PROBABILITY AND SCIENTIFIC INFERENCE

G. SPENCER BROWN

Research Lecturer of Christ Church, Oxford

LONGMANS, GREEN AND CO
LONDON · NEW YORK · TORONTO

LONGMANS, GREEN AND CO LTD
6 & 7 CLIFFORD STREET LONDON W 1
THIBAULT HOUSE THIBAULT SQUARE CAPE TOWN
605–611 LONSDALE STREET MELBOURNE C 1

LONGMANS, GREEN AND CO INC
55 FIFTH AVENUE NEW YORK 3

LONGMANS, GREEN AND CO
20 CRANFIELD ROAD TORONTO 16

ORIENT LONGMANS PRIVATE LTD
CALCUTTA BOMBAY MADRAS
DELHI HYDERABAD DACCA

First published 1957

PRINTED IN GREAT BRITAIN BY
SPOTTISWOODE, BALLANTYNE AND CO LTD
LONDON AND COLCHESTER

PREFACE

PHILOSOPHERS are sometimes thought to engage themselves in hair-splitting activities with no other purpose than to exercise the disciplines needed to perform them. It is not always suspected that the most prized advancements in scientific and other knowledge come through the application of these disciplines. Briefly, to philosophize is to criticize what is said in its relation to what is seen, and we should, therefore, set about such criticism whenever we begin to suspect that what we see is being distorted or occulted by what we say.

A well-known example of word-made obscurity lies in the concept of simultaneity fundamental to Newtonian mechanics. Einstein's mechanics begin, in the best tradition, with a destructive analysis of this concept.

All concepts are destructible and it is not always obvious which to destroy. I have been concerned with the destruction of one such concept which has, I believe, obscured and occulted to a distressing extent certain facts of observation which would otherwise have been clearly seen.

Some philosophers, struck with the recent discovery that no one set of metaphysical assumptions is necessary to communication, have fallen into the fallacy of thinking that we can communicate without any metaphysics at all. Locke fell into the same kind of error when he supposed that because what he called primary qualities were dependent for their perception upon no one sense, they were therefore qualities which matter possessed independently of sense. It is one of the greatest, if not the greatest, of discoveries in 2000

years to have seen that no one system of metaphysics is necessary; but it is a simple mistake to suppose that none is.

What must be done in the field of Probability is to replace an old metaphysical system with a new one. This means undertaking two tasks: first, to show where old assumptions were inadequate; and then, to work out better ones. Each of these tasks is a major one, and I do not think any purpose would be served in attempting them both at once. An advance in scientific theory depends ultimately upon carrying forward informed opinion, and it is an ordinary biological fact that opinion, however informed, cannot be carried through too many stages at once.

The primary task of this book, therefore, has been one of analysis; and though I have indicated how the synthesis must follow, I have not completed it. That, if the opportunity be given, will be my next task. But if anyone is impatient he has only to follow me to the end of the book and he will be ready to make the synthesis himself.

I have been criticized for not publishing before now my evidence from random number counts. This delay occurred first because of illness, after which I gave priority to the task of discovering *why* certain things had happened rather than to the task of trying to convince people at this stage *that* they had happened. I hope my colleagues will forgive this slight departure from current practice. But I should, in defence, point out that my claim, first published in 1953, was a scientific prediction and therefore public property. It has since then been open to verification or falsification by anyone willing to spend a few weeks counting. Perhaps one of the fruits of the initial withholding of the details was the subsequent striking confirmation of the prediction by a relatively hostile observer.

When I answered Mr Oram's paper I still envisaged publishing the results of my counts on the Tippett tables. I don't think I shall do this now, for, though they are in some ways more striking than the rest, I do not feel, now that we know *why* such results occur, that it is necessary to establish by repetition of similar instances *that* they occur. In any case, observations by independent observers should carry more weight than my own.

Anyone who has worked with chance machines knows very well how difficult it is not to observe certain oddities in their behaviour; it is only that classical probability, not having a place for them, has always prevented our talking about these oddities in terms of chance. The fact that the machine does something noticeably improbable has not been connected with the fact that it *logically* cannot do anything else.

I am grateful to many friends for their help in making this book. Some of these I have already thanked in the text, but there are others, perhaps not mentioned there, to whom I am equally grateful: to Sir Ronald Fisher for his early encouragement, and to Professor A. C. Hardy for enabling me to begin work on this thesis in his Department; to Mrs Eileen J. Garrett of the Parapsychology Foundation Incorporated, and Mrs K. M. Goldney of the Society for Psychical Research, for their generous practical help despite the gradual divergence of my opinions from the tradition of their field; to Mr Robin Farquharson and to my supervisor, Mr W. C. Kneale, with whom many of the ideas here presented were first discussed and clarified; to Dr Handel Davies and Mr Garry Arnott for checking some of the algebra; to the Rev. Dr Eric Mascall and Mr R. H. Dundas for kindly reading the proofs; and to Mr John Howard Barton for making the index.

I should like, finally, to thank the Governing Body of Christ Church whose patronage has enabled me to put together and publish the work I have done.

G. S. B.

Oxford, December 1956

Acknowledgements

The thesis in Chapter X was originally given in a paper to the Mathematical and Psychological Sections of the British Association in September 1954. I am grateful to the organizers for encouraging me to present it in a publishable form. The gist of Chapter XIV is taken from a paper read to the Third London Symposium on Information Theory and published by Messrs. Butterworth & Co. (Publishers) Ltd. whose permission to re-present the argument here I acknowledge with thanks. I also acknowledge with thanks the permission of the Editor of the *Journal of the Society for Psychical Research* and Mr A. T. Oram to reproduce the paper 'An Experiment with Random Numbers' and subsequent correspondence.

CONTENTS

I

THE REAL WORLD

We distinguish between real people and characters in fiction, between real happenings and dreams, between real and toy soldiers, between real languages and Esperanto; we also talk about reality and the real world as if to distinguish it from the various forms of unreality implicit in fictional, substitute, toy or dream worlds. There is a further convention that the adjective 'real' somehow implies existence; and that by antithesis unreal worlds do not exist. This is clearly contradictory, for we cannot suppose that normal people wish to deny the existence of substitutes, toys, dreams or fiction, nor can we suppose they have no wish to keep up appearances. The confusion springs from the double valency of the adverb 'really'. The statement 'Merely apparent things don't really exist' is tautologous if 'really' qualifies their mode of existence, for merely apparent things no more exist as real things than circles exist as squares. But the word 'really' can qualify backwards as well as forwards, and in its backward capacity does no more than lend emphasis to the word 'don't'. With a better-behaved word the bother does not arise. We do not find it difficult to distinguish between the statements 'Complications don't happen simply', which is a tautology, and 'Complications simply don't happen', which is not. The assumption of the non-existence of the unreal can be blamed thus upon a logical pun.

Suppose there is a firm which makes toy mice. This firm employs the best inventors to make their products resemble the living animal. After years of endeavour they succeed in producing a toy mouse which is in every way indistinguishable from a real one. Once it has been made it runs about, reproduces and eats cheese just like any other mouse. In other words, it does everything that is *expected* of it, and therefore achieves reality. For, by Occam's razor, a fake that is good enough to be true must be true.

Let us now imagine the contrary: real mice deteriorate into automata. We must suppose the deterioration to take place slowly enough for the mice to carry the term 'mice' through the process. A few zoologists of an older generation might remark with disapproval that when they were young mice did not have to be wound up; but the majority, even if they were convinced by such tales, would continue for convenience' sake to call the automata 'mice'; and the younger generation would know no better.

In both these examples reality is seen as a function of expectedness. When a mouse does what we expect of mice as such it is real, but expectedness can occur only when what we observe does not change. If a mouse appeared substantially different in size, shape and colour every time we looked at it, we should have no name for it. There is no reason other than the common usage of the term 'thing' to suppose that there are not things which do this. They cannot be real things because they never do what we expect; which is another way of saying that they change too often or too quickly. The degenerating mouse, though changing, is changing slowly enough to build up expectation. It does not change noticeably from hour to hour, or even from day to day. It is by the slow-

ness of its change that it retains its reality. The reality of a phenomenon is a function of its expectedness as such, but its expectedness is an inverse function of its rate of change.

Science is concerned with the discovery of constants: it is the study of the changeless. If I drop a bomb from my top storey window, it will fall to the ground with an ever-increasing speed. This change of speed is anathema to the scientist. He may not rest content until he has found a way of picturing it changelessly. In this case he has not far to seek. The speed of the bomb may change, but the rate at which it changes (called the acceleration) does not. The function 32 *feet per second per second* is a constant which describes not only the behaviour of my bomb, but also that of other bombs dropped in the vicinity.

We talk of the function 32 *ft per sec*^{2} as if it were *absolutely* constant, but a little reflection shows that it is not so. The mass of the earth is slowly increasing as it picks up meteorites and interstellar dust. We may thus expect g, the acceleration due to gravity, to increase as time goes on. If we formalize this increase in terms of a further 'constant', we have no reason to suppose that this further 'constant' may not itself be changing. Our attempt at a perfect description of the acceleration due to gravity has ended in a regress.

It may seem that the regress could be broken by the following means. We suppose that statements involving the concept g are statements dependent on given masses, distances, and other factors known to be 'relevant'. Given all the relevant factors, we are in a position to formulate a constant which does not change. But the problem is now seen to be purely linguistic; any change in the constant made necessary by observation and experiment can be blamed upon our faulty assessment

of the *relevant* conditions under which the constant should be observed. In other words there is always a 'real' constant to which our observations tend: it just happens that when we think we have found it we discover afterwards that what we have found is only an approximation to it.

This latter way of talking is analogous to the philosophy of the thing-in-itself, or 'the reality beneath the appearance'. It could be called 'the constant beyond the approximation'. Such an assumption is indeed part of the scientific attitude and its convenience for some purposes remains undoubted. We shall discuss its usefulness later, but for the moment we must emphasize that the laws of nature are merely the descriptions we have made of structures which have been found to change only very slowly. We have, in fact, no evidence for the existence of any structure which does not change at all.

