thus all systems which do not include the truism would also be identical. Since this is impossible, we may conclude that there is no analog, or at least no complete analog, in the algebra of systems, to contradiction in the algebra of propositions.[22]

10. The Entropy of Systems

Among the propositions belonging to any system, there are some which may be said to form its *irreducible set.* These propositions are like all the rest in being implied by others of the system, but they are different in that they themselves imply no propositions of the system except, of course, that each one implies itself.

Every proposition belonging to a system implies at least one proposition of the irreducible set. If it belongs to the irreducible set, it still implies itself. If it does not belong to that set, it implies at least one other proposition of the system. This, in turn, either belongs to the irreducible set or implies another, and so on in a chain of implication which can end only with a proposition of that set.

The irreducible set is exhaustive for, if it were other than an exhaustive set, all of its propositions could be false, and then all the propositions of the system would be false, because a false proposition is implied only by a false proposition. But it is impossible that all the propositions of the system should be false, for every system is exhaustive by definition.

The irreducible set is thus described by the three following principles:

No proposition of the irreducible set implies any proposition of the system except itself. (10.i)

Every proposition of the system implies a proposition of the irreducible set. (10.ii)

The irreducible set is exhaustive. (10.iii)

The system is composed of the propositions of the irreducible set, together with every other proposition which, immediately or remotely, implies one of that set. The irreducible set thus determines what propositions belong to the system. So also does any set of propositions which includes the irreducible set and is included in the system, because every proposition which implies one of such a set belongs to the system, and no proposition belongs to the system without implying one of such a set. It is accurate, therefore, as it is also convenient, to speak of a set, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, as defining the system **A** if these conditions are satisfied, and to call it a *defining set* of the system. A defining set is thus described by the rules:

All the propositions of the irreducible set belong to every defining set and all the propositions of every defining set belong to the system. (10.iv)

Every defining set is exhaustive. (10.v)

From these rules and the definitions of the systems, $\mathbf{A} \vee \mathbf{B}$ and $\mathbf{A} \cdot \mathbf{B}$, there follows, almost directly, the theorem:

If the set of propositions, $\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_m$,
defines the system **A** and the set, $\mathbf{b}_1, \mathbf{b}_2, \cdots \mathbf{b}_n$,
the system **B**, then the set,

$$\mathbf{a}_1, \mathbf{a}_2, \cdots \mathbf{a}_m, \mathbf{b}_1, \mathbf{b}_2, \cdots \mathbf{b}_n,$$

defines the system $\mathbf{A} \vee \mathbf{B}$, and the set,

$$\mathbf{a}_1 \cdot \mathbf{b}_1, \mathbf{a}_1 \cdot \mathbf{b}_2, \cdots \mathbf{a}_1 \cdot \mathbf{b}_n,$$
$$\mathbf{a}_2 \cdot \mathbf{b}_1, \mathbf{a}_2 \cdot \mathbf{b}_2, \cdots \mathbf{a}_2 \cdot \mathbf{b}_n,$$
$$\cdots$$
$$\mathbf{a}_m \cdot \mathbf{b}_1, \mathbf{a}_m \cdot \mathbf{b}_2, \cdots \mathbf{a}_m \cdot \mathbf{b}_n,$$

defines the system $\mathbf{A} \cdot \mathbf{B}$. (10.vi)

No relation of this kind holds universally among the irreducible sets of the systems, **A**, **B**, $\mathbf{A} \vee \mathbf{B}$ and $\mathbf{A} \cdot \mathbf{B}$. It is for this reason that defining sets will play a greater part than irreducible sets in the discussion to follow.

A system has an unlimited number of defining sets, of which the irreducible set is the most exclusive and the system itself is the most inclusive. All the defining sets of a given system, however, have the same entropy, which is that of the irreducible set. This is because every proposition of a defining set which does not belong to the irreducible set implies one of its propositions and therefore, by the theorem (8.iii), contributes nothing to the entropy. The system being one of its own defining sets, we thus have the principle:

The entropy of a system is the entropy of any of its defining sets. (10.vii)

Any exhaustive set of propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, defines a system **A** and its entropy may therefore be denoted, in accordance with this principle, simply by $\eta(\mathbf{A} \mid \mathbf{h})$.

Let us now find an expression for the entropy of $\mathbf{A} \vee \mathbf{B}$, having recourse for this purpose to Eq. (8.4). In this equation, $(\mathbf{a}_i \cdot \mathbf{b} \mid \mathbf{h})$ can be replaced by $(\mathbf{a}_i \mid \mathbf{b} \cdot \mathbf{h})\ (\mathbf{b} \mid \mathbf{h})$ and hence $\ln (\mathbf{a}_i \cdot \mathbf{b} \mid \mathbf{h})$ by $\ln (\mathbf{a}_i \mid \mathbf{b} \cdot \mathbf{h}) + \ln (\mathbf{b} \mid \mathbf{h})$. All the other terms involving conjunctions of **b** can be replaced similarly. The resulting equation is

$$\begin{aligned}
\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m, \mathbf{b} \mid \mathbf{h}) = {} & -\textstyle\sum_i (\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}) \\
& + \textstyle\sum_i \sum_{j>i} (\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{h}) \ln (\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{h}) - \ldots \\
& \pm (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_m \mid \mathbf{h}) \ln (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_m \mid \mathbf{h}) \\
& + (\mathbf{b} \mid \mathbf{h})[\textstyle\sum_i (\mathbf{a}_i \mid \mathbf{b} \cdot \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{b} \cdot \mathbf{h}) \\
& - \textstyle\sum_i \sum_{j>i} (\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{b} \cdot \mathbf{h}) \ln (\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{b} \cdot \mathbf{h}) \\
& + \ldots \mp (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_m \mid \mathbf{b} \cdot \mathbf{h}) \ln (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_m \mid \mathbf{b} \cdot \mathbf{h})] \\
& - (\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{b} \mid \mathbf{h})[1 - \textstyle\sum_i (\mathbf{a}_i \mid \mathbf{b} \cdot \mathbf{h}) + \sum_i \sum_{j>i} (\mathbf{a}_i \cdot \mathbf{a}_j \mid \mathbf{b} \cdot \mathbf{h}) \\
& \qquad - \ldots \pm (\mathbf{a}_1 \cdot \mathbf{a}_2 \cdot \ldots \cdot \mathbf{a}_m \mid \mathbf{b} \cdot \mathbf{h})].
\end{aligned}$$

If we now let $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$ be the exhaustive set of propositions which defines the system **A**, the series outside the brackets in the right-hand member is equal simply to $\eta(\mathbf{A} \mid \mathbf{h})$, the coefficient of $(\mathbf{b} \mid \mathbf{h})$ to $-\eta(\mathbf{A} \mid \mathbf{b}\cdot\mathbf{h})$, and the coefficient of $-(\mathbf{b} \mid \mathbf{h}) \ln (\mathbf{b} \mid \mathbf{h})$ to $1 - (\mathbf{a}_1 \vee \mathbf{a}_2 \vee \ldots \vee \mathbf{a}_m \mid \mathbf{b}\cdot\mathbf{h})$, which is equal to zero. Thus we have

$$\eta(\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m, \mathbf{b} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h}) - (\mathbf{b} \mid \mathbf{h})\eta(\mathbf{A} \mid \mathbf{b}\cdot\mathbf{h}).$$

Any set of propositions which includes an exhaustive set such as $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$ is itself exhaustive and therefore defines a system. Let the system defined by $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m, \mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_k$ be denoted by $\mathbf{C}_k$, where k has values from 0 to n and the set, $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_n$, defines the system **B**. By the equation just given we see that

$$\eta(\mathbf{C}_{k+1} \mid \mathbf{h}) = \eta(\mathbf{C}_k \mid \mathbf{h}) - (\mathbf{b}_{k+1} \mid \mathbf{h})\eta(\mathbf{C}_k \mid \mathbf{b}_{k+1}\cdot\mathbf{h}).$$

From this it may be proved that

$$\begin{aligned}\eta(\mathbf{C}_k \mid \mathbf{h}) = \eta(\mathbf{C}_0 \mid \mathbf{h}) &- \textstyle\sum_i(\mathbf{b}_i \mid \mathbf{h})\eta(\mathbf{C}_0 \mid \mathbf{b}_i\cdot\mathbf{h}) \\ &+ \textstyle\sum_i\sum_{j>i}(\mathbf{b}_i\cdot\mathbf{b}_j \mid \mathbf{h})\eta(\mathbf{C}_0 \mid \mathbf{b}_i\cdot\mathbf{b}_j\cdot\mathbf{h}) - \ldots \\ &\pm (\mathbf{b}_1\cdot\mathbf{b}_2\cdot\ldots\cdot\mathbf{b}_k \mid \mathbf{h})\eta(\mathbf{C}_0 \mid \mathbf{b}_1\cdot\mathbf{b}_2\cdot\ldots\cdot\mathbf{b}_k\cdot\mathbf{h}).\end{aligned}$$

The proof is by a mathematical induction so similar to the one given in Chapter 5 that it would be repetitious to give it here.

From the definition of $\mathbf{C}_k$ it is evident that $\mathbf{C}_0 = \mathbf{A}$ and $\mathbf{C}_n = \mathbf{A} \vee \mathbf{B}$. Thus, by letting k be equal to n in the preceding equation, we have an equation for $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h})$ in terms of the system **A** and the propositions, $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_n$, which define the system **B**. It may be written as

$$\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h}) - \eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h}), \tag{10.1}$$

where $\eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h})$, called a *conditional entropy*[23] or, more specifically, the conditional entropy of the system **A** on the system **B**, is defined by the equation,

$$\begin{aligned}\eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h}) = &\textstyle\sum_i(\mathbf{b}_i \mid \mathbf{h})\eta(\mathbf{A} \mid \mathbf{b}_i\cdot\mathbf{h}) \\ &- \textstyle\sum_i\sum_{j>i}(\mathbf{b}_i\cdot\mathbf{b}_j \mid \mathbf{h})\eta(\mathbf{A} \mid \mathbf{b}_i\cdot\mathbf{b}_j\cdot\mathbf{h}) + \ldots \\ &\pm (\mathbf{b}_1\cdot\mathbf{b}_2\cdot\ldots\cdot\mathbf{b}_n \mid \mathbf{h})\eta(\mathbf{A} \mid \mathbf{b}_1\cdot\mathbf{b}_2\cdot\ldots\cdot\mathbf{b}_n\cdot\mathbf{h}).\end{aligned} \tag{10.2}$$

As the notation implies, the value of $\eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h})$ is determined by the systems **A** and **B** independently of the choice of defining sets. For, according to the theorem (10.vii), the values of $\eta(\mathbf{A} \mid \mathbf{h})$ and $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h})$ are independent of this choice. It follows that so also is their difference, which is equal to $\eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h})$ by Eq. (10.1).

If we exchange **A** and **B** in Eq. (10.1), except in $\mathbf{A} \vee \mathbf{B}$, where their order is immaterial, we obtain the equation,

$$\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{B} \mid \mathbf{h}) - \eta(\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h}). \tag{10.3}$$

If, in this equation or Eq. (10.1), we make **A** and **B** equal, we see that:

The conditional entropy of a system on itself is zero. (10.viii)

Other theorems are obtained by combining other rules of Boolean algebra with these equations. For example, if we replace **A** by $\mathbf{A}\cdot\mathbf{B}$ in Eq. (10.1), we find, because $(\mathbf{A}\cdot\mathbf{B}) \vee \mathbf{B} = \mathbf{B}$, that

$$\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{B}\cdot\mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{h}).$$

We may replace **h** in this equation by $\mathbf{A}\cdot\mathbf{h}$ without making the equation invalid. Doing so, we find, because $\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{B}\cdot\mathbf{A}\cdot\mathbf{h})$ is the conditional entropy of $\mathbf{A}\cdot\mathbf{B}$ on itself and therefore zero, that

$$\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h}) = \eta(\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h}).$$

The exchange of **A** and **B**, except in $\mathbf{A}\cdot\mathbf{B}$, gives

$$\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{B}\cdot\mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h}).$$

Combining this result with the equation just obtained for $\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{h})$, we have the equation,

$$\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{h}). \tag{10.4}$$

By adding to the members of this equation the corresponding members of Eq. (10.1) and by subtracting from them the corresponding members of Eq. (10.3), we obtain two others:

$$\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{h}) + \eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{h}), \quad (10.5)$$

$$\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{h}) - \eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h}). \quad (10.6)$$

By mathematical induction based on Eq. (10.5), it is now fairly simple to obtain expressions for the entropies of conjunctions and disjunctions of any number of systems. The proof will be omitted and only the equations given. Let $\mathbf{A}_1, \mathbf{A}_2, \ldots \mathbf{A}_M$ be any systems. Then

$$\eta(\mathbf{A}_1\cdot\mathbf{A}_2\cdot\ldots\cdot\mathbf{A}_M \mid \mathbf{h}) = \sum_i \eta(\mathbf{A}_i \mid \mathbf{h}) - \sum_i \sum_{j>i} \eta(\mathbf{A}_i \vee \mathbf{A}_j \mid \mathbf{h}) + \sum_i \sum_{j>i} \sum_{k>j} \eta(\mathbf{A}_i \vee \mathbf{A}_j \vee \mathbf{A}_k \mid \mathbf{h}) - \ldots \pm \eta(\mathbf{A}_1 \vee \mathbf{A}_2 \vee \ldots \vee \mathbf{A}_M \mid \mathbf{h}) \quad (10.7)$$

and

$$\eta(\mathbf{A}_1 \vee \mathbf{A}_2 \vee \ldots \vee \mathbf{A}_M \mid \mathbf{h}) = \sum_i \eta(\mathbf{A}_i \mid \mathbf{h}) - \sum_i \sum_{j>i} \eta(\mathbf{A}_i\cdot\mathbf{A}_j \mid \mathbf{h}) + \sum_i \sum_{j>i} \sum_{k>j} \eta(\mathbf{A}_i\cdot\mathbf{A}_j\cdot\mathbf{A}_k \mid \mathbf{h}) - \ldots \pm \eta(\mathbf{A}_1\cdot\mathbf{A}_2\cdot\ldots\cdot\mathbf{A}_M \mid \mathbf{h}). \quad (10.8)$$

11. Entropy and Relevance

There are many arguments concerned only with systems definable by mutually exclusive propositions or, at most, with such systems and others which are Boolean functions of them. In such an argument, let a system **A** be defined by mutually exclusive propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$. Because they are mutually exclusive, no more than one of them can be true and, because they define a system and therefore form an exhaustive set, at least one of them must be true. The set therefore contains one and only one true proposition. As a rule, however, the hypothesis of the argument gives only enough information to assign probabilities to the propositions and not enough to distinguish the true one from the others. Although one of them is true and the rest are false, none is ordinarily certain on the hypothesis and none is impossible.