II

WORLDS AND MODELS

ANYONE who stops being a modern sophisticated philosopher and becomes for a while an old-fashioned wondering one soon finds it difficult not to wonder why there are any worlds at all. This sort of wonderment is expressible chiefly in relation to one or other of the real worlds, mystical, scientific, social, historical, etc., none of which has the appearance of necessity. (There is no logical contradiction in supposing them to be different from how we actually find them.) Take, for example, the astronomical world.

The framework is a kind of thin space; the space is infested with globular bodies moving about in it; the surface of one at least of these globular bodies crawls with living things; the living things go about mostly eating one another; and so on. 'But this is monstrous!' we think; 'It could have been *anything* else.' Very well, then. Let us try something else.

First, we had better make a list of some desirable properties for a universe to have. For instance, if we are of a serious turn of mind, we shall probably want to make it real. Next, perhaps, we should try to make it neither morally nor aesthetically offensive; and to the latter end we might like it to be symmetrical. Answering all these descriptions we could have, for example, a universe with nothing at all in it. Nobody could say that a nothing-universe was in any way offensive;

and it would have the advantage of being highly symmetrical. Furthermore, there is no doubt that, of all the universes we could construct, this one really would be changeless, and therefore, presumably, tremendously real. Unfortunately, once we had made it, there would be no one in it to expect it, and this would interfere with its reality. The nothing-universe, otherwise so desirable, is marred by this blemish. Let us try another.

Equally symmetrical, perhaps, would be the solid or everything-universe. It would be of heavy and stable construction, built to last, and, like the nothing-universe, containing no unreality. The trouble with this kind of universe, so full of everything, would be the difficulty of distinguishing anything. For example, its thick consistency would prevent our moving about in it, and this would make it difficult to distinguish between objective and subjective phenomena.

The universe we live in seems to be something between these two extremes. It is rather like an everything-universe where, mercifully, we are prevented from noticing everything at once. Observations in this universe could be considered as parallel runs between observer and observed. If I say that the desk before me remains the same desk, my statement can be interpreted, with apologies to William of Occam, as saying that whenever I change, my desk changes too. If my desk and I were both suddenly reduced to half size while the rest of the room uncooperatively stayed as it was, I might at first be at a loss to decide whether I had grown smaller or my room had grown larger.

Exactly what we notice can plausibly be ascribed to how, and especially how fast, we ourselves can change. We notice, for example, things which change as slowly as or more

slowly than we do, but not in general things which change much more quickly. Thus the faster we can change, the more we can notice.

If we take a cinematograph of a plant at, say, one frame a minute, and then show this moving picture speeded up to 30 frames a second, the plant appears to behave like an animal. When something is placed near it, it clearly perceives it and reacts to it. It is obviously a sentient being. Why, then, does it not ordinarily appear conscious? The answer is, perhaps, because it thinks too slowly. To beings which reacted eighteen hundred times as quickly as we reacted, we might appear as mere unconscious vegetables. Indeed, the beings who moved so quickly would be justified in calling us unconscious, since we should not normally be conscious of their behaviour. Such glimpses of it as might appear from time to time would mean nothing. A tree can no more perceive me walking past it than I can see a bullet flying past me. I might perceive certain events in the wake of the bullet, such as a broken arm; and similarly, if my passage were destructive enough, the tree might eventually perceive certain events in my wake, such as a broken branch. But what is fast for a tree is slow and boring for me, whereas what is normal speed for me is something out of this world for a tree.

We have spoken glibly of thinking and consciousness in a tree. If we think the matter over more carefully, we may come to doubt that trees think at all, although it does appear that, without stretching the word, they feel. If we put a cat in a cage with a latch to let itself out, it makes what appear to be random movements all over the cage until it accidentally lifts the latch. If we put a man, or even a monkey, in a similar cage, he might make a few random movements at first; but

then he might retire into a corner to 'think things out'. After this he might go straight to the latch and open the cage. We note also that in subsequent experiments the cat now goes straight to the latch and lets itself out as quickly as the man. It is the initial behaviour which is different.

We describe this difference by saying that the man and the monkey have 'insight' into the problem, whereas the cat has none. The cat does not think the matter out, but moves about at random until something advantageous happens. We can say therefore that it is not sufficiently *conscious* of the problem to *think out* a solution. But what is this thinking out process?

If we are to accept what psychologists say, to think one's way out of a difficulty is to construct inside oneself a kind of working model of the difficulty which one then treats as if it were the difficulty itself. For example, the man in the cage somehow builds inside himself a working model of the cage upon which he experiments much as the cat experiments on the real cage. The man thus has an obvious advantage over the cat; experimenting with something small and conveniently near, he can complete his operations on it more quickly than the cat which is experimenting with something relatively large and cumbersome outside itself. But we have no evidence that the conditions differ in other respects. There is no reason to suppose, for instance, that the man's experiments on the model are any less random than the cat's experiments on the cage. Indeed, if we have to construct a calculating machine which, for example, plays chess, we have to build into it some sort of randomizing or scanning device to enable it to experiment with the model chess set which has been constructed within it. So, if this is what thinking is, the machine clearly thinks. Equally, it must be *conscious* of the

particular problems it has been built to solve. But, though conscious, it is improbable that it has any *feelings*.

Apparently these remarks do not apply to a tree, however much we speed it up. The tree may react as quickly as we like, and can therefore be said to feel, but there is not much evidence that it ever thinks things out in order to decide between alternatives. It does not appear to make models. Like, say, the jelly-fish, it is clearly a sentient being but probably not a conscious one.

We see that, given a modicum of insight, the extent to which we can subsequently know things depends upon the speed at which we can move and react. A very quick cat might open the cage sooner than we could think out how to do it. Anyone who could move infinitely fast would be in a position to know everything, because to him nothing would move. He would have an infinite time in which to learn it. And if he were also allowed to move bits of the universe himself, he would not only be omniscient, but also omnipotent, since he would have as long as he liked to beetle about altering things.

We have seen that science seeks to reduce change to an unchanging formula. Wherever there *has been* change, such formulae can always be found; but they do not always apply to the future. When the change itself changes, we need a new formula.

Now science is not interested in all unchanging things. Initial unchanging observations are often taken for granted. The scientist does not so much seek descriptions of the changeless as changeless descriptions of the changing. Indeed, he even seeks change so that he may exercise his capacity to freeze it in a formula.

Forms of change which are not immediately obvious may often be rendered so by slowing up, speeding up or other such processes which bring them within the range of the scientist's normal perception. The rules of such 'processing' are simple. Besides the obvious adjustment in rate of change, we have adjustments in size. What is too large is made smaller. What is too small is made larger. We have models on a very much reduced scale to represent, say, the solar system; we have maps at 1 inch to the mile to represent the surface of the earth. We have microscopes and telescopes to enlarge what would otherwise be very small in the visual field. Complexities are also processed; what is too complicated we simplify. This is a purpose of statistics. Sometimes, even, we complicate what is too simple. For example, some Freudian explanations of simple behaviour might be described as over-complicated. This processing, or building of working models of the right size which run at the right speed and are reasonably simple, is the first aid to thought and possibly even the substance of thought itself. And it is characteristic of *experts* in any field that they be accomplished processers.

We are now ready to differentiate between the task of the historian and that of the scientist. The scientist, we have seen, is concerned with recording in a changeless way phenomena which are still changing; whereas the historian as such is concerned only with recording changes that have already stopped. The historian is not concerned to find a formula which will work from henceforward for all time. If he ever found such a formula, no more records would be necessary and he would lose his job. It is not history which repeats itself, but science. The scientist begins by looking at the welter of change and fixing in formulae whatever parts of it

he can. History is what is left over after the scientist has taken his pick.

History is therefore more fundamental than science. It is our first appreciation of things. But its study is not urgent. That which does not change, such as the past, is not dangerous. As such, it cannot harm us. But we must beware of what changes. And in order that we may adapt ourselves to it, our senses must be quick to sense it. The compound eye of a housefly is specially adapted to the perception of movement; stationary objects are ill represented in its multiple picture. What does not move cannot usually catch the fly. The fly therefore disregards it. That is why flypaper is so effective. The perception and understanding of movement is necessary to the survival of most animals. 'If it moves, salute it' is a biological imperative whose fundamental application is not confined to the quarterdeck.

Science never catches up with history. For every observation which repeats itself, there are countless observations which do not. If science is an examination of the lawfulness of nature, then history is an account of its chaos. Both scientists and historians make models of facts, and in a moment we shall see how their models differ.

Suppose we say that a yellow flame and caustic white fumes is a model for '$4Na + O_2 \rightarrow 2Na_2O$'. This, someone is sure to say, is the wrong way round. We should say that the formula is the symbol or model for the combustion of sodium and not the reverse. But are we always as sure as this which is the model? Voltage, we might say, is analogous to water-pressure. But if I were familiar with electricity and unfamiliar with hydrostatics, I might use the concept of voltage to understand that of pressure in a fluid. Which is the model and which is the

thing modelled depends on the point of view. A map 1 inch to the mile is a model of a part of the surface of the earth. But what about a map 1 mile to the mile or, worse, 10 miles to the mile? And what if I have made a perfect scale-model of a flea, one-tenth natural size? Whenever it is convenient we use one part of nature as a model or symbol of another. A mango tastes rather like a peach, and a peach tastes rather like a mango; neither of these tastes is the more fundamental. If we want to know about something new, we ask 'What is it *like*?' If there is no convenient natural model to hand, we construct one. A natural model is often called an analogy; constructed ones are called descriptions, theories or models. But in all this procedure, there is no invariable one-way relationship between the model and the modelled, the symbol and the symbolized; each can be either. 'Wir machen uns Bilder der Tatsachen.' 'Das Bild ist eine Tatsache.'

Moreover, there is no important distinction between a model which is entirely 'in our heads' and one which is partially or wholly outside. The shop assistant who gives us change for a ten-shilling note may construct in her head a model of half-crowns, florins, shillings, etc.; and, having worked out the problem by what we call mental arithmetic, select the correct change. Alternatively, she might take pencil and paper to help her with her calculations. There is nothing fundamentally different about these two operations; in the latter she is employing an extra feed-back circuit, making external traces which can be returned to the central nervous system through vision; whereas in the former case all the feed-back circuits are internal. She has yet a third course, that of counting out the change in terms of the coins themselves. This method involves least thought since it dispenses with

most of the model. It is, on a higher plane, the cat-and-cage method.

Nor is there a need for any model to be literally like the thing modelled. Symbolization is a form of modelling, but we can say with some plausibility that the word 'cow' does not look like a cow. Equally, the sentence 'The cow has climbed a tree' means but does not resemble a certain kind of improbable fact. But there is a sense in which the word 'cow' does resemble a cow. It resembles it in the sense that a thundery sky looks like rain. In general, there are two aspects of likeness. Likeness consists, first, in associations made instinctively and, secondly, in associations made through use and training. When Wittgenstein says that in order to be a picture a fact must have something in common with what it pictures, he is only making a tautology.

Thus, to say that science looks for models is to say that science looks for likenesses. But this is not enough, for so does history. The difference is that science looks for functional likenesses. It is not primarily interested in looking for names or rigid descriptions; it has a special kind of description of its own called a *function*. And a function is a working or workable model of a changing or changeable state of affairs.

We have seen that we seek information about a new thing through the question 'What is it like?'. We may note in passing that it must be logically impossible to obtain information about the whole universe. The universe by definition includes everything, so we cannot expect an answer to the question 'What is it like?', because there is nothing left out for us to liken it to.

To find or make up likenesses is to find or make up relationships, which is another way of saying what modellers are

doing. A relationship is a way of classifying. If I throw a die thirty-six times and the six-face turns up twenty times, I can describe or model this state of affairs by saying 'The die is biased'. If I now take a different die and break it open, finding inside a weight displacing its centre of gravity away from the six-face, I can make a similar model: 'The die is biased'. With a third die I note that it has a small six-face and a large one-face. Again my model is 'The die is biased'. All these observations, for the purpose at hand, are classed the same. But they need not be.

Suppose I now picture the thirty-six throws with twenty sixes as a black and white pattern. I then take a book from the shelf and open it, noting a further black and white pattern. I watch a game of chess, and again my model is 'black and white pattern'. Into this latter class only one of the former examples goes. But which of these classifications is right and which is wrong? The right classification is merely that which is the most convenient for the purpose at hand; we make what we consider to be useful classifications, and persuade others to do so.

III

THE CLASSICAL PROBLEM

LET us consider two forms of argument.

(1) *If* all swans are white
Then some swans are white.

The conclusion that some swans are white is clearly a valid inference from the premiss that all swans are white. The general rule for valid inferences is that their conclusions may say as much as or less than, but not more than, their premisses. The conclusion above says less than its premiss and the inference is therefore valid.

(2) *If* some swans are white
Then all swans are white.

Here the conclusion says more than the premiss and the argument is clearly invalid.

The former argument is an example of what is called tautology. The essential property of a tautology is that it is always true. For example, the statement 'If all swans are white, then some swans are white' remains true even if it happens to be the case that all swans are black.

The opposite of a tautology is a contradiction. Argument (2) is not a contradiction. Here is a contradiction:

(3) *If* all swans are white
Then some swans are not white.

A contradiction is false whatever is the case. Statement (3) is false whether swans are black, white or indifferent.

We see that neither a tautology nor a contradiction can give us any information. This is because they are respectively true and false whatever is the case. Thus statements (1) and (3) give us no information, whereas statement (2) can do. Whereas (1) must be true and (3) must be false, statement (2) can be either; it is contingent. This means that its truth or falsity depends upon the particular circumstances of the case; we cannot tell which it is by just looking at it. Take the statement

(4) *If* some of the balls in this bag are white
Then all of them are.

In the contingency of my having arranged that the bag was to be filled either with all white balls or with all coloured balls, the statement is true. In the contingency of my having arranged for it to be filled with some balls of each kind, it is false.

Now let us reconsider statement (2). Here the contingency is not of human arrangement, but of natural arrangement. That we consider Nature to be in some sense arranged is illustrated by the mention of the 1953 total eclipse of the sun in *The Times*'s column 'To-day's Arrangements'. That we consider the sense in which the word 'arrangements' was here used to be slightly odd is illustrated by the letters to the Editor which followed. Arrangements by man and the higher animals are one thing, arrangements of natural phenomena are another; and even if we assume a creative God it is not now

fashionable to think that He arranges the swans on a lake as we might arrange billiard balls in a bag.

But the assumption that we shall find, in some parts of nature at least, the appearance of their having been arranged is what enables us, on seeing a part of a pattern, to predict the rest of it. The statement 'If some swans are white, then all swans are white' reflects this procedure. As we have already pointed out, it is an invalid argument. We cannot conclude from the premiss 'Some swans are white' that all swans are white. Yet, so it appears, we are constantly making this sort of argument in science. Having taken several samples of sugar and found that each sample dissolves in water at 15° C, we conclude that all such samples of sugar will dissolve in water at 15° C.

The question that at once occurs is that if it is an invalid argument which reflects inductive procedure, why should it be just this type of invalidity? Why not, for example, the type of invalidity exemplified by the sequence

(5) *If* all swans are white
Then some cats are black.

In this example the contingency is total: there is no connexion whatever between the premiss and the conclusion. Whereas in example (2) the conclusion did at least imply the premiss, although the premiss did not imply the conclusion, in the last example there is no implication either way. This completely disconnected form of argument does not normally represent induction. It does sometimes, but this is only when we have reached our wits' end and are willing to try anything. Normally we are concerned with a type of invalidity which, for some reason, does not seem so shocking. This may be

because there is a simple additional premiss which renders example (2) valid; namely, that events are generally arranged so that if one of them has a particular quality, then so have all the rest. There is no such general principle which could apply in example (5).

The extra premiss of this kind is usually called the Law of Uniformity of Nature or the Law of Universal Causation. Attempts to justify bringing it in generally use the argument that, unless there were some such law, it would not be possible to know anything by induction; but as we clearly do know some things, then the Law of Uniformity must be true.

The trouble with this argument is that if we are asked to justify it we can only appeal to experience; and such an appeal is itself inductive, since it presupposes that our past experience in the application of the law will continue to hold in future. We are reduced, therefore, to justifying inductive procedure by induction, and this is begging the question. But there is worse to come.

I have before me a desk. My general experience of this desk is that various objects such as my writing-paper and my telephone rest upon it without falling to the floor. I find also that if I bring my hand down from above, it stops and meets resistance when it touches the surface of the desk. Suppose one day I bring down my hand as usual and find that it passes through the top of the desk. My first reaction is one of surprise. I become suspicious, thinking that the Law of Uniformity of Nature might have been broken. (The Law demands that the previous solidity of my desk will continue.) I look round at other objects in my room and am disturbed to find that some of them have disappeared while others have changed their shape. My feeling of disturbance is aggravated

by a small object, probably a mandrake, appearing from the fireplace and walking towards me over the carpet.

The classical procedure in such cases, though seldom remembered in the confusion of the moment, is symbolic of our general attitude. It is to pinch ourselves. This is a sort of casting vote. If I pinch myself and find that it does not hurt, then clearly this is an additional example of non-uniformity and I decide that I am not observing a real or natural world at all, but having a dream. And who would be so silly as to cite a dream as evidence against the Uniformity of Nature?

But this is precisely the point at issue. It is by its non-uniformity, and this only, that we recognize the dream; and conversely we recognize nature by its uniformity. The Law of Uniformity cannot be falsified; for when our observations are not uniform, we say we are not observing nature. In other words, it is only *real* things which we allow to count in verifying the Law. But this makes the Law of Uniformity a tautology. Laws of nature are information, but tautologies say nothing. Therefore the Law of Uniformity cannot be a law of nature.

IV

DISSOLUTION OF THE CLASSICAL PROBLEM

I STICK a pin into a balloon and it bursts. Pleased by this discovery, I stick pins into other balloons, which also burst. After a time I am tempted to say, 'I know something about balloons. If you stick pins into them they burst.' 'How do you know?' asks a sceptic. 'I have tried', I reply. 'How do you know they didn't just *happen* to burst when you stuck pins into them?' he asks. 'Because I did it lots of times', I say. If he still looks sceptical, I am tempted to bet him half-a-crown that the next balloon I stick a pin into will also burst.

These remarks summarize our attitude in the inductive set-up. First, we make a chance discovery of a conjunction of events. Secondly, we look out for and find—or even try to engineer—its recurrence. Thirdly, we assume that the two events conjoined are somehow causally related, which means that we are willing to bet that they will appear together in the next instance we observe.

Suppose now I see a round coloured shape which I think is a balloon; and suppose that, advancing towards it with my pin, I discover not only that it does not burst, but also that it is not a balloon at all but a Chinese lantern.

'You owe me half a crown!' says the sceptic who, it turns out, has been watching all the time.

'Not at all', I say crossly. 'This object is clearly not a balloon, so the bet is off.' He asks me why I think it is not a balloon, and I point out its crinkliness, its failure to burst, etc. 'But supposing', says the sceptic disconcertingly, 'it *had* burst?' 'Then', I say, 'it might *just* have been a balloon. But it didn't, so it wasn't.' 'You are being unfair', says the sceptic. 'You are making it so that you can't lose your bet. You are saying that if a thing doesn't burst when you stick a pin into it, then it can't be a balloon.'

The sceptic is right; this is exactly what I am doing. But this is current scientific practice. If we discover that several samples of compound *A* melt at x° C, we not only induce that all samples of compound *A* melt at this temperature, but if we find a sample which melts at some different temperature we simply say that it is not, after all, a sample of compound *A*. This is another elementary but fundamental procedure for making our inductive predictions turn out right. Where we can't use the dream-excuse, we merely discount as *irrelevant* the cases where they turn out wrong.

It can of course be said in these cases that we have ceased altogether to make an inductive inference, since now the conclusion has become a means of *recognizing the premiss*. A proper induction must be falsifiable. It must tell us something we don't know already; if it can't be wrong it can't tell us anything. The example above has succeeded in doing what philosophers have felt ought to be done for years, that is, it has reduced induction to deduction, and this won't do at all. Let us consider the balloon again.

I think I see a balloon. But as I approach, it is not a balloon which appears before me, but a Chinese lantern. Disappointed, I remark, 'A moment ago this object was a balloon; but

before I had time to stick my pin into it, it turned into a Chinese lantern'.