In the same argument, let the system **B** be defined by mutually exclusive propositions, $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_n$, so that in this set also there is one and only one true proposition. Let us observe, moreover, that the conjunction $\mathbf{a}_i \cdot \mathbf{b}_j$ is true only if $\mathbf{a}_i$ and $\mathbf{b}_j$ are both true, and therefore there is one and only one true proposition among all the conjunctions of a proposition of one set with one of the other. Hence the system $\mathbf{A} \cdot \mathbf{B}$, which these conjunctions define, also is a system defined by mutually exclusive propositions.

If, starting from the hypothesis **h**, we find the true proposition in the set defining **A** and then, with whatever help this discovery may provide, we find the true proposition in the set defining **B**, we shall have found the true proposition in the set defining $\mathbf{A} \cdot \mathbf{B}$. The information to be obtained in the first step, in which we are to find the true proposition among those defining **A**, is measured by the entropy $\eta(\mathbf{A} \mid \mathbf{h})$. If $\mathbf{a}_i$ should be the proposition found true in this step, the additional information to be obtained in the second step, in which we are to find the true proposition among those defining **B**, would be that measured by the entropy $\eta(\mathbf{B} \mid \mathbf{a}_i \cdot \mathbf{h})$. The probability that this will be in fact the required information is $\mathbf{a}_i \mid \mathbf{h}$ and the *a priori* estimate of the information is therefore $\Sigma_i(\mathbf{a}_i \mid \mathbf{h})\eta(\mathbf{B} \mid \mathbf{a}_i \cdot \mathbf{h})$. This is simply the conditional entropy, $\eta(\mathbf{B} \mid \mathbf{A} \cdot \mathbf{h})$, as may be seen by exchanging the roles of **A** and **B** in Eq. (10.2) and making use of the assumption that **A** is defined by mutually exclusive propositions.

Thus the information to be obtained in the two steps is measured by $\eta(\mathbf{A} \mid \mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{A} \cdot \mathbf{h})$ and, since we expect with this information to have found the true proposition among those defining $\mathbf{A} \cdot \mathbf{B}$, we infer that

$$\eta(\mathbf{A} \cdot \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{A} \cdot \mathbf{h}).$$

The equality of $\mathbf{A} \cdot \mathbf{B}$ and $\mathbf{B} \cdot \mathbf{A}$ allows us to exchange **A** and **B** on the right without exchanging them on the left and so to obtain the equation,

$$\eta(\mathbf{A} \cdot \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{B} \mid \mathbf{h}) + \eta(\mathbf{A} \mid \mathbf{B} \cdot \mathbf{h}),$$

which we have already seen as Eq. (10.4).

Equating these two expressions for $\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{h})$, we have

$$\eta(\mathbf{A} \mid \mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h}) = \eta(\mathbf{B} \mid \mathbf{h}) + \eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h}),$$

and we see that the amount of information is the same whether we find first the true proposition among those defining **A** and then the one among those defining **B** or choose the opposite order.

Transposing terms in this equation, we obtain

$$\eta(\mathbf{A} \mid \mathbf{h}) - \eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h}) = \eta(\mathbf{B} \mid \mathbf{h}) - \eta(\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h}).$$

Comparison with Eq. (10.1) shows that each of these expressions is equal to the entropy, $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h})$, of the disjunction, whence we have

$$\eta(\mathbf{A} \mid \mathbf{h}) = \eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) + \eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h})$$

and

$$\eta(\mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h}).$$

The term $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h})$, common on the right to both of these equations, measures the information to be obtained whether we are finding the true proposition among those defining **A** or **B**. The additional information to be obtained is different in the two cases but is measured in either by one of the conditional entropies. If we are to find the true proposition among those defining **A**, we require the additional information measured by $\eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h})$ but, if among those defining **B**, we require that measured by $\eta(\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h})$.

From another point of view, $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h})$, considered as the difference, $\eta(\mathbf{A} \mid \mathbf{h}) - \eta(\mathbf{A} \mid \mathbf{B}\cdot\mathbf{h})$, measures the information relevant to the discovery of the true proposition among those defining **A** which we expect to obtain from the corresponding discovery in respect to **B**. More briefly, it can be said to measure the relevance of **B** to **A**. Alternatively, as the difference, $\eta(\mathbf{B} \mid \mathbf{h}) - \eta(\mathbf{B} \mid \mathbf{A}\cdot\mathbf{h})$, it measures the relevance of **A** to **B**. It measures, therefore, the mutual relevance of the two systems.

If any of the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, is certain on the hy-

pothesis $\mathbf{b}_j \cdot \mathbf{h}$ or, in other words, if $\mathbf{b}_j$ implies one of these propositions, then $\eta(\mathbf{A} \mid \mathbf{b}_j \cdot \mathbf{h}) = 0$ by the theorem (8.iv). Consequently, if each of the propositions, $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_n$, implies one of the set defining $\mathbf{A}$, $\eta(\mathbf{A} \mid \mathbf{B} \cdot \mathbf{h}) = 0$ and $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h})$. No system, not even $\mathbf{A}$ itself, can be more relevant to $\mathbf{A}$ than $\mathbf{B}$ is in this case. Indeed we may note that $\eta(\mathbf{A} \mid \mathbf{h}) = \eta(\mathbf{A} \vee \mathbf{A} \mid \mathbf{h})$ and take the entropy of a single system $\mathbf{A}$ as measuring its relevance to itself.

At the other extreme is the case in which every proposition of the set defining either system is irrelevant to every one of the set defining the other. For the sake of brevity it is convenient to say in this case that the two systems are mutually irrelevant, omitting reference to the defining sets. However, it should be noticed that this is only a convenient phrase, which must not be taken to mean that every proposition belonging to either entire system is irrelevant to every one belonging to the other. The latter condition is indeed impossible. For, if $\mathbf{i}$ is any proposition of one system and $\mathbf{j}$ any proposition of the other, the conjunction, $\mathbf{i} \cdot \mathbf{j}$, because it implies both $\mathbf{i}$ and $\mathbf{j}$, is included in both systems and is obviously relevant to propositions of both. With this explanation of the irrelevance of systems, we may say, if $\mathbf{A}$ and $\mathbf{B}$ are mutually irrelevant, that

$$\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = 0.$$

To see this, we obtain, from Eq. (10.5), the equation,

$$\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{h}) - \eta(\mathbf{A} \cdot \mathbf{B} \mid \mathbf{h}), \qquad (11.1)$$

which can be written

$$\begin{aligned} \eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = & -\sum_i (\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}) \\ & - \sum_j (\mathbf{b}_j \mid \mathbf{h}) \ln (\mathbf{b}_j \mid \mathbf{h}) \\ & + \sum_i \sum_j (\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h}) \ln (\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h}), \end{aligned} \qquad (11.2)$$

because each of the systems, $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A} \cdot \mathbf{B}$, is defined by a set of mutually exclusive propositions.

The three summations on the right in this equation can be combined. For

$$\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h} = (\mathbf{a}_i \mid \mathbf{h})(\mathbf{b}_j \mid \mathbf{a}_i \cdot \mathbf{h}),$$

whence, summing over all values of j and noting that $\Sigma_j(\mathbf{b}_j \mid \mathbf{a}_i \cdot \mathbf{h}) = 1$, we find that

$$\mathbf{a}_i \mid \mathbf{h} = \sum_j(\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h}).$$

Similarly,

$$\mathbf{b}_j \mid \mathbf{h} = \sum_i(\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h}).$$

Substituting these expressions for $\mathbf{a}_i \mid \mathbf{h}$ and $\mathbf{b}_j \mid \mathbf{h}$ in Eq. (11.2), we obtain

$$\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \sum_i\sum_j[\ln (\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h}) - \ln (\mathbf{a}_i \mid \mathbf{h}) - \ln (\mathbf{b}_j \mid \mathbf{h})](\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h}).$$

When $\mathbf{A}$ and $\mathbf{B}$ are mutually irrelevant,

$$\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h} = (\mathbf{a}_i \mid \mathbf{h})(\mathbf{b}_j \mid \mathbf{h})$$

and hence

$$\ln (\mathbf{a}_i \cdot \mathbf{b}_j \mid \mathbf{h}) = \ln (\mathbf{a}_i \mid \mathbf{h}) + \ln (\mathbf{b}_j \mid \mathbf{h})$$

for all values of i and j. Thus $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = 0$.

As might be expected, this is the minimum value. To prove that it is so, let the probabilities be infinitesimally varied. For the resulting variations in the entropies, we have

$$\delta\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \delta\eta(\mathbf{A} \mid \mathbf{h}) + \delta\eta(\mathbf{B} \mid \mathbf{h}) - \delta\eta(\mathbf{A} \cdot \mathbf{B} \mid \mathbf{h}).$$

By differentiating the members of the equation,

$$\eta(\mathbf{A} \mid \mathbf{h}) = -\sum_i(\mathbf{a}_i \mid \mathbf{h}) \ln (\mathbf{a}_i \mid \mathbf{h}),$$

we obtain

$$\delta\eta(\mathbf{A} \mid \mathbf{h}) = -\sum_i[\ln (\mathbf{a}_i \mid \mathbf{h}) + 1]\delta(\mathbf{a}_i \mid \mathbf{h}).$$

$\Sigma_i\delta(\mathbf{a}_i \mid \mathbf{h}) = 0$ because $\Sigma_i(\mathbf{a}_i \mid \mathbf{h}) = 1$ and thus we have simply

$$\delta\eta(\mathbf{A} \mid \mathbf{h}) = -\sum_i \ln(\mathbf{a}_i \mid \mathbf{h})\delta(\mathbf{a}_i \mid \mathbf{h}).$$

Substituting this and similar expressions for $\delta\eta(\mathbf{B} \mid \mathbf{h})$ and $\delta\eta(\mathbf{A}\cdot\mathbf{B} \mid \mathbf{h})$ in Eq. (11.2), we see that

$$\delta\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = -\sum_i \ln(\mathbf{a}_i \mid \mathbf{h})\delta(\mathbf{a}_i \mid \mathbf{h}) - \sum_j \ln(\mathbf{b}_j \mid \mathbf{h})\delta(\mathbf{b}_j \mid \mathbf{h}) + \sum_i\sum_j \ln(\mathbf{a}_i\cdot\mathbf{b}_j \mid \mathbf{h})\delta(\mathbf{a}_i\cdot\mathbf{b}_j \mid \mathbf{h}).$$

In this equation, as in Eq. (11.2), the three summations can be combined, because $\delta(\mathbf{a}_i \mid \mathbf{h}) = \Sigma_j\, \delta(\mathbf{a}_i\cdot\mathbf{b}_j \mid \mathbf{h})$ and $\delta(\mathbf{b}_j \mid \mathbf{h}) = \Sigma_i\, \delta(\mathbf{a}_i\cdot\mathbf{b}_j \mid \mathbf{h})$. Thus

$$\delta\eta(\mathbf{A} \vee \mathrm{B} \mid \mathbf{h}) = \sum_i\sum_j[\ln(\mathbf{a}_i\cdot\mathbf{b}_j \mid \mathbf{h}) - \ln(\mathbf{a}_i \mid \mathbf{h}) - \ln(\mathbf{b}_j \mid \mathbf{h})]\delta(\mathbf{a}_i\cdot\mathbf{b}_j \mid \mathbf{h}).$$

When **A** and **B** are mutually irrelevant, the right-hand member vanishes and $\delta\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = 0$ for all possible variations of the probabilities. Moreover it is only when they are mutually irrelevant that this condition is satisfied. Therefore $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h})$ has no maximum or minimum value except zero. If zero were its maximum value, all the other values would be negative, but this is obviously untrue, since $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h})$ when $\mathbf{B} = \mathbf{A}$. Therefore zero is the minimum value and we have the theorem:

> If each of two systems is definable by a set of mutually exclusive propositions, the entropy of their disjunction is zero if they are mutually irrelevant and is otherwise positive. (11.i)

This theorem justifies a familiar type of inquiry, one in which the subject is chosen not so much for its intrinsic interest as for its relevance to another subject, more immediately interesting but less accessible to investigation. Let us identify the subject of principal interest with the system **A** and suppose that we should like to know the true proposition in the set, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_m$, but we are obliged to rely on indirect evidence. We identify the secondary subject with the system **B**, and we propose to find the true proposition in the set, $\mathbf{b}, \mathbf{b}_2, \ldots \mathbf{b}_n$, for whatever bearing its discovery may have on the primary subject. We expect, unless

there is complete irrelevance between the two subjects, that this information will be helpful, at least to some extent. We expect it to diminish rather than increase our uncertainty about the primary subject. The symbolic expression of this expectation is the inequality, $\eta(\mathbf{A} \mid \mathbf{B} \cdot \mathbf{h}) \leq \eta(\mathbf{A} \mid \mathbf{h})$, which is equivalent to $\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) \geq 0$, the symbolic expression of the theorem.

The expectation is reasonable but, like any other which is based on merely probable inference, it is liable to disappointment in the event. Such disappointments are common enough to make it a familiar remark that "we know less now than when we began."

For an artificial but simple example, let the hypothesis **h** assert that a blindfolded man puts both hands into a bag containing one white ball and two black balls and takes out one ball in each hand. Let us imagine that for some reason we are interested primarily in the color of the ball in his right hand but we can learn the color only of the ball in his left.

Information that the ball in his left hand is white will leave no uncertainty at all about the color of the ball in his right hand, because there was only one white ball in the box. By contrast, information that the ball in his left hand is black will increase the uncertainty about the color of the ball in his right hand, because it will equalize the probabilities of the two colors and thus produce an uncertainty as great as any possible with only two alternatives. Moreover the increase in the uncertainty is more probable than the decrease, because the chances are two to one that the man has a black ball in his left hand.