What is wrong with this statement? After all, left with my first impression and without subsequent examination of the object in question, I should say I had seen a balloon. That I now see a Chinese lantern, though it could mean that my first observation was mistaken, could equally well mean that the balloon had changed its state. Someone might have put a spell on it, but instead of saying this I make a retroactive reassessment; it is a Chinese lantern now, therefore it must have been one before.

But what an arbitrary assumption! Why not assume that as it was a balloon before, it must still be one now? The answer is, I suppose, that we must consider convenience and bow to the present; and if later on the object in question turns back into a balloon, we must not then suggest that it had ever been anything else. A resemblance will doubtless be noted between this procedure and the activities of the Ministry of Truth in Orwell's *Nineteen Eighty-four*.

I have shown history in the unflattering light of being what is left over after science has taken its pick; but scientists in their turn need not feel complacent in their questionable practice of rewriting history.

'I saw a balloon which turned into a Chinese lantern.'

'Oh, no, you didn't', says the scientist, fixing me with a glassy eye.

'Yes, I did', I say. 'I suppose someone put a spell on it.'

'This is the twentieth century. There are no such things as spells', he says sternly.

'Oh', I say, disappointed. 'When did they finish?'

'There never were such things', he says irritably. 'People only *thought* there were.'

'I suppose you are going to say next that I only *thought* this Chinese lantern was once a balloon?'

'Exactly', he says.

'But what if I say it really *was* a balloon?'

'Then', says the scientist menacingly, 'there is always the lunatic asylum.'

'So you would liquidate me for deviation?'

'Precisely', he says.

Retroactive reclassification of observations is one of the scientist's most important tools, and we shall meet it again when we consider statistical arguments. But it is notable that it bears the status merely of a linguistic convention. Either we say things are just as they seem, and that 'nature is constantly changing', or we say that they are really much more constant than they seem, their manifest inconstancy being accounted for by our repeated errors of observation. If we use the former way of speaking, our language will be full of terms like 'hobgoblin', 'leprechaun', 'poltergeist', 'magic spell', 'miracle', and so on; whereas in the other language we find terms like 'misjudgment', 'illusion', 'hallucination', 'delirium tremens', 'psychosis', etc.

Some people would like to ask which of these kinds of description was the *true* one; but there is no sense in this. Both descriptions are ways of saying what we observe, and can therefore be true. We might be able to say which was the *right* description, just as we could say which was the right way to behave at a dinner party. But this has nothing to do with truth, and the decision to use one or another type of description must be made according to the

dictates of fashion, Occam's razor or some other rule of thumb.

Examination of primitive languages shows at once that they are not designed to foster belief in the Uniformity of Nature. If we do consider nature to be uniform, then in order to support our belief we must either (*a*) not count the bits which appear otherwise (for example, laugh them off as dreams, silly mistakes, etc.) or (*b*) not talk about them, or talk about them only in whispers and in such a way as to assume that they are uniform all the time in spite of appearances. This sort of behaviour is called science.

The problem of induction and analogy is already beginning to pale, but we must return to it.

(1) *If* some swans are white
Then all swans are white.

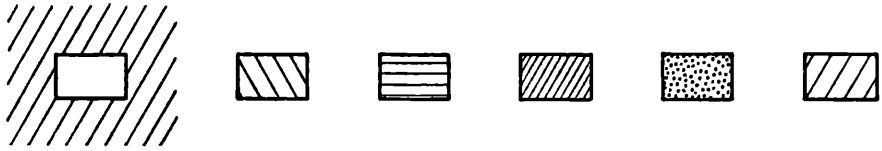
(2) $s = ut + \frac{1}{2}at^2$ (always)

These are both inductive arguments. They are both formulae applied to states of affairs. If, for example, the distance-acceleration formula ceased to apply, we could either say that acceleration, etc., had changed its character (the natural order had changed) or refuse to count such instances as did not fit the formula (as errors of observation or dreams). If all mountains gradually turned into molehills we could either say that this was what had happened or defend the natural order by saying that 'Mountains and molehills, as everybody knows, are the same thing'.

In formula (2) we note that s is a function of t. This means that, given the nature of the function and the value of t, we can *deduce* the value of s. Thus the natural phenomena represented by formula (2) look as if they are being represented

by deductive arguments. But the patterns of deductions are tautologies, and, as Wittgenstein succinctly remarks, all tautologies say the same thing, that is, nothing. Does, then, formula (2) say nothing? The answer is, I think, that it says nothing in the same way as any name, such as 'mountain' or 'unicorn' says nothing. This does not mean that we may not use it, if we wish, *of something*. And of a changing set-up we may find it convenient to use a functional name; that is all.

Let us consider what is called the 'Progressive Matrix' intelligence test.



The person tested has to interpolate the proper square into the space. The test could be given the other way round, so that from a small patterned rectangle he had to extrapolate the surroundings. Induction and analogy are to time what extrapolation and interpolation are to space. But there is no sharp division. I peep through a knot-hole in the cow shed and see a bit of a cow, from which I induce or analogize the rest of it. If I am a doctor I see a bit of a disease (a symptom) and at once recognize the whole of it. 'Recognize' is the operative word. It is the word we use when we do the operation quickly enough. We call it induction or analogy only when we do it very slowly. Induction and analogy are the processes of recognition observed in slow motion.

V

TRUTH

ONE of the tests of an informative statement is that it need not be true. If we can see it is true by merely looking at it, then it is not information. The statement 'Ripe tomatoes are red' is true by virtue of the fact of the redness of ripe tomatoes, and would be untrue if ripe tomatoes were, say, blue.

Some people have attempted to define empirical truth as that which follows from a statement's accordance with the facts. This account of empirical truth is circular; we mean no more by 'accordance with the facts' than we do by 'truth'. Let us analyse the statement '"Ripe tomatoes are red" is true'.

The first thing we see about this complex statement is that it can be expanded or contracted without gain or loss to its sense. When I say ' "Ripe tomatoes are red" is true', I mean no more than 'Ripe tomatoes are red'. Putting 'is true' after it may serve to emphasize what I say, but does not add to its meaning. Similarly, if ripe tomatoes happen to be red, then not only is the statement 'Ripe tomatoes are red' true, but so also are the series of statements ' "Ripe tomatoes are red" is true', ' " 'Ripe tomatoes are red' is true" is true', ' " ' "Ripe tomatoes are red" is true' is true" is true', etc.

What strikes us about these transformations is that they involve different hierarchies of language. If we are to have a

language of the first order, then the words 'true' and 'false' have no significance. We use our first order language automatically to describe what we see around us. Provided we do not examine our descriptions critically, there is no question of truth or falsehood. Suppose now we observe one of our contemporaries looking at a ripe red tomato and saying, 'It is blue'. This might shock us into criticism. We should think that he was using the language wrongly, and for this we might invent the word 'false' to describe his statement. When we had found such a word for the wrong use of language, we should have to find another for its right use.

The concept of truth thus arises first through any use of a language which might be misleading. The well-known paradox of the liar depends upon this critical function of the concept of truth. The statement 'This statement is false' is paradoxical because if it is true then it must be, as it says, false. But if it is false, it cannot be false. The paradox can appear only if we are confused enough to imagine that truth and falsehood are properties which can be described in a single non-hierarchical language. If we accept the words 'true' and 'false' as applicable to a language in a critical appraisal of its description of a state of affairs, then we at once see that the statement 'This statement is false' is not allowable; for here there is no question of an appraisal of the way a state of affairs is being described. There is a piece of language and nothing else. It describes nothing, and so may not be true or false. It is in a class with tautology and contradiction, though it is not either.

I am sitting in a room with a friend, and we notice a cat walking across the floor. The next day I say to my friend, reminding him of this instance, 'Yesterday my cat walked

across the floor.' 'Yes', he might say, 'that is true.' Suppose now it turned out that my cat had been with neighbours all day yesterday, who had fed it. And suppose that at the time we had thought it walked across the floor, they had been feeding it and said so. Now we should have to modify our views about the truth of the statement 'My cat walked across the floor yesterday'. The least drastic kind of modification would be to suppose either that it was not my cat walking across the floor, or that the neighbours were feeding someone else's cat by mistake. In other words, either myself or my neighbours must have been wrong. As both of us thought we were right, then clearly neither of us could assert what we did with certainty.

It is impossible to imagine an empirical statement which could not be put in doubt by some such circumstances, so we must arrive at the fairly obvious conclusion that no empirical statement is certainly true. Empirical uncertainty disturbs some people, and many attempts have been made to get rid of it. Descartes, for example, in his statement 'Cogito ergo sum' imagined he had done the trick. Actually, 'trick' is about right, for his certainty is only logical. If the word 'I' is to have any meaning, the object of its designation must exist. The verb *to be* unqualified denotes only existence. The word 'I' is self-referential and therefore must refer to an existing person. To put 'am' after 'I' therefore adds nothing. The statement 'I am' already follows from the statement 'I think' before we have, as it were, had time even to say 'think'.

One of the most recent attempts to find certainty in empirical knowledge has been by the phenomenalists in their use of what they call the sense-datum language. Here some observations are supposed to be certain. It may not be possible for me

to say with certainty that my table is black, because I might be having hallucinations. But if I make the remark 'I am seeing black now', this, it is held, is certain; for I am supposed to be capable of describing what I see whether it be a hallucination or not.

There are several objections to this point of view. In the first place, what is 'now'? I must have been seeing black for some time to be able to remark 'I am seeing black now'. Indeed, if 'now' represented the present time as known to physics—i.e. a single point in time—then I could not say with certainty 'I am seeing black now', because it might have ceased to be black before I had finished my sentence. Supposing that whenever I saw black I emitted the single word 'black'. If black came suddenly and lasted only a very short time then when I said 'black' it would have ceased to be black and I should be lying. In these circumstances, the only way I could tell the truth would be by chance. I could, of course, make sure of telling the truth sometimes by saying 'black' all the time, but this would not be giving any information away. It would be like making a tautology, covering the whole universe with the quality blackness in order to be right about some of its parts. Equally, no one cares about being right by chance. He wants to know beforehand that he is going to be right. Information derived by chance is considered to be of as little value as tautologous truth. If I were asked where one of my colleagues happened to be, I should not be considered helpful if I said that he was either in his rooms or somewhere else. But I might be even less relied upon if I were known in such circumstances to make a wild guess which might happen to be right.

We see, then, that even the sense-datum language need not provide us with true information. The next question we

should ask is, Does it provide us with *any* information? What is the information contained in 'I am seeing black now'? Either it says something or it doesn't. If it says something, then it must refer to some quality or observation, in this case blackness, which someone besides myself can also sometimes see. If he cannot, then my sentence means nothing to him. But if it means anything, then we have at once left the sense-datum language and are talking in terms of objective qualities or even material objects. It now begins to become clear why phenomenalists, who use, or suppose they use, the sense-datum language, have nevertheless found it always impossible practically to translate any material object sentence into their language. They maintain that it can be done in principle, and that the only difficulties in their way are complications. Someone who could see through these complications, they hold, would be able to do it. But we have shown above that (*a*) the sense-datum language can communicate no information, and is therefore not a language, and (*b*) that even if it could communicate information, that information would not necessarily be true.

If sense data are an unsuitable basis for scientific conversation, we must try to find something better. The material object hypothesis is more useful, besides being of historical interest. It is not, of course, impugned by the idealistic criticism of Berkeley, but is still somewhat too specialized for the purpose we have in mind.

If I am summoned before the magistrates for 'failing to observe a constable's signal', it is no defence for me to prove that at the time I observed to my companion, 'By Gad there's a constable's signal!' It is not my lack of any visual experience which annoys the constable; it is my lack of

appropriate reaction. The word 'observation' covers a wide field, referring as it may to an experience, to a reaction or remark, or even to a material object itself. And it is observation which is, I suggest, the primitive concept of science.[1] Observations form a wider class than material things, but are no less objective. I can observe through my senses, or through extensions of my senses in instruments (e.g. pointer readings), or through further extensions in accounts by other people of their sensations. It may be complained that the word 'observation' is ambiguous, signifying either the experience or the remark; but this is exactly what the scientist wants; he has no use for any experience without any remark; experience without remark is not science.

[1] Cf. von Mises, R., *Positivism* (Cambridge, U.S.A., 1951), p. 147.

VI

MEASURING PROBABILITY

CONSIDER the statement 'heads or not-heads'. It covers all possibilities and is therefore tautologous, saying nothing. The statement 'heads' by itself covers a definite class of possibilities and is therefore informative. Thus the statement

'*If* heads
Then heads or not-heads'

is itself tautologous because its premiss, which says something, says more than its conclusion, which says nothing. If we represent 'Heads' by h and 'Not-heads' by $\bar{h}$, the strong 'or' by $\not\equiv$ and the *if then* implication by $\supset$, we can rewrite the statement thus

$$h . \supset . h \not\equiv \bar{h} \qquad (1)$$

We cannot, of course, say with certainty 'If heads or not-heads, then heads'. We might be right sometimes but we are making a wild guess: and in logic, wild guesses do not follow. Rewriting the latter statement

$$h \not\equiv \bar{h} . \supset . h! \qquad (2)$$

illustrates succinctly what we have done; the shriek-mark is not of course to be interpreted as factorial h, but as a sign drawing attention to the unusual boldness and riskiness of the romantic assertion we have just made.

Supposing I say 'Heads your team bats, not-heads mine does'. I then toss the coin in a carefully selected spot so that it falls down a drain. 'Not-heads', I say, 'so we bat.' This is unfair. Occurrences like the coin's falling down a drain or getting hit by a meteorite and vaporized are not supposed to count. The proper procedure is not 'heads you bat, not-heads we do', but 'heads you bat, *tails* we do'. The strict probability model, then, is

$$h \not\equiv t . \supset . h! \qquad (3a)$$

or

$$h \not\equiv t . \supset . t! \qquad (3b)$$

There are in this case two assertions we can make, one of which must be right because we are not going to count anything else.

Suppose now we put six snooker balls in a bag: black, pink, blue, green, brown, yellow. The probability of drawing the black one is conventionally $\frac{1}{6}$.

$$p \not\equiv q \not\equiv r \not\equiv s \not\equiv t \not\equiv u . \supset . p! \qquad (4)$$

Note that this time to be sure of being right we must make no less than six separate statements. Thus the probability of any conclusion in this sort of entailment is the reciprocal of one more than the number of disjunctions in the premiss. If the probability relation is, as Keynes and Wittgenstein suggest, a kind of partial entailment, we could rewrite (4) thus

$$p \not\equiv q \not\equiv r \not\equiv s \not\equiv t \not\equiv u . \supset_{\frac{1}{6}} . p$$

But it is, I suggest, over-restricting the meaning of 'probability' to say that it is merely a sort of entailment. The word is used also to designate a *property* of *events*. 'The probability of the event called heads is $\frac{1}{2}$' is a perfectly significant statement. It is also meaningful to speak of the probability of an

observation: 'The probability of observing a parhelion is small.'

We have put our coloured snooker balls into a bag and define the probability of drawing the black one as $\frac{1}{6}$. If we now put in our hand and draw out, say, a white rabbit we do not revise our estimate of the probability of drawing a black ball. The probability-field has been defined in terms of balls only, and white rabbits do not count. But if every time we put in our hand we drew out a black ball, then we might eventually revise our estimate of the probability of this occurrence. Similarly, if in a long series of draws the black ball appeared much less than $\frac{1}{6}$ of the times, we should again be inclined to revise our estimate of the probability of its appearance.

One thing we are not allowed to do in all these procedures: we are not allowed first to take a good look into the bag, and then to draw which ball we please. The choice must be *random.*

VII

PROBABILITY AS AN INDICATOR

THE concept of randomness arises partly from games of chance. The word 'chance' derives from the Latin *cadentia* signifying the fall of a die. The word 'random' itself comes from the French *randir* meaning to run fast or gallop. We shall see later how this seemingly unconnected concept of speed came to be associated with the behaviour of a chance machine.

The concept of randomness is basic to the concept of probability itself, and the study of probabilities was first inspired by the need to calculate betting strategies in games of chance. Here calculations are based on a concept of probability strictly interpreted as the ratio of the number of cases considered to the number of cases considered possible. For example, in tossing a coin, two cases, heads or tails, are considered possible. If we consider one of these cases, heads, the probability of its occurrence must be $\frac{1}{2}$. Another example: suppose we are asked the probability of throwing 11 or more with two dice at a single throw. The number of possible occurrences (i.e. those which we are going to count) is clearly 36. (For each of the six faces the first die can turn up, the second can turn up one of six.) Out of these 36 possibilities we consider the following. First, the one die reads 6 and the other 6. Secondly, the one reads 6 and the other 5. And

lastly, the one reads 5 and the other 6. We thus define the probability of a score of 11 or more in a single throw of two dice as $\frac{3}{36} = \frac{1}{12}$. This basically simple method is adhered to in all comparisons of probabilities in games of chance.

We have assumed so far that the chance-machines we have been using are *unbiased*. This means, roughly speaking, that they will not 'in the long run' favour unduly any of the cases considered possible. For a biased machine we need to alter the model slightly: this we shall do in a later chapter.

But suppose, now, we wish to discover whether or not a particular coin is biased. We devise an experiment whereby the coin is tossed up and allowed to fall freely 100 times. If as a result of each of these 100 tosses it falls heads 99 times, we have little doubt that it is biased towards heads. But if it falls heads only 51 times we do not feel that the experiment had provided much evidence of a bias towards heads. These extreme examples provide an obvious indication or lack of it that the coin is biased. But how are we to assess results ranging somewhere between these extremes? Suppose, for example, the coin falls heads 60 times and tails 40 times. Should we now be justified in believing it to be biased? The answer to this is not obvious and demands some analysis of the question.

First of all, what do we mean by 'biased'? The reason we don't suspect a coin of being biased when it falls 51 times heads and 49 times tails is that we don't suspect on this evidence that the coin will, on future occasions, almost invariably give us an excess of heads over tails. On the other hand, an excess involving 99 heads and only 1 tails is so large that it leads us to expect that at least some sort of excess will occur in similar tossing experiments. These expectations may be partly intuitive, but they are at least partly the result of

previous experience. We know that large excesses tend to repeat themselves more than small ones.

The commonly used criterion of whether an excess of heads or tails is likely to repeat itself on future occasions is an interesting one. A combination-classification is made of all the possible arrangements of heads and tails which could occur as a result of tossing the coin 100 times. The classes are then arranged in order of bias; into the first class go all arrangements, into the second class go arrangements containing at most 49 heads or 49 tails, into the third class go arrangements containing at most 48 heads or 48 tails, and so on up to arrangements containing at most no heads or no tails. Any particular arrangement of heads and tails found in practice is placed in the largest possible class. The ratio of all possibilities in this to all possibilities is then taken as the probability of a bias of or greater than the order observed in the particular case. If this probability is reasonably small, say $\frac{1}{20}$, the particular bias is suspected of being causal rather than chance and the result is said to be 'significant'. To say such a result is significant is to suppose that in similar circumstances a similar excess will recur. But there is no obvious reason why this should happen.

It is commonly supposed that, if the criterion of significance is taken at $p = \frac{1}{20}$, then, having achieved in our first experiment a result which gives this exact value, we should expect to be wrong about the recurrence of a similar-sensed excess only once in twenty such initial experiments. But our final p-value depends entirely on what sort of bias we are looking for in the series of heads and tails; and in the case in question we have considered only elementary bias. There are many kinds of compound bias (bias of pairs, trios, and so on)

which we might consider relevant; and if we did consider any of these biases relevant, we should have to alter our p-value accordingly. So it appears that not only is the *level* of our significant p-value an arbitrary choice, but also that the p-value itself arbitrarily depends upon what kind of bias we are considering.

Suppose I have an acquaintance called Mr X. Mr X rarely comes to see me, but arrives on a particular occasion when I have a guest whom he very much wants to meet. His behaviour is open to two explanations: either it is one of his rare visits which luckily for him happens to coincide with the visit of the person he wants to meet; or it is a special visit causally arranged through leakage of information about my other visitor. Suppose now we have to decide whether his visit was causal or otherwise and are prevented by etiquette from asking him directly. Thus set, the circumstances of the problem provide us with two relevant criteria upon which to make our decision. These are as follows.