To discuss this example in formal terms, let $\mathbf{a}_1$ assert that the ball in his right hand is white, $\mathbf{a}_2$ that it is black, $\mathbf{b}_1$ that the ball in his left hand is white, $\mathbf{b}_2$ that it is black. Then

$$\mathbf{a}_1 \mid \mathbf{h} = \tfrac{1}{3}, \qquad \mathbf{a}_2 \mid \mathbf{h} = \tfrac{2}{3}$$

and

$$\eta(\mathbf{A} \mid \mathbf{h}) = -(\mathbf{a}_1 \mid \mathbf{h}) \ln (\mathbf{a}_1 \mid \mathbf{h}) - (\mathbf{a}_2 \mid \mathbf{h}) \ln (\mathbf{a}_2 \mid \mathbf{h}) = \ln 3 - \tfrac{2}{3} \ln 2,$$

whereas

$$\eta(\mathbf{A} \mid \mathbf{b}_1 \cdot \mathbf{h}) = 0 \text{ and } \eta(\mathbf{A} \mid \mathbf{b}_2 \cdot \mathbf{h}) = \ln 2.$$

Also

$$\mathbf{b}_1 \mid \mathbf{h} = \tfrac{1}{3}, \qquad \mathbf{b}_2 \mid \mathbf{h} = \tfrac{2}{3}$$

and

$$\eta(\mathbf{A} \mid \mathbf{B} \cdot \mathbf{h}) = (\mathbf{b}_1 \mid \mathbf{h})\eta(\mathbf{A} \mid \mathbf{b}_1 \cdot \mathbf{h}) + (\mathbf{b}_2 \mid \mathbf{h})\eta(\mathbf{A} \mid \mathbf{b}_2 \cdot \mathbf{h}) = \tfrac{2}{3} \ln 2.$$

Therefore

$$\eta(\mathbf{A} \vee \mathrm{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h}) - \eta(\mathbf{A} \mid \mathbf{B} \cdot \mathbf{h})$$
$$= \ln 3 - \tfrac{4}{3} \ln 2 = \tfrac{1}{3} \ln {}^{27}\!/_{16} > 0.$$

The uncertainty about the color of the ball in the man's right hand is measured in each case by the entropy of **A**. It is measured by $\eta(\mathbf{A} \mid \mathbf{h})$ if the color of the ball in his left hand is unknown, by $\eta(\mathbf{A} \mid \mathbf{b}_1 \cdot \mathbf{h})$ if the ball in his left hand is known to be white, and by $\eta(\mathbf{A} \mid \mathbf{b}_2 \cdot \mathbf{h})$ if it is known to be black. In the former case the entropy of **A** is decreased by the additional information, whereas in the latter case it is increased. Although the decrease is only half as probable as the increase, it is more than twice as great, and it therefore counts for more in the expectation, as is shown by the fact that $\eta(\mathbf{A} \mid \mathbf{B} \cdot \mathbf{h})$ is less than $\eta(\mathbf{A} \mid \mathbf{h})$.

From Eq. (11.1) and the theorem (11.i), there follows directly another theorem:

> If each of two systems is definable by a set of mutually exclusive propositions, the entropy of their conjunction is equal to the sum of their entropies if the systems are mutually irrelevant, and otherwise is less.[24] (11.ii)

12. A Remark on Chance

The essentials of chance, or, at any rate, the characteristics essential to its discussion in this essay, are two in number. One is the coincidence of two or more events or, more exactly, the con-

junction of two or more systems of propositions. The other is a limitation of knowledge, in consequence of which the events or systems are mutually irrelevant.

Both features are admirably illustrated by a stanza in one of Sir Walter Scott's poems:

> "O, Richard! if my brother died,
> 'Twas but a fatal chance,
> For darkling was the battle tried,
> And fortune sped the lance."[25]

In these lines a lady is trying to console her husband, who has, as they believe, killed her brother in combat. The coincidence of events, literal and physical in this example, is between the point of her husband's lance and a vital part of her brother's person. The impediment to knowledge, equally literal and physical, is the darkness in which the battle was fought, and the irrelevance it imposed on the events is implied in the words, "fortune sped the lance." The implication is that, because her husband could not see what he was doing, the fact that he aimed his lance in a certain direction had no relation to the fact that her brother, at that instant, was in the way and vulnerable. If this appears to involve the lady in some exaggeration, it is no more than would readily be allowed under the circumstances to the heroine of a romantic ballad.

If it was by chance, in this example, that the brother died, it would have been by chance also if he had lived.[26] In a more familiar example, if it is by chance that a coin falls heads, it is equally by chance that it falls tails. Although it is convenient in ordinary speech to associate chance with the actual event, it is truer to the concept to relate it to a set of possible alternatives, of which the actual event is one. The set may comprise only two alternatives, such as life and death or heads and tails, or it may include more, but in any case it is exhaustive and the alternatives are mutually exclusive. Hence the set of propositions, each of which asserts one of the alternatives, defines a system of the kind considered in the chapter before this one. It is possible, there-

fore, and reasonable to associate chance with systems of propositions rather than with single propositions or events. Indeed such an association is necessarily implied if, as we have just supposed, an essential feature of chance is irrelevance. For, as was pointed out at the end of Chapter 4, if two propositions are mutually irrelevant, each is irrelevant to the contradictory of the other and the contradictories are also mutually irrelevant. Thus irrelevance is a relation between a pair, at least, of mutually exclusive propositions and another such pair. It is a relation, therefore, between systems, because each pair, being exhaustive, defines a system. Of course there can be irrelevance also between systems defined by more than two propositions.

It may still be questioned whether irrelevance is an invariable characteristic of chance, and indeed it is not explicitly present in every case. There seems, however, to be at least an implication of it in every occurrence attributed to chance by common usage. For example, it will sometimes be said, "That was only chance," when someone has performed an astonishing feat. Although the assertion of irrelevance is not explicit here, it becomes more evident if the speaker adds in explanation, "I doubt if he could do it again." The meaning of the added remark is that the first performance of the feat, if it were a proof of skill, would create a presumption of success at a second trial, but, if it were a matter only of chance, there would be no such presumption and success at the second trial would be as unexpected as it was at the first. The expression of doubt in the second remark makes explicit an implication of irrelevance already present in the first.

Chance may therefore be described as a condition under which two or more systems of propositions are mutually irrelevant. If **A** and **B** are the systems, their mutual irrelevance is expressed by either of the equations,

$$\eta(\mathbf{A} \vee \mathbf{B} \mid \mathbf{h}) = 0$$

or

$$\eta(\mathbf{A} \cdot \mathbf{B} \mid \mathbf{h}) = \eta(\mathbf{A} \mid \mathbf{h}) + \eta(\mathbf{B} \mid \mathbf{h}).$$

This description is still incomplete, because irrelevance is not all we mean when we speak of chance. What else we mean is hard to say precisely, but we seem always to associate chance with an irrelevance which is not merely present in the argument but is produced by an impediment to knowledge inseparable from the circumstances on which the argument rests. The circumstances may be brought about intentionally, as they are in games of chance and in many statistical studies. Thus cards are shuffled until all knowledge of their prior arrangement becomes irrelevant to any inference about the order in which they will be dealt afterwards. Or, like the darkness in the ballad, the circumstances may be those of time and place. Or, again, they may be inherent in the nature of things, as when we call radioactive decay a matter of chance and mean that no possible observation will enable us to say in what order the atoms of a radioactive element will disintegrate and no method exists for separating those which will disintegrate early from the others which will outlast them.

It has often been said that when we speak of chance, sometimes of "blind chance", we are only giving an external embodiment to our own ignorance.[27] This may be true, but it should be noted that we do not ascribe to chance all the coincidences of whose causes we are ignorant, but only some of them. Moreover we conceive our ignorance in these cases not as altogether private and subjective but rather as something which the given situation imposes on us and would impose equally on anyone else who might be there in our stead.

III

Expectation

13. Expectations and Deviations

The idea of *expectation* began in gambling and may still be most easily explained by that example. Consider a prize of value x put up in a lottery of W chances. The holder of a single chance is said to have an expectation equal to x/W. In a lottery in which the prices of all the chances are pooled to make the prize, the expectation is the price of one chance.

Suppose now that, instead of a single prize, there are numerous prizes of different values: w_1 of value x_1, w_2 of value x_2, and so on; so that the holder of a single chance has the probability, w_r/W, of winning a prize of value x_r. His expectation is said to be $\Sigma_r x_r(w_r/W)$. If each of the W chances is sold at this price, the total receipts will be $\Sigma_r x_r w_r$ and will thus be just enough to pay for all the prizes.

The definition is easily generalized from this example. Let x be a quantity which, on the hypothesis $\mathbf{h}$, can have any one of a number of values. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be an exhaustive set of mutually exclusive propositions such that x has the value x_r if $\mathbf{x}_r$ is true. Let the expectation of x on the hypothesis $\mathbf{h}$ be denoted by $\langle x \mid \mathbf{h} \rangle$. In analogy with the example of the lottery, it is defined by the equation,

$$\langle x \mid \mathbf{h} \rangle = \sum_r x_r(\mathbf{x}_r \mid \mathbf{h}). \qquad (13.1)$$

If, in the set, $\mathbf{x}_1, \mathbf{x}_2, \ldots$, there is a proposition which ascribes to x the value zero, its probability obviously contributes nothing

to the expectation. It is convenient, however, to consider this proposition, when it has any probability, as always included in the set, so that we may employ theorems which are valid only for exhaustive sets and may refer on occasion to the system $\mathbf{X}$, which the propositions, if they form an exhaustive set, define.

A quantity which has only one value possible on the hypothesis $\mathbf{h}$ is a constant in every argument from that hypothesis, and the proposition which asserts that value is certain. If C is any such quantity, it follows immediately from Eq. (13.1) that

$$\langle C \mid \mathbf{h} \rangle = C.$$

If A is another quantity constant on the hypothesis $\mathbf{h}$ and x is any variable, Ax has the value Ax_r when $\mathbf{x}_r$ is true. Hence, by Eq. (13.1),

$$\langle Ax \mid \mathbf{h} \rangle = A \langle x \mid \mathbf{h} \rangle.$$

Now let y be a quantity to which propositions, $\mathbf{y}_1, \mathbf{y}_2, \ldots$, ascribe values $y_1, y_2, \ldots$ Then $x + y$ has the value $x_r + y_s$ when $\mathbf{x}_r \cdot \mathbf{y}_s$ is true and, by Eq. (13.1),

$$\langle (x + y) \mid \mathbf{h} \rangle = \sum_r \sum_s (x_r + y_s)(\mathbf{x}_r \cdot \mathbf{y}_s \mid \mathbf{h}).$$

This may be written

$$\langle (x + y) \mid \mathbf{h} \rangle = \sum_r [x_r (\mathbf{x}_r \mid \mathbf{h}) \sum_s (\mathbf{y}_s \mid \mathbf{x}_r \cdot \mathbf{h})] + \sum_s [y_s (\mathbf{y}_s \mid \mathbf{h}) \sum_r (\mathbf{x}_r \mid \mathbf{y}_s \cdot \mathbf{h})].$$

The propositions, $\mathbf{y}_1, \mathbf{y}_2, \ldots$, are mutually exclusive and form an exhaustive set. Therefore $\Sigma_s (\mathbf{y}_s \mid \mathbf{x}_r \cdot \mathbf{h}) = 1$ and, similarly, $\Sigma_r (\mathbf{x}_r \mid \mathbf{y}_s \cdot \mathbf{h}) = 1$. Hence

$$\langle (x + y) \mid \mathbf{h} \rangle = \sum_r x_r (\mathbf{x}_r \mid \mathbf{h}) + \sum_s y_s (\mathbf{y}_s \mid \mathbf{h}) = \langle x \mid \mathbf{h} \rangle + \langle y \mid \mathbf{h} \rangle.$$

Thus the expectation of the sum of two quantities is equal to the sum of their expectations.

By combining the three results just obtained, we see that

$$\langle (Ax + By + C) \mid \mathbf{h} \rangle = A \langle x \mid \mathbf{h} \rangle + B \langle y \mid \mathbf{h} \rangle + C,$$

where x and y are any quantities and A, B and C are any constants. More generally, we have the theorem,

The expectation of a linear function of any quantities is equal to the same linear function of the expectations of the quantities. (13.i)

When all the expectations involved in a given discussion are reckoned on the same hypothesis, the symbol for the hypothesis may, without confusion, be omitted from the symbols for the expectations. Thus, with the omission of the symbol $\mathbf{h}$, the preceding equation may be written in the form,

$$\langle Ax + By + C \rangle = A\langle x \rangle + B\langle y \rangle + C.$$

The simpler notation will be used henceforth except when reference to the hypothesis is necessary in order to avoid ambiguity.

For functions which are not linear, there is no theorem corresponding to (13.i). For example, the expectation of the product of two quantities is not, in general, equal to the product of their expectations. The expectation of the product xy is given by

$$\langle xy \rangle = \sum_r \sum_s x_r y_s (\mathbf{x}_r \cdot \mathbf{y}_s \mid \mathbf{h}),$$

whereas the product of the expectations is given by

$$\langle x \rangle \langle y \rangle = \sum_r x_r (\mathbf{x}_r \mid \mathbf{h}) \sum_s y_s (\mathbf{y}_s \mid \mathbf{h}) = \sum_r \sum_s x_r y_s (\mathbf{x}_r \mid \mathbf{h})(\mathbf{y}_s \mid \mathbf{h}).$$

The most frequently encountered case in which these two expressions are equal is that in which every proposition of the set, $\mathbf{x}_1, \mathbf{x}_2, \ldots$, is irrelevant to every one of the set, $\mathbf{y}_1, \mathbf{y}_2, \ldots$, or, as it may be said more briefly, the systems $\mathbf{X}$ and $\mathbf{Y}$ are mutually irrelevant. In this case, $\mathbf{x}_r \cdot \mathbf{y}_s \mid \mathbf{h} = (\mathbf{x}_r \mid \mathbf{h})(\mathbf{y}_s \mid \mathbf{h})$ and the expectation of the product is given by the same expression as the product of the expectations. The case in which one of the quantities, x or y, is constant and the proposition which states its value is therefore certain, is a special instance of this irrelevance, according to the discussion at the end of Chapter 3.

The difference of any quantity from its expectation, for example, $x - \langle x \rangle$, is called the *deviation* of the quantity. The

product of the deviations of x and y is given by

$$(x - \langle x \rangle)(y - \langle y \rangle) = xy - x\langle y \rangle - \langle x \rangle y + \langle x \rangle \langle y \rangle$$

and is therefore a linear function of the quantities, xy, x and y. Hence it follows, by the theorem (13.i), that

$$\langle (x - \langle x \rangle)(y - \langle y \rangle) \rangle = \langle xy \rangle - \langle x \rangle \langle y \rangle. \qquad (13.2)$$

If the deviations of x, whether positive or negative, are predominantly associated with deviations of y of the same sign, it follows from this equation that $\langle xy \rangle$ is greater than $\langle x \rangle \langle y \rangle$, whereas, with the opposite association of signs, it is less. In the case of mutual irrelevance, and exceptionally in other cases, $\langle xy \rangle$ and $\langle x \rangle \langle y \rangle$ are equal.