The first criterion is simply a measure of the rarity of Mr X's visits. The rarer they have been in the past, the more we suspect that his present visit is not a chance one but causally arranged. The other criterion consists in my knowledge of the intensity of Mr X's desire to meet my other visitor: the greater his desire, the less inclined I shall be to believe that his visit is a chance one. So, in making my decision, I must weigh together these two criteria.

The former criterion will be seen to be the simple measure of statistical significance. It would in fact be possible for me to give it a p-value: if, say, Mr X's visits in the past had worked out at an average of one every hundred days, the 'significance' of his visit to-day could be given at $p = 0{\cdot}01$.

But this does not end the matter. I must now bring to bear upon the problem my knowledge relevant to the second criterion: that of the reasons why Mr X should make a causally arranged visit on this occasion. If I have no reason to believe he wants to meet my other guest, or if I think he cannot know my other guest is with me, I shall be inclined, in spite of the 'significance' of the first criterion, to suspect that Mr X's visit is chance after all. But if I believe he strongly desires to see my other guest, and have reason to suspect that information about this guest's presence has leaked, then I might suspect Mr X's visit to be causally arranged even though he normally came to see me quite often.

This example throws an interesting light upon our usage of the concept of chance. Our enquiry about Mr X's visit begins with the assumption that he is a sort of chance-machine. We go on thinking in these terms until he does something very improbable. Then we assume he is not a chance-machine after all. D'Alembert's axiom that very small probabilities are really zero I take to be an illustration of this usage. Theoretically there is a chance of 2^{-100} that an unbiased coin will fall heads on every one of a hundred tosses. But if this event were actually observed, we should not call the coin unbiased. It is inconsistent to say a particular event has a certain positive probability if we are going to disallow the event whenever it occurs. Indeed, if this is what we are going to do, then our ascription of a zero probability to the event is quite accurate: and to say it has a probability of 2^{-100} is wrong.

Thus a chance-machine is allowed by the properties we ascribe to it to give results which fall only within a certain range. If they fall outside this range, we at once cease to call it a chance-machine. The minimum probability observable

depends on how far we can count in an experimental lifetime. All probabilities less than this are the same, that is, zero.

The behaviour of Mr X throws light also upon our concept of causality. While we are thinking of him as a chance-machine we do not thereby assume that his rare visits are without causation. We assume he has his own reasons for coming, though we may not be interested in them. We assume, indeed, that all his visits are in some respect or by some reference causally arranged. We nevertheless continue to call them chance visits as long as their causal arrangement appears to be without reference to events which are of any interest to *us*. It is only when Mr X's arrangements appear to become predictable by reference to our own arrangements that we cease to think of Mr X as a chance-machine. The concept of chance aligns with that of independence and shares its relativity.

The concept of causality arises as a description of dependence in sentient or conscious beings. I behave in particular ways, and notice the *effects* of my behaviour on people and animals. I notice also its less reverberating effects upon inanimate objects. It is a sophisticated step to extend the concept to wholly non-living worlds, but it can usefully be taken. I might make a series of observations to discover whether the height of the cloud tops was in any way related to the height of the mercury in the barometer and discover that it was. I could then speak of a causal arrangement involving no living organisms.

The use of statistical investigation upon entirely non-living material is rare. This is partly due to a bias in our interests. Sir Ronald Fisher's famous tea-cup experiment[1] in which he

[1] Fisher, Ronald A., *The Design of Experiments* (London, 1947).

lays the foundations of modern statistical investigation is devised as an attempt to discover whether there is causal dependence or only chance relationship between the method of preparation of cups of tea and a person's responses to them. In outline it is as follows.

A lady claims that she can tell by tasting whether the milk or the tea infusion was poured into the cup first. She is presented with eight cups of tea, four of one kind and four of the other. She is told that four cups have had the milk put in first, and asked to pick them out by tasting. We now assume that the lady can't tell the difference; we imagine that she turns the experimental set-up into a chance-machine. Inspection of the problem will show that there are $\binom{8}{4} = 70$ possible ways of choosing four cups out of eight. One only of these will be the right way so we can define the lady's chance of getting all four cups right in a single trial as $\frac{1}{70}$. Thus if, in these circumstances, she did get all four right, we might consider the result remarkable enough to lead us to suspect that the experimental set-up was not a chance-machine after all.

But there still remains one problem. So far we have taken no account of the order in which the various cups of tea were presented. It is well known that, left to themselves, people tend to guess according to certain patterns. We will suppose for simplicity's sake that the lady will tend, if she can't tell the difference, to place the cups in the order: milk, milk, milk, milk, tea, tea, tea, tea. We will suppose further that the experimenter himself has a natural tendency to arrange the cups of tea in this order. The result of this set-up would be that the lady would tend to guess right whether or not she could tell the difference by tasting. The object of the

experiment would not be achieved, and the test of the significance would be vitiated.

The fashionable method for resolving this difficulty is to randomize the order in which the cups of tea are presented. Indeed, some form of randomization is necessary before any test of significance may be validly introduced. Thus, in Fisher's words, 'the two modifications of the test beverage are to be prepared "in random order". This, in fact, is the only point in the experimental procedure in which the laws of chance, which are to be in exclusive control of our frequency distribution, have been explicitly introduced. The phrase "random order" itself, however, must be regarded as an incomplete instruction, standing as a kind of shorthand symbol for the full procedure of randomization, by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated.' As at the end of the last chapter, we are here again left hanging upon the concept of randomization; we shall devote the next few chapters to its elucidation.

VIII

THE RANDOM SERIES

I

LET us suppose we have an old car, and that on one occasion when we are unable to start the engine Mr X is in the car. If our attention were drawn to this conjunction of events we should probably say that they 'had nothing to do with one another', that they just 'chanced' to be coincidental and that their correlation was 'not significant'.

Now let us suppose another case. Each of two cricket captains wishes to put in his team to bat first. As this is impossible, the problem is to find a decision which will be acceptable to both. Here if either captain were allowed to decide whose team was to bat first, the other might regard the decision as unsatisfactory and the rule by which it was made as unfair. So the decision is made by neither captain, but by 'chance'; a coin is spun and allowed to decide for them. Here the decision might be equally unsatisfactory to one of the captains, but he accepts it as fair. Later we shall see why. For the moment we shall return to the case of the car.

Suppose that in the last six occasions when we tried to start the car, it started well enough on all but three of these occasions and suppose, moreover, that these three occasions were the only occasions when Mr X was in the car. We should now become suspicious. What we should begin to

suspect would be some sort of 'real' or 'causal' relationship between the car's not starting and Mr X's being in it. We might say, 'This doesn't look like chance!' Of course we should admit that it might still very well be chance, but we should have begun to feel that it might also very well not be.

What, then, do we mean when we say that we think the car's behaviour is 'chance' (or, as we sometimes loosely say, 'due to chance')? We mean that we expect its relationship with some other event along with which we have observed it will not continue if we observe other instances. Similarly, if we say that its behaviour is 'not chance', we mean that we expect the association to be continued. If we examine a series of falls of a die and find a high proportion of sixes, we tend to say, 'This is not chance'. Here the logical procedure has been to divide up the series into groups. Each of these groups has been observed, and it is noted that all of them have more sixes than anything else. After a while we expect and are able to predict this property of the next group we examine. The two conjoined events here are groups of falls of a particular die and preponderance of sixes.

Now, in the case of our car, there would come a time when we should begin to *suspect* some correlation of a real nature between Mr X and its not starting. And by 'correlation of a real nature' we mean no more than that it is an observation we shall be able to make again. In order, in fact, to verify our suspicions, the only thing we can do is to give the observation an opportunity to repeat itself. So, if we try to start the car again with Mr X inside and it still does not start, but starts perfectly well without him, we shall say our suspicions were *justified*, and this curious association of data is not chance after

all. The classification, be it noted, applies not only to our latest (confirmatory) observations of the car and Mr X; it is also retroactive, in that all our previous associations of Mr X with the car's not starting, although previously called chance associations, are now reclassified as not chance. The unreal becomes real, and Mr X was guilty all the time.

But let us suppose that further observations, instead of justifying our suspicions about Mr X, show no correlation at all between his presence and our inability to start the car. How do we behave? After a few observations we begin to suspect that we might, after all, have been wrong about Mr X; and later, when more observations have 'reduced' the original data to 'insignificance', we assume that we were in fact wrong, and reclassify the observations of the past accordingly. The real becomes unreal, and we apologize to Mr X.

Let us now suppose a rather different case. Out of the last 111 attempts to start the car, all have been successful except those on the 73 occasions when Mr X was in it. But in making what now seems the quite justified prediction that the car will not start again with Mr X in it, there appears in future no relationship at all between Mr X's presence and the starting of the car, thus confounding our prediction. Here no amount of later evidence will incline us to say that we were wrong about Mr X's past relationship with the car. Instead we are inclined to say that the new observations, which do not support the original judgment, are merely evidence for Mr X's *changed character*. We say, in other words, that the new set of observations come from a *different population*. It is only when the habit of expectancy built up is not very strong that we make, as in the previous case, the sort of reclassification of the data which entirely absolves Mr X.

We will examine these two kinds of classification in greater detail.

The sets of observations which are to be classified are distinguishable only in the following way. The first set leads us to make a prediction about what will happen in further sets of similar observations. If this prediction is not substantiated in subsequent sets, the correlation between the events cited may be reclassified from 'real' to 'chance'. But if the bias observed in the first set is very pronounced, we do not reclassify it as 'chance' if subsequent sets fail to confirm it. Instead we say that the new observations are of a different kind and therefore not relevant. In the former case the observations of the original set are reclassified into the class of the observations of the subsequent sets; and in the latter case the observations of the original set are not reclassified, so that the subsequent sets are said to be of different observations. These rules of classification are important for the concept of a random series.

If a series is to remain homogeneous, then the concept of its 'randomness' is analysable in terms of the unexpectedness or unpredictability of each of its members in turn. But if the series is broken up, on the nature of its data alone, into, say, two series, one 'random' and the other not, then the usual meaning of 'random' in this case is 'possessing no discernible pattern'. The former concept applies to unexpected individual observations, and is usually covered by the primitive meaning of the word 'chance'. The latter is usually included in the meaning of the word 'random' in the expression 'random series'. It is sometimes assumed that the two concepts are equivalent, or at least that the latter implies the former. Both of these views, as I shall argue in the next

section, are wrong. At best they provide a working approximation only for certain series which are neither very short nor very long.