When x and y are the same quantity, the preceding equation becomes

$$\langle (x - \langle x \rangle)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2. \qquad (13.3)$$

The left-hand member of this equation can not be negative and it follows therefore that the expectation of the square of a quantity can not be less than the square of its expectation. They are equal only in the extreme case in which the quantity is constant, its expectation is equal to its only possible value and its deviation is therefore zero. A very small value of $\langle (x - \langle x \rangle)^2 \rangle$ indicates that values of x much different from $\langle x \rangle$ are very improbable. On the other hand, if the more probable values of x are widely different from one another and hence from $\langle x \rangle$, the probable values of $(x - \langle x \rangle)^2$ are large and so also therefore is $\langle (x - \langle x \rangle)^2 \rangle$. The extreme example of this kind is that in which the only possible values of x are two constants, C and $-C$, and these are equally probable. In this case, $\langle x \rangle = 0$ but $\langle (x - \langle x \rangle)^2 \rangle = C^2$. It is evident from this discussion that the expectation of the square of the deviation of a quantity is a convenient measure of the dispersion of its probable values. It is not a very discriminating one, in that it tells us nothing about the probabilities of single values. but it is often adequate, especially when it is small

and our only need is to be assured that the dispersion is within tolerable limits.

An equation useful as a lemma is

$$\sum_r x_r(\mathbf{x}_r \mid \mathbf{h})(\mathbf{a} \mid \mathbf{x}_r \cdot \mathbf{h}) = (\mathbf{a} \mid \mathbf{h})\langle x \mid \mathbf{a} \cdot \mathbf{h}\rangle, \qquad (13.4)$$

where $\mathbf{a}$ is an arbitrary proposition.

This equation is easily proved. We have

$$(\mathbf{x}_r \mid \mathbf{h})(\mathbf{a} \mid \mathbf{x}_r \cdot \mathbf{h}) = (\mathbf{a} \mid \mathbf{h})(\mathbf{x}_r \mid \mathbf{a} \cdot \mathbf{h}),$$

since these are both expressions for $\mathbf{x}_r \cdot \mathbf{a} \mid \mathbf{h}$. Multiplying by x_r and summing with respect to r, we immediately obtain the lemma.

If, in this lemma, we replace $\mathbf{a}$ by each in turn of an exhaustive set of propositions, $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_n$, and sum over all of them, we obtain

$$\sum_r [x_r(\mathbf{x}_r \mid \mathbf{h}) \sum_i (\mathbf{b}_i \mid \mathbf{x}_r \cdot \mathbf{h})] = \sum_i (\mathbf{b}_i \mid \mathbf{h})\langle x \mid \mathbf{b}_i \cdot \mathbf{h}\rangle. \qquad (13.5)$$

If the propositions of the set are mutually exclusive, $\Sigma_i(\mathbf{b}_i \mid \mathbf{x}_r \cdot \mathbf{h}) = 1$ and the left-hand member is equal simply to $\langle x \mid \mathbf{h}\rangle$. In this case, therefore,

$$\langle x \mid \mathbf{h}\rangle = \sum_i (\mathbf{b}_i \mid \mathbf{h})\langle x \mid \mathbf{b}_i \cdot \mathbf{h}\rangle. \qquad (13.6)$$

This is a special case of a more general equation, valid for any exhaustive set of propositions, whether or not they are mutually exclusive. It is

$$\begin{aligned}\langle x \mid \mathbf{h}\rangle &= \sum_i (\mathbf{b}_i \mid \mathbf{h})\langle x \mid \mathbf{b}_i \cdot \mathbf{h}\rangle - \sum_i \sum_{j>i} (\mathbf{b}_i \cdot \mathbf{b}_j \mid \mathbf{h})\langle x \mid \mathbf{b}_i \cdot \mathbf{b}_j \cdot \mathbf{h}\rangle \\ &+ \sum_i \sum_{j>i} \sum_{k>j} (\mathbf{b}_i \cdot \mathbf{b}_j \cdot \mathbf{b}_k \mid \mathbf{h})\langle x \mid \mathbf{b}_i \cdot \mathbf{b}_j \cdot \mathbf{b}_k \cdot \mathbf{h}\rangle - \ldots \\ &\pm (\mathbf{b}_1 \cdot \mathbf{b}_2 \cdot \ldots \cdot \mathbf{b}_n \mid \mathbf{h})\langle x \mid \mathbf{b}_1 \cdot \mathbf{b}_2 \cdot \ldots \cdot \mathbf{b}_n \cdot \mathbf{h}\rangle. \qquad (13.7)\end{aligned}$$

To prove this equation, we replace $\mathbf{a}$ in Eq. (13.4) successively by $\mathbf{b}_i \cdot \mathbf{b}_j$, $\mathbf{b}_i \cdot \mathbf{b}_j \cdot \mathbf{b}_k, \ldots$ and sum over all the different combinations of unequal values of $i, j, k, \ldots$ The members of the equations so obtained are alternately subtracted from and added to those of Eq. (13.5). In this way we obtain an equation of which the right-

hand member is the same as that of Eq. (13.7) and the left-hand member is

$$\sum_r \{x_r(\mathbf{x}_r \mid \mathbf{h})[\sum_i(\mathbf{b}_i \mid \mathbf{x}_r \cdot \mathbf{h}) - \sum_i \sum_{j>i}(\mathbf{b}_i \cdot \mathbf{b}_j \mid \mathbf{x}_r \cdot \mathbf{h}) + \ldots \pm (\mathbf{b}_1 \cdot \mathbf{b}_2 \cdot \ldots \cdot \mathbf{b}_n \mid \mathbf{x}_r \cdot \mathbf{h})]\}.$$

The bracketed quantity, by which $x_r(\mathbf{x}_r \mid \mathbf{h})$ is multiplied, is equal to $\mathbf{b}_1 \vee \mathbf{b}_2 \vee \ldots \vee \mathbf{b}_n \mid \mathbf{x}_r \cdot \mathbf{h}$ and hence to 1, because the set, $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_n$, is exhaustive. Thus the whole expression is reduced to $\Sigma_r\, x_r(\mathbf{x}_r \mid \mathbf{h})$, which is equal to $\langle x \mid \mathbf{h} \rangle$, and thus Eq. (13.7) is proved.

There is an evident likeness between this equation and Eq. (10.2), which defines the conditional entropy.

14. The Expectation of Numbers

There are times when we have a statistical interest, rather than an interest in detail, in respect to some group of propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M$. It may be more feasible or it may be more urgent to concern ourselves with the number of true propositions in the group than with the question as to which are true and which false. For example, a body of citizens may be urging their City Council to enact some ordinance and $\mathbf{a}_I$ may be the proposition, "The Councilman for the Ith District will vote for the ordinance." The citizens will be more interested in the prospect of a majority vote than in the composition of the majority. Or a public health official, trying to control an epidemic, will be obliged to forecast the incidence of the disease. These are examples of the expectation of numbers. In the ordinary case, as in these examples, the propositions have some similarity of meaning which makes it natural to associate them as members of one group. There are statistical theorems, however, which do not depend for their proof on the nature of any such resemblance or even on its existence but, on the contrary, are valid for propositions assembled in any way, even a capricious one.

Let $\mathbf{h}$ denote the hypothesis common to all the calculations and let m be the number of true propositions in the group, $\mathbf{a}_1$, $\mathbf{a}_2, \ldots \mathbf{a}_M$. If the propositions were mutually exclusive, m could not be greater than 1 and, if they formed an exhaustive set, it could not be less, but neither assumption is to be made here and all the integers from 0 to M are possible values of m. The first theorem to be proved is

$$\langle m \mid \mathbf{h} \rangle = \sum_I (\mathbf{a}_I \mid \mathbf{h}), \tag{14.1}$$

where the summation is over all the propositions in the group.

The proof is by a mathematical induction, in which the expectation of the number of true propositions in the original group, $\mathbf{a}_1$, $\mathbf{a}_2, \ldots \mathbf{a}_M$, is compared with the like expectation in the group, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M, \mathbf{a}_{M+1}$, identical with the first except that it includes one more proposition, $\mathbf{a}_{M+1}$. Let us denote the number of true propositions in the first group by m_M and in the second group by m_{M+1}, and let m_{M+1} be substituted for x in Eq. (13.6). Then the propositions, $\mathbf{a}_{M+1}$ and $\sim\mathbf{a}_{M+1}$, making, as they do, an exhaustive set of mutually exclusive propositions, may replace the set, $\mathbf{b}_1$, $\mathbf{b}_2, \ldots \mathbf{b}_n$, in the same equation. With these substitutions we obtain:

$$\langle m_{M+1} \mid \mathbf{h} \rangle = (\mathbf{a}_{M+1} \mid \mathbf{h}) \langle m_{M+1} \mid \mathbf{a}_{M+1} \cdot \mathbf{h} \rangle + (\sim\mathbf{a}_{M+1} \mid \mathbf{h}) \langle m_{M+1} \mid \sim\mathbf{a}_{M+1} \cdot \mathbf{h} \rangle. \tag{14.2}$$

If $\mathbf{a}_{M+1}$ is true, there is one more true proposition in the group which includes it than in the group which excludes it. The expectation of m_{M+1} on the hypothesis $\mathbf{a}_{M+1} \cdot \mathbf{h}$ is therefore greater by 1 than that of m_M on the same hypothesis. Thus

$$\langle m_{M+1} \mid \mathbf{a}_{M+1} \cdot \mathbf{h} \rangle = \langle m_M \mid \mathbf{a}_{M+1} \cdot \mathbf{h} \rangle + 1.$$

If, on the other hand, $\mathbf{a}_{M+1}$ is false, the number of true propositions is the same in both groups and hence

$$\langle m_{M+1} \mid \sim\mathbf{a}_{M+1} \cdot \mathbf{h} \rangle = \langle m_M \mid \sim\mathbf{a}_{M+1} \cdot \mathbf{h} \rangle.$$

Substituting these expressions in Eq. (14.2), we obtain

$$\langle m_{M+1} \mid \mathbf{h} \rangle = [(\mathbf{a}_{M+1} \mid \mathbf{h}) \langle m_M \mid \mathbf{a}_{M+1} \cdot \mathbf{h} \rangle + (\sim\mathbf{a}_{M+1} \mid \mathbf{h}) \langle m_M \mid \sim\mathbf{a}_{M+1} \cdot \mathbf{h} \rangle] + (\mathbf{a}_{M+1} \mid \mathbf{h}).$$

By the use of Eq. (13.6) again, we see that the expression in brackets is equal to $\langle m_M \mid \mathbf{h} \rangle$ and thus we find that

$$\langle m_{M+1} \mid \mathbf{h} \rangle = \langle m_M \mid \mathbf{h} \rangle + (\mathbf{a}_{M+1} \mid \mathbf{h}).$$

Assuming now, for the sake of the induction, that Eq. (14.1) holds for the group of M propositions, we have provisionally

$$\langle m_M \mid \mathbf{h} \rangle = \sum_{I=1}^{M} (\mathbf{a}_I \mid \mathbf{h})$$

and therefore, by the result just obtained,

$$\langle m_{M+1} \mid \mathbf{h} \rangle = \sum_{I=1}^{M} (\mathbf{a}_I \mid \mathbf{h}) + (\mathbf{a}_{M+1} \mid \mathbf{h}) = \sum_{I=1}^{M+1} (\mathbf{a}_I \mid \mathbf{h}).$$

Thus, if Eq. (14.1) holds for one value of M, it is proved for the next higher value and therefore for all higher values. When $M = 1$, there is the probability $\mathbf{a}_1 \mid \mathbf{h}$ that $m = 1$ and the probability $\sim\mathbf{a}_1 \mid \mathbf{h}$ that $m = 0$. It follows immediately, by Eq. (13.1), that $\langle m_1 \mid \mathbf{h} \rangle = \mathbf{a}_1 \mid \mathbf{h}$, in agreement also with Eq. (14.1). Thus the induction is completed and the theorem is proved.

If we denote by n the number of true propositions in a second group, $\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_N$, we have, in analogy to Eq. (14.1),

$$\langle n \mid \mathbf{h} \rangle = \textstyle\sum_J (\mathbf{b}_J \mid \mathbf{h}). \tag{14.3}$$

Now, among the conjunctions,

$$\begin{aligned} &\mathbf{a}_1 \cdot \mathbf{b}_1, \ \mathbf{a}_1 \cdot \mathbf{b}_2, \ldots \mathbf{a}_1 \cdot \mathbf{b}_N, \\ &\mathbf{a}_2 \cdot \mathbf{b}_1, \ \mathbf{a}_2 \cdot \mathbf{b}_2, \ldots \mathbf{a}_2 \cdot \mathbf{b}_N, \\ &\ldots \\ &\mathbf{a}_M \cdot \mathbf{b}_1, \ \mathbf{a}_M \cdot \mathbf{b}_2, \ldots \mathbf{a}_M \cdot \mathbf{b}_N, \end{aligned}$$

the number of true ones is mn, because every conjunction of one of the m true propositions of the first group with one of the n true propositions of the second group is true. All the others are false,

because each is a conjunction either of two false propositions or of one false and one true, and in either case is false itself. Hence it follows that

$$\langle mn \mid \mathbf{h} \rangle = \sum_I \sum_J (\mathbf{a}_I \cdot \mathbf{b}_J \mid \mathbf{h}). \tag{14.4}$$

The proof can easily be extended to apply to the product of the numbers of true propositions in more than two groups.

According to Eq. (13.2), the product of the deviations of m and n has an expectation given by

$$\langle (m - \langle m \rangle)(n - \langle n \rangle) \rangle = \langle mn \rangle - \langle m \rangle \langle n \rangle,$$

and therefore, by Eqs. (14.1), (14.3) and (14.4),

$$\langle (m - \langle m \rangle)(n - \langle n \rangle) \rangle = \sum_I \sum_J [(\mathbf{a}_I \cdot \mathbf{b}_J \mid \mathbf{h}) - (\mathbf{a}_I \mid \mathbf{h})(\mathbf{b}_J \mid \mathbf{h})]. \tag{14.5}$$

If the two groups of propositions are mutually irrelevant, $\mathbf{a}_I \cdot \mathbf{b}_J \mid \mathbf{h} = (\mathbf{a}_I \mid \mathbf{h})(\mathbf{b}_J \mid \mathbf{h})$ for all values of I and J. In this case, therefore, the expectation of the product of the deviations is zero.

When the two groups of propositions are identical, the equation becomes

$$\langle (m - \langle m \rangle)^2 \rangle = \sum_I \sum_J [(\mathbf{a}_I \cdot \mathbf{a}_J \mid \mathbf{h}) - (\mathbf{a}_I \mid \mathbf{h})(\mathbf{a}_J \mid \mathbf{h})] \tag{14.6}$$

and thus gives the expectation of the square of the deviation of m. A group of propositions can not be completely irrelevant to itself (except in the trivial case in which every proposition is either certain or impossible) but each proposition can be irrelevant to every one except itself. With this degree of irrelevance, $(\mathbf{a}_I \cdot \mathbf{a}_J \mid \mathbf{h}) = (\mathbf{a}_I \mid \mathbf{h})(\mathbf{a}_J \mid \mathbf{h})$ for all unequal values of I and J. All the terms of the summation on the right in Eq. (14.6) therefore vanish, except those in which $J = I$, and thus the summation becomes single-fold. Since also $\mathbf{a}_I \cdot \mathbf{a}_I = \mathbf{a}_I$, the equation becomes

$$\langle (m - \langle m \rangle)^2 \rangle = \sum_I (\mathbf{a}_I \mid \mathbf{h})[1 - (\mathbf{a}_I \mid \mathbf{h})] = \sum_I (\mathbf{a}_I \mid \mathbf{h})(\sim \mathbf{a}_I \mid \mathbf{h}). \tag{14.7}$$

The symmetry on the right between the inferences and their contradictories shows that the square of the deviation in the number of false propositions has the same expectation as in the number of true ones. This is a consequence of the fact that every excess in the number of true propositions above its expectation is accompanied by an equal deficiency in the number of false ones, and the squares of the two deviations are thus equal and equally probable.

If we denote m/M, the proportion of true propositions to the total number, by μ, Eq. (14.1) becomes

$$\langle \mu \rangle = \Sigma_I (\mathbf{a}_I \mid \mathbf{h})/M. \tag{14.8}$$

Thus $\langle \mu \rangle$ is equal to the arithmetical average of all the probabilities.

Let us now denote by D_I the difference, $(\mathbf{a}_I \mid \mathbf{h}) - \langle \mu \rangle$, between one probability and the average of all, so that $\Sigma_I D_I = 0$. Replacing m by μM and $(\mathbf{a}_I \mid \mathbf{h})$ by $\langle \mu \rangle + D_I$ in Eq. (14.7), we find that

$$\langle (\mu - \langle \mu \rangle)^2 \rangle = \frac{\langle \mu \rangle (1 - \langle \mu \rangle)}{M} - \frac{\Sigma_I D_I^2}{M^2}.$$

Because $\Sigma_I D_I^2/M^2$ can not be negative, it follows from this equation that $\langle (\mu - \langle \mu \rangle)^2 \rangle$ can not be greater than $\langle \mu \rangle (1 - \langle \mu \rangle)/M$, the value which it attains when all the propositions are equally probable. Moreover, the maximum value of $\langle \mu \rangle (1 - \langle \mu \rangle)$, attained when $\langle \mu \rangle = \frac{1}{2}$, is $\frac{1}{4}$. Therefore

$$\langle (\mu - \langle \mu \rangle)^2 \rangle \leq \frac{1}{4M}. \tag{14.9}$$

It was remarked in the chapter before this one that the expectation of the square of the deviation of a quantity measures the dispersion of its probable values and is small if the quantity is unlikely to have values appreciably different from its expectation. We therefore conclude from Eq. (14.8) and the inequality (14.9) that:

In any group of mutually irrelevant propositions, the proportion of true ones has an expectation equal to the average of the probabilities of all the propositions, and an appreciable difference between this proportion and its expectation is very improbable if the propositions are very numerous. (14.i)

This is one of a group of theorems which express, with greater or less precision, the principle known as the *law of great numbers*.[28]

15. *The Ensemble of Instances*

In the preceding chapter, the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M$, were not required to have any resemblance among themselves in order to be associated as a group. In the present chapter, we consider a more restricted case, in which the subjects of all the propositions have some common characteristic. Singly the subjects are called *instances* of this characteristic and collectively they are said to form an *ensemble* of instances. For example, a hand at cards is an instance in the ensemble of all hands dealt according to the same rules, and a particular inhabitant of North America is an instance in the ensemble of all North Americans.

Although the instances of the ensemble are all identical in the respect by which the ensemble is defined, they are not necessarily so in other respects. We suppose, indeed, that each instance is distinguishable in some way from every other and each is therefore unique in at least one particular. It is to be understood that both the common characteristic which defines the ensemble and the singular characteristics which distinguish the instances are stated in the hypothesis, **h**, of the argument. Concerning these characteristics, therefore, the hypothesis is explicit, whether in ascribing them to all the instances or in ascribing them to some or only one and denying them to the rest.

Ordinarily there are also other characteristics, concerning whose presence in any instance the hypothesis is not explicit but provides ground only for probable inference. We suppose that

it is with such a characteristic that the group of propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M$, is concerned and that the proposition $\mathbf{a}_I$ asserts that this characteristic is present in the I th instance in the ensemble. For example, $\mathbf{a}_I$ may assert that the ace of hearts is in the I th hand at cards or that the I th North American has studied Latin. M becomes a number of instances in the ensemble and m the number of these instances having the characteristic in question.

Because all the propositions have reference to the same characteristic and differ only in ascribing it to different instances, it is only the particulars which distinguish the instances that can cause inequalities among the probabilities, $(\mathbf{a}_1 \mid \mathbf{h})$, $(\mathbf{a}_2 \mid \mathbf{h})$, $\ldots$ $(\mathbf{a}_M \mid \mathbf{h})$. If these particulars are all irrelevant to the characteristic in question, the probabilities are all equal and a single symbol p may stand for any of them. In this case, the expression given for $\langle m \rangle$ by Eq. (14.1) becomes simply the sum of M terms each equal to p and we have the familiar result,

$$\langle m \rangle = Mp.$$

If also the presence of the characteristic in any instance is irrelevant to its presence in any other, so that the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M$, are all mutually irrelevant, Eq. (14.7) holds and becomes

$$\langle (m - \langle m \rangle)^2 \rangle = Mp(1 - p).$$

If m/M is denoted by μ, these equations take the form,

$$\langle \mu \rangle = p$$

and

$$\langle (\mu - \langle \mu \rangle)^2 \rangle = p(1 - p)/M.$$

Thus the probability, p, of the characteristic is not only the expectation of μ, the proportion in which it is present in M instances in the ensemble, but is also the value which this proportion will almost certainly approach as M, the number of instances, becomes very large.

It is a corollary of this principle that the average, over a large number of instances, of every quantity which satisfies certain appropriate conditions is almost certain to be nearly equal to the expectation of that quantity in a single instance. To see that this is true, let x be a quantity which, in any instance, has one of the values, $x_1, x_2, \ldots x_r, \ldots$, and let its value in one instance be irrelevant to its value in any other. Let the probability of any value, as x_r, be the same in every instance, so that we may denote it always by the same symbol, p_r. Then the expectation of x in any instance is given by

$$\langle x \rangle = \sum_r x_r p_r.$$

Among M instances in the ensemble, let the number in which x has the value x_r be denoted by m_r. The average value of x in these M instances is then given by

$$x_{\text{av}} = \sum_r x_r m_r / M.$$

If M is a very large number, m_r/M is almost certain to be very nearly equal to p_r. Therefore x_{av} is almost certain to be very nearly equal to $\langle x \rangle$.

In such a subject as statistical mechanics, in which the numbers of instances are ordinarily enormous, it is common practice to ignore the distinction between the expectation and the average, as though they were not only equal quantities but also interchangeable concepts.

When we say that a true die will show, on the average, one deuce in every six throws, we are, in effect, considering an ensemble not of single throws but of sequences of six. One such sequence is one instance in this ensemble, and the number of deuces in the sequence is a quantity whose possible values are the integers from 0 to 6. Its expectation in a single instance and its approximate average in a large number are both equal to 1. The law of great numbers, in the aspect illustrated by this example, is often called the *law of averages*.

16. The Rule of Succession

The characteristic which the proposition $\mathbf{a}_I$ ascribes to the I th instance in an ensemble was supposed, in the chapter before this one, to satisfy two rather strict conditions of irrelevance. First, its presence in any instance was assumed irrelevant to the presence of whatever singular characteristic served in the hypothesis to distinguish that instance from the others in the ensemble. Second, its presence in one instance was assumed irrelevant to its presence in any other instance. Let us now compare this case with one in which the second of these assumptions is replaced by a less stringent requirement.

For a rather trivial example, imagine a bag full of dice, all accurately squared and balanced but carelessly stamped, so that some of them have two, three or more faces marked with two spots. After the dice have been thoroughly shaken in the bag, one of them is to be drawn and thrown a number of times.

On an hypothesis which identifies the die, whether as correctly stamped or as stamped defectively in a specific way, the conditions of irrelevance assumed in the preceding chapter are satisfied in respect to throwing deuces. The probability of a deuce in any single throw is equal to the ratio of the number of faces marked with two spots to 6, the total number of faces. Moreover, no inference from the result in one throw can alter the probabilities of the results possible in any other, for, except for defects in marking, the dice are true.

If, on the other hand, the die is not identified in the hypothesis except as having been drawn from the bag of mixed dice, the results of different throws are not mutually irrelevant. For example, if any of the dice in the bag had every face stamped with two spots, a long run of deuces will make it very likely that the die drawn was one of them, and a deuce on the next throw thereafter, though not quite certain, will be very nearly so. The result even of a single throw will contribute something to the iden-

tification of the die and thus change the probability of a deuce on the next throw. If it is a deuce, it will somewhat increase the probability that the die has more than one face with two spots. If it is not a deuce, it will eliminate the possibility that all six faces are so marked and it will make some changes in the probabilities of the other possible markings.

Generalizing from this example, we consider an ensemble of instances defined by a common characteristic, which is not itself identified, however, except as one of a set of mutually exclusive alternatives. In the example, the ensemble consists in the throws of the die, the common characteristic is that the same die is thrown in all the instances, and the alternatives are distinguished by the different markings of the dice in the bag from which one was drawn to be thrown. If we distinguish the alternatives in the general case by numbers, 1, 2, ... w, and denote by $\mathbf{p}_r$ the proposition which names the rth alternative as the common characteristic of all the instances, then, in the example, $w = 6$ and $\mathbf{p}_r$ asserts that the die drawn has r faces marked with two spots. In the general case we suppose that the hypothesis $\mathbf{h}$ assigns a probability to each of the propositions, $\mathbf{p}_1$, $\mathbf{p}_2$, ... $\mathbf{p}_w$, and that these propositions form an exhaustive set, so that $\Sigma_r(\mathbf{p}_r \mid \mathbf{h}) = 1$. In the example, $\mathbf{p}_r \mid \mathbf{h}$ is the fraction of the dice in the bag that have r faces marked with two spots.

We now consider a characteristic which we expect to be present in some instances in the ensemble and absent from others, and we denote, as heretofore, by $\mathbf{a}_I$ the proposition which ascribes this characteristic to the Ith instance. In the example of the dice, $\mathbf{a}_I$ asserts that the Ith throw of the die is a deuce. Just as, in the example, when the marking of the die is specified, the results of successive throws are mutually irrelevant as well as equally probable, so, in the generalization, when one of the propositions, $\mathbf{p}_1$, $\mathbf{p}_2$, ... $\mathbf{p}_w$, is asserted in the hypothesis, we attribute mutual irrelevance and equal probability to each of the inferences, $\mathbf{a}_1$, $\mathbf{a}_2$, ... $\mathbf{a}_I$, ... If we denote $\mathbf{a}_I \mid \mathbf{p}_r \cdot \mathbf{h}$ by p_r in the generalization, then $p_r = r/6$ in the example.

On these assumptions let us seek an expression for $\mathbf{a}_{M+1} \mid \mathbf{m} \cdot \mathbf{h}$, where $\mathbf{m}$ asserts that the number of true propositions in the group, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M$, is m. Thus, in the example, we suppose that the die has been thrown M times and has shown m deuces, and we seek, with this information, to know the probability of a deuce on the next throw. In the generalization, we suppose that M instances in the ensemble have been examined and the characteristic under consideration has been found present in m of them, and we seek its probability in the next instance.

Although $\mathbf{m}$ states the number of true propositions in the group, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M$, it does not say of any particular proposition whether it is among the m true ones or the $M - m$ false ones. Let us denote by $\mathbf{m}^*$ a more specific proposition, which not only asserts, as $\mathbf{m}$ does, that there are m true propositions in the group but also, as $\mathbf{m}$ does not, specifies which propositions are true and, by exclusion, which are false. Consider first the probability of $\mathbf{m}^*$ on the hypothesis $\mathbf{p}_r \cdot \mathbf{h}$. Because, on this hypothesis, each of the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_M$, has the probability p_r and they are all mutually irrelevant, the probability that two of them, as $\mathbf{a}_I$ and $\mathbf{a}_J$, are both true is p_r^2 and that they are both false is $(1 - p_r)^2$, whereas the conjunctions which specify one as true and the other as false have probabilities given by the equation,

$$\mathbf{a}_I \cdot \sim\mathbf{a}_J \mid \mathbf{p}_r \cdot \mathbf{h} = \sim\mathbf{a}_I \cdot \mathbf{a}_J \mid \mathbf{p}_r \cdot \mathbf{h} = p_r(1 - p_r).$$

We assume irrelevance in all the possible conjunctions by which some of the propositions are specified as true and some as false and thus, continuing the same reasoning, we see that

$$\mathbf{m}^* \mid \mathbf{p}_r \cdot \mathbf{h} = p_r^m(1 - p_r)^{M-m}.$$

To find an expression for $\mathbf{m}^* \mid \mathbf{h}$, we equate the two expressions for $\mathbf{m}^* \cdot \mathbf{p}_r \mid \mathbf{h}$ and thus have

$$(\mathbf{m}^* \mid \mathbf{h})(\mathbf{p}_r \mid \mathbf{m}^* \cdot \mathbf{h}) = (\mathbf{m}^* \mid \mathbf{p}_r \cdot \mathbf{h})(\mathbf{p}_r \mid \mathbf{h}).$$

Substituting in this equation the expression just obtained for $\mathbf{m}^* \mid \mathbf{p}_r \cdot \mathbf{h}$, we find that

$$(\mathbf{m}^* \mid \mathbf{h})(\mathbf{p}_r \mid \mathbf{m}^* \cdot \mathbf{h}) = p_r^m(1 - p_r)^{M-m}(\mathbf{p}_r \mid \mathbf{h}),$$

whence, summing over all values of r, we obtain

$$\mathbf{m}^* \mid \mathbf{h} = \sum_r p_r^m(1 - p_r)^{M-m}(\mathbf{p}_r \mid \mathbf{h}).$$

The conjunction, $\mathbf{a}_{M+1} \cdot \mathbf{m}^*$, specifies as false the same $M - m$ propositions as $\mathbf{m}^*$ and as true the m propositions so specified by $\mathbf{m}^*$ and one more, $\mathbf{a}_{M+1}$. Therefore, by analogy with the equation just found for $\mathbf{m}^* \mid \mathbf{h}$, we have

$$\mathbf{a}_{M+1} \cdot \mathbf{m}^* \mid \mathbf{h} = \sum_r p_r^{m+1}(1 - p_r)^{M-m}(\mathbf{p}_r \mid \mathbf{h}).$$