Let us consider again the two cricket captains spinning a coin. It has been agreed between them that no form of decision in which the result is predictable by either of them is satisfactory. Of the alternative means of deciding they might have a fight, run a race, or play ping-pong: but it is simpler and quicker to spin a coin. Less acceptable would be for one of them to cast a die and say 'If this die turns up six your team bats, if not mine does', since it is our general experience that six does not often turn up on a die. We know that it does sometimes, but if we had to make a prediction we should normally predict that it wouldn't. Thus the criterion of satisfactory decision is absent. But if both captains were ignorant of the relevant behaviour of dice, then the decision, as far as *they* were concerned, would be quite satisfactory, though it might appear unfair to onlookers who knew about dice.

Let us now imagine a case where one of the captains has won the toss fifteen times running. We should begin to suspect that he was cheating in that he *could* predict beforehand the outcome of the toss. If we discovered also that he was using a double-headed coin, this would add to our suspicions. But if it could be proved that he certainly did not know that the coin was faulty, then we could not properly accuse him of cheating; the rule that the outcome of the toss should be unpredictable by either captain would not have been broken, and the results of the fifteen spins would have been chance results, that is, they would each have been unpredicted. That the faulty coin happened to favour a particular team would still have been a matter of luck.

The fact that cheating depends not on the actual machinery used but on the ability of the cheater to predict what it will do is illustrated in the imaginary case of a person who could, by lightning observation and calculation, predict and call the fall of a coin while it was still spinning in the air. He would be just as guilty of cheating as a person who made his predictions by means of a double-headed coin. But there is an important social difference between the two cases. Double-headed coins, being more common than lightning calculators, are specifically guarded against in the rules of fair play, and our censure of them is therefore stronger because of precedent. Also, there would be a tendency to be lenient to the lightning calculator through admiration, where we should not admire the person with a double-headed coin.

The distinction between the two meanings of 'chance' is now clear. The results of each of the fifteen spins of the coin before we began to suspect jiggery-pokery were 'chance' in that they were not expected (predicted) by the observer or observers. Had there come a time, say, as here, after fifteen heads running, when the observer had begun to predict what came next, then the events beginning from where he had predicted them would not on any view be called 'chance'. The events before this point might *have been* called chance, though convention has it that even they should be relabelled in the light of later events. But anyway, as individual events, they were 'chance' in the sense of not generally expected when they happened. What we have done when we have reclassified them as 'not chance' is to use the word 'chance' in a different sense; it is the series itself we are now speaking about, and it is certainly not chance as a series, since it shows an obvious pattern, namely, all its members being the same.

But the argument from the premiss, 'The series is not chance' to the conclusion, 'Its separate events were not chance' is not good because it is a logical pun on the word 'chance' which has a different meaning in each proposition.

II

In the previous section we distinguished two senses of 'chance', the first applicable to discrete events and the second, often called 'randomness', applicable only to series of events. For the sake of clarity we shall speak of the former as 'primary chance' or 'primary randomness', and the latter as 'secondary chance' or 'secondary randomness'.

An event is primarily random, then, in so far as within the framework of possibilities we are considering, we cannot be sure either of its occurrence or of its non-occurrence. The only relevant criterion of primary randomness is that we are able to guess. Note that the concept can only apply to an event which is yet to happen. It is strictly meaningless to use the expression for past events *which we now know*, since these events cannot now be guessed; all we can say is that if they *were* guessable, then they *were* primarily random. Note also that the expression 'primary randomness' is applicable only to *classes* of events in so far as it is applicable to each member of the class as an individual. Note also, finally, that the concept bears resemblance to ethical concepts such as 'good' in that it admits of analysis in subjective terms. For example, since an event may be expected (and therefore unguessable) by one person and unexpected but guessable by another, there is no contradiction when the latter says 'It is random' and the former says 'No, it is not random'.

Secondary randomness is a more objective concept, though difficulties arise if we try to press the idea of its objectivity too far. It is a property belonging only to series of events or observations when the series are themselves taken as units; and in this sense a random series is a series with no discernible pattern. Note that the objectivity of this definition is not as firm as we might like it to be. If by 'colour' we meant 'spectrum colour' we should not define a grey object as 'having no discernible colour'; simply 'having no colour' would be sufficient. Nor does it help us if we define this type of randomness—which Professor von Wright[1] has called 'mathematical randomness'—as 'insensitivity to selections according to mathematical rules': since now whether a series is random or not depends entirely on what rules we adopt for testing it.

It seems that the practice of not distinguishing between these two meanings of the word 'random' has led to a number of puzzles. I shall try to show below that the two concepts, though sometimes roughly applicable to the same set of data—the one to the individual items as such and the other to the set as such—are neither of them inferable from the other, and that they are, moreover, ultimately incompatible.

Suppose we have control of a series of noughts and ones which are presented to a subject who is telling us his expectations, if any, of what the series will do next; and suppose that we arrange the ones and noughts after each guess so that only about half of the subject's guesses are right. Each event in this series is now by definition primarily random to this subject, but obviously by this means we could produce a series which was not secondarily random even to the subject to whom each event in it was primarily random.

[1] von Wright, G. H., *Mind*, **49**, 279 (1940).

Let us now take another case. Suppose we are given a series of 1,000 ones and noughts together with the information that it is secondarily random. Suppose also that the test for secondary randomness is here rigorously defined in terms of frequency: 40 noughts more or less than 500 being sufficient to exclude the series from the class of series which are secondarily random. After observing the first 900 of these digits we find in them 500 noughts; but we know that the whole series is secondarily random; therefore, it will not be rejectable on the test for secondary randomness; therefore we can predict that the last 100 observations in the series will contain at least 22 more ones than noughts. But this means that we can predict ones with some success; therefore these observations are not completely primarily random with respect to ones and noughts.

Thus, though in some cases an observer may be able both to call the several events in a series primarily random and to call the series itself secondarily random, there are certain cases where, by virtue of the fact that he is enabled to call the separate events of a series primarily random, he becomes unable to call the series as a whole secondarily random, and *vice versa.*

Attempts to form a more precise single concept of randomness generally leave the empirical problem untouched. For example, Professor von Wright [1] gives the following definition of 'absolute randomness'.

> The distribution of A in H is called random, if there is no property G such that the relative frequency of A in H & G is different from the relative frequency of A in H.

[1] von Wright, G. H., *A Treatise on Induction and Probability* (London, 1951), p. 229.

Obviously, in practice, a difference of a small amount between the relative frequencies of A in H & G and A in H does not prevent our saying that the distribution of A in H is random. But this is a matter of practical rather than logical convenience, since we know that if we make a rigid criterion of difference here we put ourselves in the same logical difficulties as before: whatever criterion of 'significance' we take—that is, whatever arbitrary difference between the relative frequencies of A in H & G and A in H we use to prevent our asserting that the distribution of A in H is random—we know that if we take enough of such series and the criterion remains *unbroken* in all of them, then there comes a time when we may *not* call such series random. In other words we suppose that if we take a sufficient number of 'random' series, we can find among them any given permutation of their elements. The expression of this theorem is well known in terms of a monkey with a typewriter. It strikes the keys at random and eventually, in the long run, writes all the Shakespeare sonnets, or something equally unlikely. I shall refer to this as the monkey theorem. It has a converse, which is equally applicable in this case.

In the definition cited above, randomness of A in H is made dependent on there being *no property* G such that the relative frequency of A in H & G is different from that of A in H. To this there is the clear objection that such a property G, if we look far enough, can always be found. Lord Russell[1] gives the following illustration.

> Take, for example, the numbers of all the taxis that I have hired in the course of my life, and the times when I have hired them. We

[1] Russell, Bertrand, *Human Knowledge, Its Scope and Limits* (London, 1948), p. 329.

have here a finite set of integers and a finite number of corresponding times. If n is the number of the taxi that I hired at the time t, it is certainly possible, in an infinite number of ways, to find a function f such that the formula

$$n = f(t)$$

is true for all values of n and t that have hitherto occurred.

This is the converse of the monkey theorem. The only practical meaning that can be given to the von Wright definition of randomness is not, therefore, that there is no property G, etc., but simply that we do not expect to find such a property by selection *at random*. We are thus reduced to interpreting randomness in terms of randomness, which means, as J. M. Robertson[1] long ago pointed out, that we have succeeded only in 'measuring chance in terms of chance'.

An attempt to skirt this obvious difficulty was made by von Mises [2] who, to avoid application of the monkey theorem converse, defined all random series, which are instances of what he called 'collectives', as being infinitely long. A collective he defined as a series in which the ratio of the elements present tends to a certain limit as the series continues, and from which any place-selection generates a new series with the same elements tending to the same limiting ratio. But there are serious objections to this view. Apart from the practical difficulties of testing an infinitely long series for randomness, my colleague Mr Kneale [3] and others have pointed out that both the concepts of limit and of infinity have been used hitherto only for series for which we have formal rules of construction; and since it is agreed that a

[1] Robertson, J. M., *Letters on Reasoning* (London, 1902), p. 108.
[2] von Mises, Richard, *Probability, Statistics and Truth* (London, 1939).
[3] Kneale, William, *Probability and Induction* (Oxford, 1949).

random series may not have formal rules of construction, it is doubtful whether either the concept of limit or that of infinity can be given any meaning in this definition. Dr Fry also has criticized the use of a limit concept and shown [1] that in a sequence of independent events such a limit can always be exceeded: that is, there is no limit.

In a random series, not only the atomic events are supposed to be primarily random; all the molecular events, such as groups of ten atomic events considered as a unit, are supposed to be primarily random as well. As a consequence, these molecular events cannot all be secondarily random (monkey theorem). If we assume that the molecular units (such as all the patterns of ten atomic units) will behave, within the framework of possibilities, just as unpredictably as the atomic units themselves, we find we have a method which enables us to calculate the relative frequency of a given pattern type (for example the type consisting of groups of ten adjacent units with more than seven units the same) in a given series or segment. This is the basis of Bernoulli's law of large numbers; and it is only to the extent that we can make the above assumption of *some* natural series that present theory can lead to any reliable predictions. But let us observe the consequences of this assumption as the series lengthens.