These two results can be combined to give an expression for $\mathbf{a}_{M+1} \mid \mathbf{m}^* \cdot \mathbf{h}$, for

$$\mathbf{a}_{M+1} \mid \mathbf{m}^* \cdot \mathbf{h} = (\mathbf{a}_{M+1} \cdot \mathbf{m}^* \mid \mathbf{h})/(\mathbf{m}^* \mid \mathbf{h}),$$

and hence we have

$$\mathbf{a}_{M+1} \mid \mathbf{m}^* \cdot \mathbf{h} = \frac{\sum_r p_r^{m+1}(1 - p_r)^{M-m}(\mathbf{p}_r \mid \mathbf{h})}{\sum_r p_r^m(1 - p_r)^{M-m}(\mathbf{p}_r \mid \mathbf{h})}.$$

It is to be noted that the expression on the right in this equation depends on the number of propositions specified as true and false but not on the way in which they are specified. Thus $\mathbf{a}_{M+1}$ has the same probability for all the specifications consistent with the given numbers. Hence it follows that it has this probability also if the propositions are not specified but only the numbers are given, as they are by the proposition $\mathbf{m}$. Thus, although the propositions $\mathbf{m}$ and $\mathbf{m}^*$ are quite different, their difference is irrelevant to $\mathbf{a}_{M+1}$ and therefore they are interchangeable in the hypothesis when $\mathbf{a}_{M+1}$ is the inference. Hence $\mathbf{a}_{M+1} \mid \mathbf{m} \cdot \mathbf{h} = \mathbf{a}_{M+1} \mid \mathbf{m}^* \cdot \mathbf{h}$, and the solution of the problem with which we have been concerned is the equation,

$$\mathbf{a}_{M+1} \mid \mathbf{m} \cdot \mathbf{h} = \frac{\sum_r p_r^{m+1}(1 - p_r)^{M-m}(\mathbf{p}_r \mid \mathbf{h})}{\sum_r p_r^m(1 - p_r)^{M-m}(\mathbf{p}_r \mid \mathbf{h})}. \qquad (16.1)$$

This equation can be expressed as a relation among expectations, for we may regard $p_1, p_2, \ldots p_w$ as the possible values of a

single quantity p and $\mathbf{p}_r$ as a proposition which ascribes to p the value p_r. In the example of the dice, p is the probability of throwing a deuce when it is known what die is being thrown. In general it is the probability of $\mathbf{a}_I$ (for an arbitrary value of I) when it is known which of the alternatives, $\mathbf{p}_1, \mathbf{p}_2, \ldots \mathbf{p}_w$, is true. With this understanding, the right-hand member of Eq. (16.1) appears as the ratio of the expectations of two functions of p. To express $\mathbf{a}_{M+1} \mid \mathbf{m}\cdot\mathbf{h}$, the left-hand member, as an expectation also, we equate the two expressions for $\mathbf{a}_{M+1}\cdot\mathbf{p}_r \mid \mathbf{m}\cdot\mathbf{h}$, and so obtain

$$(\mathbf{a}_{M+1} \mid \mathbf{m}\cdot\mathbf{h})(\mathbf{p}_r \mid \mathbf{a}_{M+1}\cdot\mathbf{m}\cdot\mathbf{h}) = (\mathbf{a}_{M+1} \mid \mathbf{p}_r\cdot\mathbf{m}\cdot\mathbf{h})(\mathbf{p}_r \mid \mathbf{m}\cdot\mathbf{h}) = p_r(\mathbf{p}_r \mid \mathbf{m}\cdot\mathbf{h}),$$

whence, summing with respect to r, we see that

$$\mathbf{a}_{M+1} \mid \mathbf{m}\cdot\mathbf{h} = \textstyle\sum_r p_r(\mathbf{p}_r \mid \mathbf{m}\cdot\mathbf{h}) = \langle p \mid \mathbf{m}\cdot\mathbf{h}\rangle.$$

Thus Eq. (16.1) can be written

$$\langle p \mid \mathbf{m}\cdot\mathbf{h}\rangle = \frac{\langle p^{m+1}(1-p)^{M-m} \mid \mathbf{h}\rangle}{\langle p^m(1-p)^{M-m} \mid \mathbf{h}\rangle}. \tag{16.2}$$

In some examples, p is not limited to discrete values but has a continuous range. In such a case, Eq. (16.2) requires no change, but the summations in Eq. (16.1) must be replaced by integrals. If we denote by $f(p)\,dp$ the probability on the hypothesis $\mathbf{h}$ that p has a value within the infinitesimal range dp, the equation becomes

$$\mathbf{a}_{M+1} \mid \mathbf{m}\cdot\mathbf{h} = \frac{\int_0^1 p^{m+1}(1-p)^{M-m}f(p)\,dp}{\int_0^1 p^m(1-p)^{M-m}f(p)\,dp}. \tag{16.3}$$

If $f(p)$ is constant in the integrations, the integrals take known forms and the equation becomes simply

$$\mathbf{a}_{M+1} \mid \mathbf{m}\cdot\mathbf{h} = \frac{m+1}{M+2}. \tag{16.4}$$

This is Laplace's *rule of succession*.[29]

Only in exceptional cases, however, can $f(p)$ reasonably be assumed constant. This assumption requires, if the range of values of p from 0 to 1 be divided into equal elements, that p is just as likely, on the hypothesis **h**, to have a value in one element as another. Artificial hypotheses can be constructed which satisfy this requirement, but actual circumstances seldom do so. It is not from these exceptional cases that the rule of succession derives its utility but from the much more numerous cases in which the rule can be shown to hold approximately when M, the number of known instances, is very large. It holds in the latter cases, not because of an assumed indifference of the hypothesis to the value of p, which is the ground on which it has usually been justified, but because, when M is very large, the expression given in Eq. (16.3) for $\mathbf{a}_{M+1} \mid \mathbf{m}\cdot\mathbf{h}$ is indifferent, or very nearly so, to the form of $f(p)$. In other words, the rule is useful not because $f(p)$ has commonly a particular form but because, when M is large enough, its form hardly matters.

17. *Expectation and Experience*

To obtain the rule of succession in its wider use, we eliminate m from Eq. (16.3), denoting m/M by μ, and so find the equation in the form,

$$\mathbf{a}_{M+1} \mid \mathbf{m}\cdot\mathbf{h} = \frac{\int_0^1 p[p^\mu(1-p)^{1-\mu}]^M f(p)\,dp}{\int_0^1 [p^\mu(1-p)^{1-\mu}]^M f(p)\,dp}.$$

By differentiating the function $p^\mu(1-p)^{1-\mu}$, with respect to p while keeping μ constant, we find that it has its maximum value when $p = \mu$. When this function is raised to the power M, as it is in the integrands in the equation, the maximum stays at the same value of p and, as M is increased, the factor by which the maximum exceeds the other values increases exponentially. It

follows, when M is very large, that the integrands are negligible except for values of p in the near neighborhood of μ, whatever the form of the function $f(p)$, provided only that it is not very much smaller in this neighborhood than elsewhere. The values of the integrals are therefore sensitive to the form of $f(p)$ only in this neighborhood and, unless it is there a very rapidly varying function of p, it may be replaced in the integrands by $f(\mu)$. As μ is a constant in the integration, $f(\mu)$ can now be taken outside the integral signs. There, as a common factor of numerator and denominator, it is eliminated from the equation. The result is again the rule of succession, which is approximated, when M is very large, by the equation,

$$\mathbf{a}_{M+1} \mid \mathbf{m}\cdot\mathbf{h} = m/M.$$

Thus, in determining probabilities in the ensemble, the accumulation of instances prevails, in the long run, over the prior evidence, and the fraction of instances in which a characteristic is found present becomes, as the instances are multiplied, the probability of the characteristic in a new instance.

It is still important, however, to remember the two requirements of irrelevance by which this conclusion was made possible. The first is that the instances be differentiated from one another only by particulars irrelevant to the presence of the characteristic whose probability is in question. The importance of this requirement can be seen in an example taken from Peirce:

> "About two per cent of persons wounded in the liver recover,
> This man has been wounded in the liver;
> Therefore there are two chances out of a hundred that he will recover."[30]

What counts here is the particular by which "this man" is to be identified. If he is not identified at all except as someone wounded in the liver, he remains an anonymous, undifferentiated member of the population whose injury defines the ensemble. In this case, the statement that "there are two chances out of a

hundred that he will recover" is scarcely if at at all more than a tautology which repeats in other words the statement that "about two per cent of persons wounded in the liver recover." But, if he is identified in any more discriminating way, the statement about his chances of recovery depends for its validity upon the irrelevance between the proposition which identifies him and the inference that he will recover. If he is identified as the patient of a skillful surgeon, his chances will not be the same as if he were attended by a tribal medicine man. If he is Prometheus, his chances can be estimated only by comparing the prognosis of wounds of the liver inflicted by the vultures of Zeus with that of injuries more conventionally incurred.

The second requirement for proving the rule is that of mutual irrelevance among the propositions, $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_I, \ldots$, which was assumed to hold on each of the alternatiye hypotheses, $\mathbf{p}_1 \cdot \mathbf{h}, \mathbf{p}_2 \cdot \mathbf{h}, \ldots \mathbf{p}_w \cdot \mathbf{h}$. A celebrated calculation by Laplace provides an example in which this requirement was not satisfied. Accepting historical evidence for the past occurrence of 1,826,213 sunrises, he used the rule of succession to estimate the probability of the next as $\frac{1{,}826{,}214}{1{,}826{,}215}$. This calculation ignores the fact that, if one sunrise failed to occur as expected, this would, on any credible hypothesis, change the probability of the one expected to follow it.[31]

In this chapter so far, and the one before it, we have been concerned with examples at two extremes. In the example of the dice, considered in the preceding chapter, the required conditions of irrelevance are fully met. By contrast, in the example of the sunrise, they are not met at all, and the calculation from the rule of succession is, in this example, a travesty of the proper use of the principle. Between these extremes we carry on the familiar daily reasoning by which we bring our experience to bear on our expectation. In an ordinary case, we are obliged, under the given circumstances, whatever they are, to anticipate an unknown event. We look to experience for occasions in which the circum-

stances were similar and where we know the event which followed them. We determine our expectation of a particular event in the present instance by the frequency with which like events have occurred in the past, allowing as best we can for whatever disparity we find between the present and the former circumstances.

The ensemble is the conventional form for this reasoning. Some cases it fits with high precision, others with low, and for some it is scarcely useful. Suppose that someone is reading a book about a subject which he knows well in some respects but not in others, and that he finds, among the author's assertions, instances both true and false in the matters he knows about. If he finds more truth than error in these matters, he will judge that an assertion about an unfamiliar matter is more probably true than false, other things being equal. His reasoning has the same character as an application of the rule of succession but not the same precision. In the algebra of propositions,

$$\mathbf{a} = \mathbf{a} \vee (\mathbf{b} \cdot \sim\mathbf{b}) = (\mathbf{a} \vee \mathbf{b}) \cdot (\mathbf{a} \vee \sim\mathbf{b})$$

for every meaning of **a**, and thus there is no proposition so simple that it can not be expressed as the conjunction of others. Hence there is no unambiguous way of counting the assertions in a discourse. Although it is possible often to recognize true and false statements and sometimes to observe a clear preponderance of one kind over the other, yet this observation can not always be expressed by a ratio of numbers of instances, as it must be if the rule of succession is to be applicable.

In every case in which we use the ensemble to estimate a probability, whether with high precision or low, we depend on the similarity of the circumstances associated with the known and unknown events. It seems strange, therefore, that Venn, who defined probability in terms of the ensemble, should have excluded argument by analogy from the theory, as he did in the passage quoted in the first chapter. For every estimate of probability made by that definition is an argument by analogy.

18. A Remark on Induction

Inductive reasoning, when the term is used broadly, is any reasoning in which the verification of one or more propositions is adduced as an argument for the truth, or at least the probability, of a proposition which implies them. For example, we see leaves moving and infer that the wind is blowing, or we hear the whistle of a locomotive and infer that a train is coming.

The argument depends on the equality of the two expressions for the probability of a conjunctive inference. Let **g** be a proposition which, on the hypothesis **h**, implies another proposition, **i**. Equating the two expressions for $\mathbf{g}\cdot\mathbf{i} \mid \mathbf{h}$, we have

$$(\mathbf{g} \mid \mathbf{h}\cdot\mathbf{i})(\mathbf{i} \mid \mathbf{h}) = (\mathbf{i} \mid \mathbf{h}\cdot\mathbf{g})(\mathbf{g} \mid \mathbf{h}),$$

whence

$$\frac{\mathbf{g} \mid \mathbf{h}\cdot\mathbf{i}}{\mathbf{i} \mid \mathbf{h}\cdot\mathbf{g}} = \frac{\mathbf{g} \mid \mathbf{h}}{\mathbf{i} \mid \mathbf{h}}.$$

To say that **g** implies **i** is to say that $\mathbf{i} \mid \mathbf{h}\cdot\mathbf{g} = 1$ and thus

$$\mathbf{g} \mid \mathbf{h}\cdot\mathbf{i} = \frac{\mathbf{g} \mid \mathbf{h}}{\mathbf{i} \mid \mathbf{h}}. \tag{18.1}$$

By this equation, $\mathbf{g} \mid \mathbf{h}\cdot\mathbf{i} > \mathbf{g} \mid \mathbf{h}$ unless $\mathbf{g} \mid \mathbf{h} = 0$ or $\mathbf{i} \mid \mathbf{h} = 1$. The reasons for these two exceptions are obvious. If $\mathbf{g} \mid \mathbf{h} = 0$, **g** is an impossible inference to begin with and no accumulation of evidence will make it possible. If $\mathbf{i} \mid \mathbf{h} = 1$, **i** is implied by **h** and its verification, since it gives no information which was not already implicit in **h** alone, can not change the probability of **g**. In all other cases, Eq. (18.1) shows that the verification of any proposition **i** increases the probability of every proposition **g** which implies it.

Moreover, the smaller is $\mathbf{i} \mid \mathbf{h}$, the prior probability of **i**, the greater is $(\mathbf{g} \mid \mathbf{h}\cdot\mathbf{i})/(\mathbf{g} \mid \mathbf{h})$, the factor by which its verification in-

creases the probability of **g**. For example, when Fresnel's memoir on the wave theory of light was being considered for a prize of the French Academy, Poisson, who was one of the judges, pointed out the implication that the circular shadow of a disk, intercepting light from a fine source, would have a small bright spot at its center. This had never been seen and its existence therefore appeared very improbable. When Fresnel performed the experiment and showed the bright spot, the unexpectedness of the result made it so much the stronger evidence for the theory which implied it.[32]

For another example, we may consider Macbeth's reasoning about the witches who hailed him on the desolate heath as thane of Glamis and Cawdor and thereafter king. At first he was incredulous and said,

> "By Sinel's death I know I am thane of Glamis;
> But how of Cawdor? the thane of Cawdor lives,
> A prosperous gentleman; and to be king
> Stands not within the prospect of belief,
> No more than to be Cawdor."

Farther along the way he met King Duncan's messengers and learned that he had in truth become thane of Cawdor. So he was persuaded that the witches knew what they were talking about and the more so because the prediction just confirmed had been so improbable before.

Returning to the formal argument, let **j** be another proposition implied by **g** on the hypothesis **h** but not implied by **h** alone or by **h·i**. Then, by the same reasoning as before, we find that

$$\mathbf{g} \mid \mathbf{h \cdot i \cdot j} > \mathbf{g} \mid \mathbf{h \cdot i},$$

and if **k** is yet another proposition implied by **g·h** but not by **h·i·j**,

$$\mathbf{g} \mid \mathbf{h \cdot i \cdot j \cdot k} > \mathbf{g} \mid \mathbf{h \cdot i \cdot j}.$$

Thus **g** becomes more probable with the verification of each of the propositions, **i**, **j**, **k**, which it implies. This cumulative effect of

successive verifications is important in the special case of inductive reasoning often distinguished by the briefer name *induction.*

Induction has to do with an ensemble of instances. It is reasoning in which the observed presence of a given characteristic in some instances in the ensemble is made an argument for its presence in all of them. Because the conclusion expressed by Eq. (18.1) holds for inductive reasoning generally, it holds in this special case. Therefore a proposition which ascribes the given characteristic to all the instances is made more probable by its verified presence in some, with only the two obvious exceptions already noted.

The ensemble which is made the subject of an induction is ordinarily unlimited in the number of its instances. The argument is aimed at establishing a universal principle, valid under given circumstances no matter how many times they are encountered or produced. Certainty is hardly to be expected in such an argument, for it would be surprising if a principle could be proved valid in an infinite number of instances by being verified in a finite number. In some cases, however, certainty is approximated when the number of verified instances is very large.

For example, let the subject of the induction be such an ensemble as was described in Chapter 16. Let the characteristic whose probabilities on the alternative hypotheses are the possible values of p be the one which $\mathbf{g}$ ascribes to every instance in the ensemble. Then $\mathbf{g}$ is included among the alternative propositions and $p = 1$ when $\mathbf{g}$ is certain. Let $\mathbf{i}$ assert that this characteristic has been found present in every one of M instances examined. Then, in Eq. (16.2), $\mathbf{m} = \mathbf{i}$ and $m = M$, and the equation becomes

$$\langle p \mid \mathbf{i}\cdot\mathbf{h} \rangle = \frac{\langle p^{M+1} \mid \mathbf{h} \rangle}{\langle p^{M} \mid \mathbf{h} \rangle}.$$

As M increases indefinitely, p^M and p^{M+1} approach zero for all values of p less than 1, and these values therefore contribute less and less to the expectations on the right in the equation. By contrast, the value 1 contributes the amount $\mathbf{g} \mid \mathbf{h}$ to each of the

expectations, whatever the value of M. Hence, unless $\mathbf{g} \mid \mathbf{h} = 0$, each of the expectations, $\langle p^M \mid \mathbf{h} \rangle$ and $\langle p^{M+1} \mid \mathbf{h} \rangle$, is nearly equal to $\mathbf{g} \mid \mathbf{h}$ when M is large enough, and $\langle p \mid \mathbf{i} \cdot \mathbf{h} \rangle$, being equal to their ratio, is nearly equal to 1. Since the maximum value of p is also 1, it follows, when p has an expectation equal, or nearly equal, to 1, that $\mathbf{g}$, the proposition which ascribes this value to p, is certain, or nearly so. Thus, if the characteristic in question is found present in every one of a large enough number of instances, it is almost certainly present in all of them.

All this has a bearing on Hume's criticism of induction. In his *Enquiry Concerning Human Understanding*, he asks the question:

> "Now where is that process of reasoning which, from one instance, draws a conclusion so different from that which it infers from a hundred instances that are nowise different from that single one?"

and he continues:

> "This question I propose as much for the sake of information, as with an intention of raising difficulties. I cannot find, I cannot imagine any such reasoning."[33]

The instances differ more among themselves, however, than is implied in Hume's question. They must differ in some respect in order to be distinguishable one from another and they may differ with respect to any characteristic except that by which the ensemble is defined. Specifically, with respect to the characteristic in question in the induction, the instances are not known to be alike until their likeness is verified by observation. This verification provides a ground for inference which was not present before. A change in the conclusion, therefore, so far from being unimaginable, is altogether reasonable, if by reasoning we mean making inferences appropriate to the premises. It would be astonishing if nothing could be inferred from the information that a characteristic is common to a hundred instances when, on prior evidence, it might have been dispersed among them in any way numerically possible.

If the criticism implied in Hume's question, on the one hand, too much ignores the differences among the instances, on the other, it stresses too much the difference which the number of instances makes in the conclusion. Whether the instances are few or many, the conclusion is the estimate of a probability and, when it changes, the change is not qualitative but quantitative and appropriate therefore to the quantitative difference between numbers of instances, of which it is the consequence. If the principle which an induction is intended to establish is possible at the beginning, it becomes gradually more probable as the number of favorable instances increases and no contrary instance is found; but, unless it is certain at the beginning, it remains uncertain, at least in some degree, after verification in any finite number of instances. If it is impossible at the beginning, no accumulation of instances can make it probable, much less certain; one instance and a hundred are in this case the same.

Hume's criticism is perhaps useful as a corrective to the opinion, occasionally maintained, that induction can not only approach certainty but can actually attain it. In any case it is valuable as emphasizing that induction, along with probable inference in general, has its own laws, which are not derived from those of deduction, and that induction therefore can not be justified as a part of necessary inference. But Hume, not content with showing that induction is not certain and not deductive, went farther and declared, in effect, that it is also not rational. In this, however, he seems simply to have identified what is rational with what is deductive and certain. That to him reasoning meant deductive reasoning and inference meant necessary inference clearly appears in a remark on argument from experience:

> "If there be any suspicion that the course of nature may change, and that the past may be no rule for the future, all experience becomes useless and can give rise to no inference or conclusion."

If we are willing to deal with probabilities rather than cer-

tainties and admit the rules of probable inference to the canon of reason, we should counterphrase this remark and say:

> If there be any possibility that the course of nature is uniform and that the past may be some rule for the future, all experience becomes useful and can give support to some inference:

> *". . . so that the whole succession of men, during the course of many ages, should be considered as a single man who subsists forever and learns continually."*[34]

Notes

1. (p. 1) Axioms of probability have been formulated in many ways by many authors in the following books and articles, and doubtless in others which have not come to my attention.

Books

Keynes, J. M., *A Treatise on Probability* (London: Macmillan, 1921).

Reichenbach, Hans, *The Theory of Probability: an inquiry into the logical and mathematical foundations of the calculus of probability.* English translation by Ernest H. Hutten and Maria Reichenbach. (Berkeley and Los Angeles: University of California Press, 1949).

Jeffreys, Harold, *Theory of Probability* (Oxford: Clarendon Press, 1st ed. 1939, 2nd ed. 1948).

von Wright, G. H., *A Treatise on Induction and Probability* (London: Routledge and Kegan Paul, 1951).

Articles

Bernstein, M. S., "An attempt at an axiomatic exposition of the principles of the calculus of probabilities" (in Russian) *Communications of the Mathematical Society of Kharkov*, Second Ser., **15** (1917).

Wrinch, Dorothy, and Jeffreys, Harold, "The nature of probability," *Phil. Mag.*, Sixth Ser., **38** (1919).

Reichenbach, Hans, "Axiomatik der Wahrscheinlichkeitsrechnung," *Math. Z.* **34** (1932).

Kolmogorov, A. "Grundbegriffe der Wahrscheinlichkeitsrechnung," *Ergebnisse der Mathematik und ihrer Grenzgebiete* **2**, 3 (1933).

Evans, H. P., and Kleene, S. C., "A postulational basis for probability," *Amer. Math. Monthly* **46** (1939).

Koopman, O., "The axioms of intuitive probability," *Annals of Math.* **41** (1940).

——— "The bases of probability," *Bull. Amer. Math. Soc.* **46** (1940).

Koopman, O., "Intuitive probabilities and sequences," *Annals of Math.* **42** (1941).

Copeland, A. H., "Postulates for the theory of probability," *Amer. J. Math.* **63** (1941).

von Wright, G. H., "Ueber Wahrscheinlichkeit, eine logische und philosophische Untersuchung," *Acta Soc. Sci. Fennica* Nova Series **A, 3,** 11 (1945).

Schrödinger, E., "The foundation of the theory of probability," I and II, *Proc. Roy. Irish Acad.* **51,** Sect. A (1947).

Jaynes, E. T., "How does the brain do plausible reasoning?" Report **421,** Microwave Laboratory, Stanford University (1957).

2. (p. 1) Venn, John, *The Logic of Chance: an essay on the foundations and province of the theory of probability,* (London and New York: Macmillan, 3rd ed. 1888) p. 124.

3. (p. 2) The opinion that the theory of probability should be restricted in this way had been advocated earlier by R. L. Ellis and by A. Cournot and it has been held since by a number of well known authors. Ellis' views were given in two papers, "On the foundations of the theory of probabilities" and "Remarks on the fundamental principles of the theory of probabilities," of which the first appeared in vol. **8** (1843) and the second in vol. **9** (1854) of the *Trans. Camb. Phil. Soc.* Both were reprinted in his *Mathematical and Other Writings* (Cambridge: Deighton, Bell and Co.; London: Bell and Daldy, 1863). The views of Cournot were given in his book, *Exposition de la Théorie des Chances et des Probabilités* (Paris: 1843). These works are cited in Keynes' *Treatise* in the course of an exposition and critical discussion of the view of probability which they express. Keynes also quotes from Venn the passage quoted in this chapter.

A recent exposition of the theory of probability as statistical frequency is that of Richard von Mises in his book, *Probability, Statistics and Truth* (2nd revised English ed., London: Allen and Unwin; New York: Macmillan, 1957. Originally published in German with the title *Wahrscheinlichkeit, Statistik und Wahrheit*).

4. (p. 4) The opinion which would comprise all kinds of probable inference in an extended logic (whether independent of the logic of necessary inference or including it as a special case) is an old one. It was expressed, for example, by Leibnitz, who wrote: "Opinion, based on probability, deserves perhaps the name knowledge also; otherwise nearly all historic knowledge and many other kinds will fall. But without disputing about terms, I hold that the *investigation of the degrees of probability* is very important, that we are still lacking in it, and that this lack is a great defect of our logics." *Nouveaux Essais sur l'Entendement Humain,* book 4, ch. 2, Langley's translation. Similar statements occur in the same work in book 2, ch. 21, and book 4, ch. 16.

The development of the calculus of probability, which was just getting under way when Leibnitz wrote, had an influence unfavorable to the acceptance of this opinion. The calculus found most of its examples in the problems first of gamesters and then of actuaries. The problems of the first kind suggested a definition of probabilities in terms of numbers of chances, those of the second, one in terms of numbers of instances in an ensemble. Neither definition was broad enough to accommodate the idea of a logic of probability which should be the art of reasoning from inconclusive evidence.

The idea persisted, however. It guided De Morgan, for example, in his *Formal Logic: or the calculus of inference necessary and probable* (London: Taylor and Walton, 1847). It was systematically developed by Keynes and strenuously championed by Jeffreys in their books cited in Note 1.

5. (p. 4) Rules of logical algebra were given by George Boole in *An Investigation of the Laws of Thought: on which are founded the mathematical theories of logic and probabilities* (London: Walton, 1854). Others later made changes in their formulation.

In his discussion of probabilities, Boole employed the definition in terms of numbers of chances, but he described an alternative possibility in the following passage, which ends ch. 17:

"From the above investigations it clearly appears, 1st, that whether we set out from the ordinary numerical definition of the measure of probability, or from the definition which assigns to the numerical measure of probability such a law of value as shall establish a formal identity between the logical expressions of events and the algebraic expressions of their values, we shall be led to the same system of practical results. 2dly, that either of these definitions pursued to its consequences, and considered in connexion with the relations which it inseparably involves, conducts us, by inference or suggestion, to the other definition. To a scientific view of the theory of probabilities it is essential that both principles should be viewed together in their mutual bearing and dependence."

6. (p. 5) Boole himself used only the signs of ordinary algebra and a number of later writers have followed his practice. It has the advantage of keeping us aware of the resemblances between Boolean and ordinary algebra. But it has the corresponding disadvantage of helping us to forget their points of contrast, and it is besides somewhat inconvenient in a discussion in which the signs of Boolean and ordinary algebra appear in the same equations. With the signs used here, which are the choice of many authors, the only required precaution against confusion is to reserve the sign • for conjunction in Boolean algebra and avoid its use as the sign of ordinary multiplication.

7. (p. 7) This duality was first pointed out by Charles S. Peirce in an article, "On an improvement in Boole's calculus of logic," *Proc. Amer. Acad. Arts and Sci.*, **7**, (1867). Later it was emphasized by E. Schröder in *Opera-*

tionskreis des Logikkalkuls (Leipzig: Teubner, 1877). It was not a feature of Boole's original algebra, because he employed the exclusive disjunctive, *either-or*, and had no sign for the inclusive disjunctive, *and/or*. The change from exclusive to inclusive disjunction was made independently by several authors, of whom W. S. Jevons was the first in his book, *Pure Logic: or the logic of quality apart from quantity* (London: Stanford, 1864).

8. (p. 12) It is interesting that vector algebra and logical algebra were developed at nearly the same time. Although Boole's *Laws of Thought* did not appear until 1854, he had already published a part of its contents some years earlier in *The Mathematical Analysis of Logic*. Hamilton's first papers on quaternions and Grassmann's *Lineale Ausdehnungslehre* were published in 1844, and Saint-Venant's memoir on vector algebra the next year.

The following quotation from P. G. Tait's *Quaternions* is apt in this connection:

"It is curious to compare the properties of these quaternion symbols with those of the Elective Symbols of Logic, as given in Boole's wonderful treatise on the *Laws of Thought*; and to think that the same grand science of mathematical analysis, by processes remarkably similar to each other, reveals to us truths in the science of *position* far beyond the powers of the geometer, and truths of deductive reasoning to which unaided thought could never have led the logician."

9. (p. 12) Many symbols have been used for probabilities. Any will serve if it indicates the propositions of which it is a function, distinguishes the inference from the hypothesis and is unlikely to be confused with any other symbol used in the same discourse with it. It should, of course, also be easily read, written and printed.

10. (p. 14) A functional equation almost the same as this was solved by Abel. The solution may be found in *Oeuvres Complètes de Niels Henrik Abel*, edited by L. Sylow and S. Lie (Christiania: Impr. de Groendahl & soen, 1881). I owe this reference to the article by Jaynes cited in Note 1.

11. (p. 29) "Bishop Blougram's Apology."

12. (p. 29) This may be the meaning of Kronecker's often quoted remark, "God made the whole numbers. Everything else is the work of man."

13. (p. 30) The *principle of insufficient reason*, invoked to justify this judgment, was so called early in the development of the theory of probability, in antithesis to the *principle of sufficient reason*. It was meant by the latter principle that causes identical in all respects have always the same effects. On the other hand, if it is known only that the causes are alike in some respects, whereas their likeness or difference in other respects is unknown, the reason

for expecting the same effect from all is insufficient. Alternatives become possible and probability replaces certainty.

In much of the early theory and some more recent, there is an underlying assumption, which does not quite come to the surface, that, in every case of this kind, alternatives can be found among which there is not only insufficient reason for expecting any one with certainty but even insufficient reason for expecting one more than another. This assumption was doubtless derived from games of chance, in which it is ordinarily valid. Its tacit acceptance, however, was probably also made easier by the use of the antithetical terms, *sufficient reason* and *insufficient reason.* The antithesis suggests what the assumption asserts, that there are only two cases to be distinguished, the one in which there is no ground for doubt and the one in which there is no ground for preference.

The term *principle of indifference,* introduced by Keynes, does not carry this implication and is besides apter and briefer.

14. (p. 31) This opinion is clearly expressed in the following quotation from W. S. Jevons:

"But in the absence of all knowledge the probability should be considered = ½, for if we make it less than this we incline to believe it false rather than true. Thus, before we possessed any means of estimating the magnitude of the fixed stars, the statement that Sirius was greater than the sun had a probability of exactly ½; it was as likely that it would be greater as that it would be smaller; and so of any other star. . . . If I ask the reader to assign the odds that a 'Platythliptic Coefficient is positive' he will hardly see his way to doing so, unless he regard them as even." *The Principles of Science: a treatise on logic and scientific method* (London and New York, Macmillan, 2nd ed. 1877).

15. (p. 31) This example is, of course, from *The Hunt of the Snark* by Lewis Carroll. Readers who wish to pursue the subject farther are referred also to *La Chasse au Snark, une agonie en huit crises,* par Lewis Carroll. Traduit pour la première fois en français en 1929 par Louis Aragon. (Paris: P. Seghers, 1949).

16. (p. 33) The influence of games of chance on the early development of the mathematical theory of probability is well described in the work of Isaac Todhunter, *A History of the Mathematical Theory of Probability from the time of Pascal to that of Laplace* (Cambridge and London: Macmillan, 1865). The theory is usually held to have begun in a correspondence on games between Pascal and Fermat. A hundred years earlier, the mathematician Cardan had written a treatise on games, *De Ludo Aleae,* but it was published after Pascal and Fermat had ended their correspondence. Cardan, according to Todhunter, was an inveterate gambler, and his interests were thus more practical and less theoretical than those of the eminent mathematicians who followed him in

the field. It is therefore not surprising that he was less disposed than they were to take for granted the equality of chances and instructed his readers how to make sure of the matter when playing with persons of doubtful character.

17. (p. 35) The word *entropy* was coined in 1871 by Clausius as the name of a thermodynamic quantity, which he defined in terms of heat and temperature but which, he rightly supposed, must have an alternative interpretation in terms of molecular configurations and motions. This conjecture was confirmed as statistical mechanics was developed by Maxwell, Boltzmann and Gibbs. As this development proceeded, the association of entropy with probability became, by stages, more explicit, so that Gibbs could write in 1889: "In reading Clausius, we seem to be reading mechanics; in reading Maxwell, and in much of Boltzmann's most valuable work, we seem rather to be reading in the theory of probabilities. There is no doubt that the larger manner in which Maxwell and Boltzmann proposed the problems of molecular science enabled them in some cases to get a more satisfactory and complete answer, even for those questions which do not seem at first sight to require so broad a treatment." (This passage is quoted from a tribute to Clausius published in the *Proceedings of the American Academy of Arts and Sciences* and reprinted in Gibbs' *Collected Works*.)

What Gibbs wrote in 1889 of the work of Maxwell and Boltzmann could not have been said of statistical mechanics as it had been presented the year before by J. J. Thomson in his *Applications of Dynamics to Physics and Chemistry*, but it applies to Gibbs' own work, *Elementary Principles in Statistical Mechanics*, published in 1902. In the comparison of these two books, it is worth noticing that Thomson mentioned entropy only to explain that he preferred not to use it, because it "depends upon other than purely dynamical considerations," whereas Gibbs made it the guiding concept in his method. As different as they are, however, these two books have one very important feature in common, which they share also with the later works of Boltzmann. This common trait is that the conclusions do not depend on any particular model of a physical system, whether the model of a gas as a swarm of colliding spherical particles or any other. Generalized coördinates were used in all these works and thus entropy was made independent of any particular structure, although it remained still a quantity with its meaning defined only in thermodynamics and statistical mechanics.

There was still wanting the extension of thought by which entropy would become a logical rather than a physical concept and could be attributed to a set of events of any kind or a set of propositions on any subject. It is true that several writers on probability had noted the need of some such concept and had even partly defined it. In Keynes' *Treatise*, for example, there is a chapter on "The weight of arguments," in which the following passage is found:

"As the relevant evidence at our disposal increases, the magnitude of the probability of the argument may either decrease or increase, according as the

new knowledge strengthens the unfavourable or the favourable evidence; but *something* seems to have increased in either case,—we have a more substantial basis on which to rest our conclusion. I express this by saying that an accession of new evidence increases the *weight* of an argument. New evidence will sometimes decrease the probability of an argument, but it will always increase its 'weight'."

This description and the attributes of weight, as he describes it in the rest of the chapter, are suggestive of, though not identical with those which have since been given to negative entropy in the theory of probability. Keynes cites two German authors, Meinong and Nitsche, as having expressed ideas on this subject somewhat similar to his.

These suggestions, however, had no influence or, at most, a very indirect one upon the assimilation of entropy in the theory of probability. This result was the product of research in a very different subject, the transmission of messages. It was accomplished by C. E. Shannon in an article, "The mathematical theory of communication," published in 1948 in the *Bell System Tech. J.* and reprinted in the book of the same title by Shannon and W. Weaver (Urbana: Univ. of Illinois Press, 1949). The transmission of messages had been the subject of mathematical analysis earlier in several articles: Nyquist, H., "Certain factors affecting telegraph speed," *Bell System Tech. J.* (1924) and "Certain topics in telegraph transmission theory," *Trans. Amer. Inst. Elect. Eng.*, **47** (1948); Hartley, R. V. L., "Transmission of information," *Bell System Tech. J.* (1928). These authors, however, did not employ the idea of entropy. Shannon not only introduced entropy in the theory of communication but also defined it in terms of the probabilities of events without limiting the definition to events of any particular kind. His work has found application in the most diverse fields and has been followed by a great deal of research by many authors. Most of this work has dealt with what has become known as information theory rather than with the general theory of probability and has therefore little direct bearing on the subject of the present essay. Reference should be made, however, to an article by A. I. Khinchin, "The entropy concept in probability theory," *Uspekhi Matematicheskikh Nauk*, 8 (1953), translated into English by Silverman and Friedman and published, with a translation of a longer paper, also by Khinchin, in the book *Mathematical Foundations of Information Theory* (New York: Dover Publications, 1957). Entropy is treated as a concept in probability also in the article by Jaynes cited in Note 1 and, in a more specialized context, in two articles by the same author entitled, "Information theory and statistical mechanics," *Phys. Rev.*, **106** and **108** (1957).

18. (p. 37) This conclusion was derived from experimentally known properties of gases by Gibbs in his work, "On the equilibrium of heterogeneous substances." It is known as Gibbs' paradox.

19. (p. 40) The logarithm of a number of alternatives as a measure of

information was used by Hartley in the paper already cited. The name *bit* as an abbreviation of *binary digit* was adopted by Shannon on the suggestion of J. W. Tukey. I do not know who first used the game of twenty questions to illustrate the measurement of information by entropy.

20. (p. 43) In statistical mechanics the condition in which the possible microscopic states of a physical system are all equally probable is called the *microcanonical distribution*. It is the condition of equilibrium of an isolated system with a given energy, and the fact that it is also the condition of maximum entropy is in agreement with the second law of thermodynamics.

21. (p. 43) A proposal to extend the meaning of such an established term as entropy calls for some justification. There is good precedent, of course, in the generalizations already made. In the work of Boltzmann and Gibbs entropy has a broader meaning than Clausius gave it, and it has a broader meaning still in the work of Shannon. The further generalization proposed here does not change its meaning in any case in which it has had a meaning heretofore. It only defines it where it has been undefined until now and it does this by reasoning so natural that it seems almost unavoidable.

22. (p. 53) Boole, in *The Laws of Thought*, applied his algebra to classes of things as well as to propositions, and it might be supposed that a system of propositions, as defined in the chapter just ended, could be considered a class of things in Boole's sense. There is indeed a likeness between them, and it is this which allows the conjunction and disjunction of systems. But in respect to contradiction the analogy fails, for the propositions which do not belong to a system **A**, although they form a Boolean class, do not constitute a system. This is because of the rule that every proposition which implies a proposition of a system itself belongs to that system. Innumerable propositions belong to the system **A** but imply propositions which do not belong to it. It is this fact which keeps the system **A** from having a system standing in such a relation to it as to be denoted by $\sim\mathbf{A}$.

23. (p. 56) In the case in which each of the systems **A** and **B** is defined by a set of mutually exclusive propositions, the definition of conditional entropy given in Eq. (10.2) is the same as Shannon's. He also gave Eq. (10.4) for the entropy of the conjunction.

24. (p. 65) This theorem has its physical counterpart in the fact that the thermodynamic entropy of a physical system is the sum of the entropies of its parts, at least so long as the parts are not made too fine. There is a system of propositions associated in statistical mechanics with every physical system, and the logical entropy of the one system is identified with the thermodynamic entropy of the other. If, in the system of propositions, there is one which is certain, the microscopic state of the physical system is uniquely determined. In a physical system of several parts, a microscopic state of the whole system

is a combination of microscopic states of the parts and the system of propositions associated with the whole system is therefore the conjunction of those associated with the parts. That the sum of the partial thermodynamic entropies is equal to the thermodynamic entropy of the system therefore implies that the microscopic state of one part is irrelevant to that of another part. This is, however, only approximately true. Insofar as it is true, it is a consequence of the short range of intermolecular forces, in consequence of which no part of the system has any influence on matter more than a minute distance beyond its boundaries. Also, in a physical system of ordinary complexity, the number of possible microscopic states is enormous and so also, therefore, is the number of propositions required to define the system of propositions. Even a high degree of relevance, if it involves only a small part of the propositions of each system, is inappreciable in the entropy. What Poincaré once called "the extreme insensibility of the thermodynamic functions" is a consequence of this characteristic.

25. (p. 66) This is a stanza from "Alice Brand," a ballad interpolated in *The Lady of the Lake.*

26. (p. 66) If we can believe the ballad, he did neither, but instead fell into an intermediate state, whence he was changed by enchantment into a grisly elf. His sister broke the spell and restored him to life in his human form. This complication, although it is essential to the theme of the ballad, seems unnecessary in the present discussion.

27. (p. 68) This view has been expressed by authors whose opinions on other subjects were widely different, as, for example:

Milton, in *Paradise Lost:* "That power Which erring men call Chance."

Hume, in *An Enquiry concerning Human Understanding:* "Though there be no such thing as *chance* in the world, our ignorance of the real cause of any event has the same influence on the understanding and begets a like species of belief or opinion."

Jevons, in *The Principles of Science:* "There is no doubt in lightning as to the point it shall strike; in the greatest storm there is nothing capricious; not a grain of sand lies upon the beach, but infinite knowledge would account for its lying there; and the course of every falling leaf is guided by the principles of mechanics which rule the motions of the heavenly bodies.

"Chance then exists not in nature, and cannot coexist with knowledge; it is merely an expression, as Laplace remarked, for our ignorance of the causes in action, and our consequent inability to predict the result, or to bring it about infallibly."

28. (p. 79) This principle was proved, in a more precise form than that given here, in the *Ars Conjectandi* of James Bernoulli, published in 1713, eight years after his death. His proof applied only to the case in which all the probabilities are equal. The general proof was published in 1837 by Poisson

in a work entitled, *Recherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile: précédées des règles générales du calcul des probabilités.* The name *law of great numbers* is due to Poisson also.

29. (p. 86) This rule was published in a memoir of the French Academy of Sciences in 1774 and again in Laplace's *Essai Philosophique sur les Probabilités.* An English translation, by Truscott and Emory, of the *Essai* has recently been reprinted by Dover Publications. The name *rule of succession* was given to Laplace's principle by Venn in his *Logic of Chance.* Venn, however, denied the practical validity of the principle, as many other authors have done before and since. Todhunter in his *History* quotes the following passage from an essay by the mathematician Waring, published in Cambridge in 1794:

"I know that some mathematicians of the first class have endeavoured to demonstrate the degree of probability of an event's happening n times from its having happened m preceding times; and consequently that such an event will probably take place; but, alas, the problem far exceeds the extent of human understanding; who can determine the time when the sun will probably cease to run its present course?" Keynes in his *Treatise* concludes a long discussion of the rule with the remark, "Indeed this is so foolish a theorem that to entertain it is discreditable." In *A Treatise on Induction and Probability,* von Wright calls it "the notorious Principle of Succession." The proper quarrel, however, is not with the derivation of the principle but only with its misuse. This, it must be admitted, has sometimes been outrageous.

30. (p. 88) This quotation is from Peirce's essay, "A theory of probable inference," which was included in the book, Johns Hopkins *Studies in Logic,* edited by Charles S. Peirce (Boston: Little, Brown and Co., 1883). The essay has been reprinted in Peirce's collected papers published by the Harvard University Press and the selections from his writings published in London by Routledge and Kegan Paul and in New York by Dover Publications.

31. (p. 89) So many authors since Laplace have criticized this calculation that it is only fair to recall his own criticism of it. After quoting odds of 1,826,214 to 1 in favor of the next sunrise, he adds: "But this number is incomparably greater for him who, recognizing in the totality of phenomena the principal regulator of days and seasons, sees that nothing at the present moment can arrest the course of it." (Translation by Truscott and Emory.)

32. (p. 92) The incident is described (although Poisson is not identified by name) in the memoir on Fresnel written by François Arago and published in his *Oeuvres Complètes* (Paris: Gide et Baudry; Leipzig: Weigel, 1854).

33. (p. 94) Section IV, part II. The quotation which follows this one is from the same section and part.

34. (p. 96) From *New Experiments on the Vacuum* by Blaise Pascal, English translation from *The Living Thoughts of Pascal presented by François Mauriac* (New York and Toronto: Longmans, Green & Co., 1940).

Index

D

E

N

P

R

S

T

U

V

W