Suppose that in a random series of digits we take units of five digits as our events. Since the series is random, we have no reason to expect any particular group such as '12345' any more than any other such as '32213' or '00000'. There are, as we can easily calculate, exactly 100,000 different permutations of five standard digits, allowing for repetitions. So in

[1] Fry, Thornton C., *Probability and its Engineering Uses* (New York, 1928), pp. 88–91.

a million such groups of digits we should expect to find roughly ten groups of each—that is, about ten 00000's, about ten 00001's, ten 00002's and so on. Why roughly *ten*? Because there are 100,000 possibilities to be divided between 1,000,000 instances, so that if, say, 00000 consistently appeared about thirty times in every million such groups, we might begin to expect it in preference to other groups; and if these expectations turned out to be justified, then the group 00000 would not be the primarily random event we are supposing it to be. Why *roughly* ten? Because if we knew it to be exactly ten and we had examined, for example, the first 99,990 of our five-digit units and found no groups of five noughts, then we should know that the last ten groups were all groups of five noughts; the final ten observations would thus have no element of primary randomness whatever.

Let us suppose a very long series divided up into groups of a million digits each. As there are $10^{1000000}$ different groups of this kind, we shall expect to find, if they are primarily random, about ten of any particular kind in a series of $10^{1000001}$ of them. Thus in a random series of $10^{1000007}$ digits we should expect to find about ten subseries of a million consecutive noughts.

Now let us consider an observer with a machine for making random numbers, having arrived at the beginning of one of these subseries of a million consecutive noughts. Will he be calling the series random? If he is accustomed to checking long series of about $10^{1000007}$ digits, he might. But if he is a normal observer, dying at about 70, he will be mildly surprised after five consecutive noughts; after ten he will begin to suspect the machinery; after twenty he will call for his laboratory assistants to see to it; and, if he happens to be

compiling a table of random numbers for scientific uses, he will certainly regard the records from where the noughts began as unpublishable.

This predicament can be approached another way. Suppose we are organisms with a very slow rate of observation, such that we rarely count beyond five of any event in a lifetime. Alternate ones and noughts might appear random to us, but a series of, say, two noughts might seem to be quite significantly long.

From whichever end we look at it the dilemma is the same. In order to give a practical interpretation of probability theory for scientific purposes, we have to assume the primary randomness of molecular events; but the moment we do this our random series contains limitless possibilities of predictable repetition which we cannot call 'random' in any ordinary sense of the word; when, therefore, one of these possibilities in our random series begins to be realized, we do everything we can to stop it by fiddling with the machinery, and to hush it up by suppressing its publication as such. Sir Ronald Fisher and Dr Yates, having produced for publication some random numbers[1] which failed to pass the test, altered them until they did. Professor Kendall and Mr Babington Smith[2] took the other course and suppressed 10,000 of theirs.

It becomes obvious, then, that the concept of randomness, instead of growing more satisfactory in the consideration of longer series, tends instead to grow less so; and that in a series of infinite length it becomes absolutely contradictory: that is to say, in an infinite series the impossible will certainly happen.

[1] Fisher, Ronald A. & Yates, F., *Statistical Tables for Biological, Agricultural and Medical Research* (London, 1949).

[2] Kendall, M. G. & Babington Smith, B., *Tables of Random Sampling Numbers* (Tracts for Computers XXIV) (Cambridge, 1939).

IX

THE PARADOXES OF PROBABILITY

In the last chapter we showed up the ultimate self-contradiction of the concept of randomness. This can be further exemplified as follows. We have a randomizing machine which produces a series of ones and noughts. We require for experimental purposes a random series of 16 ones and noughts. We start the machine which now gives us a series of 16 noughts. We of course reject this series as unsuitable and suspect the machine of being biased. It is returned to the makers for adjustment. When it comes back we have a very long experiment for which we require a random series of 2,000,000 ones and noughts. We leave the machine running all night, but on checking through the 2,000,000 ones and noughts it produces we are surprised to find not a single run of 16 noughts. Again we suspect it of being biased and send it back.

But what is its designer to say to all this? First we send it back because it produces 16 noughts in a row. Very well: he puts in a device to prevent its doing this. We then send it back because it never produces 16 noughts in a row. What is he to do now? First of all we use a specific criterion to reject the series the machine produces, and then we use the absence of this very criterion to reject another series it produces. It seems we are never satisfied.

But if the designer were careful enough to note exactly when we became dissatisfied with the machine's performance, then he might hit upon an ingenious idea. He would see that we didn't like 16 noughts appearing all together in a very short series, but he would see also that we appeared to want 16 noughts together if the series were very long. He might therefore fix a gadget to his machine which inhibited long runs of ones or noughts in short series but encouraged them in longer series. It would be a sort of meter device, and the size of run allowed would depend upon how long the machine had been working.

But a randomizer is just a machine which, in this case, is as likely to produce a one as a nought. Moreover, however many noughts or ones have gone before, it is still supposed to be just as likely next time to produce a one as it is to produce a nought. Similarly for the pairs 00, 01, 10 and 11; the machine is supposed to be just as likely to produce any one of these as it is to produce any of the others. Again for trios, quartettes, and so on.

We measure likelihood by what has already happened; when we see 16 noughts together and nothing else, we cannot say, on this evidence, that noughts and ones are equally likely. Noughts are clearly much more likely. But when we consider very long series, we have room for many repetitions of quite long patterns such as 16 noughts in a row. Now, as there are only 65,536 different patterns of 16 ones and noughts in a row, a series of about ten million random ones and noughts should contain about ten of each pattern if the original condition of equal likelihood is to be maintained.

Suppose, then, we allow our criteria for randomness to vary according to the length of the series we see. In other words

we are driven willy-nilly to accept the gadget which the designer has at last hopefully put on the machine, which prevents 16 noughts in a row ever occurring in, say, a series as short as 30. 15 noughts running, then, is all that this machine is going to allow in a series only 30 units long. For our next experiment we need, say 30 random digits. We start it up and it gives us a couple of ones to begin with, after which it goes on to produce 15 noughts. It is a good randomizer and must therefore keep within its limits. Its next figure therefore cannot be a nought and must be a one. It is not, therefore, equally likely to be a nought. The criterion of independence of atomic events is absent and the machine cannot therefore be a good randomizer.

It is probably this awkward contradiction that makes the average writer on probability expose to such savage ridicule the punter who believes in the maturity of chances. I quote from *The Science of Chance* by Mr. Horace C. Levinson.

> Suppose that you play at heads or tails with a certain Mr Smith, the amount bet at each toss of the coin being fixed at $1, and that Mr Smith has allowed you to decide, before each toss, whether you will bet on heads or on tails. At a certain point in the game let us assume that heads has turned up ten consecutive times; how will you place your next bet, on heads or on tails? The doctrine of 'the maturity of the chances' insistently advises you to bet on tails, 'for', it says, 'heads has come more than its share, at least in the last few tosses, and therefore there is more chance for tails, which will help to restore the balance.'

Mr Levinson produces a number of arguments to demonstrate the absurdity of this doctrine. The most convincing one—that the theory of chances assumes that each event in a

random series is independent of the others and that one may not argue from this premiss of independence to a conclusion of dependence—he misses, and falls back finally on an appeal to experiment. It is perhaps unfortunate that the only experiment he cites is one where he has to admit that 'the laws of chance were *not* followed'. He goes on to say:

> It is a curious, but not necessarily an astonishing fact that gamblers have a second doctrine that precisely contradicts that of 'the maturity of the chances.' According to this second principle you are advised to place your bets on the chances that have appeared *most* frequently. It would seem appropriate to call this maxim 'the immaturity of the chances.' From what I am able to make out of the accounts of this doctrine it appears that it should be followed when the maturity doctrine does not work. The decision as to when one of these theories ceases to work, and the other becomes valid, would seem rather difficult in practice.

Mr Levinson's arguments against the maturity of chances hypothesis, in keeping with those of other theorists, all skirt the main problem. Granted that there is no need after, say, an excess of heads, for a subsequent excess of tails to 'restore the balance'; as Mr Levinson points out, if the original excess remains constant it will have less and less effect on the ratio of heads to tosses as the length of the series increases. But this is not altogether relevant. It takes no account of the need for every randomizer to keep the series it produces within the bounds of the tests for randomness we are going to make on them. When the excess of a particular element has reached a large enough value, we then have to decide whether the machine is a good randomizer or biased.

Suppose in roulette we are betting on red or black. Suppose also, like all good statisticians, we determine upon a criterion

of significance for bias of the roulette. We will take a simplified criterion in terms of runs of reds or blacks to illustrate the point. The principle will at once be seen to apply to other criteria.

Suppose our betting is limited to a day, and we decide that a run of 20 reds or 20 blacks is decidedly 'significant' and that if one occurs we shall agree to stigmatize the roulette as a bad randomizer. During the course of our betting a run of 19 reds appears. We are required to place our next bet. Shall we place it on red or on black? Our decision depends now upon our acceptance or rejection of the machine as a randomizer. And this decision has now become expressible in our next bet. Suppose we decide to bet upon another red. This means that we believe the 20th red of the run will turn up and therefore that the machine is biased. So, if we believe that the machine is still a good randomizer, we can express this belief only by betting on black. Our belief therefore demands at some stage a maturity of chances betting policy.

We bet, therefore, on black; but yet another red turns up. Our maturity of chances policy (consistent with our belief in the machine as a randomizer) has failed, and we have only one course left, which is to assume that the machine is biased in favour of reds and to place our future bets on that colour. This, at any rate, is no more than a policy of common sense, and it is difficult to see why Mr Levinson should refer to it in terms of such archness and veiled hostility.

It is also difficult to follow Mr Levinson when he says that the gambler should base his policy upon clear-cut experiment. Surely there is nothing so clear-cut in probability as the evidence that when a machine has produced a continuous excess of one element in a series, it tends to go on doing so.

Philosophers are familiar with the paradoxical announcement. A notice is put up saying that there will be a fire practice at a particular time one day next week, but in order that the practice should come as a surprise the day on which it occurs will be kept secret. If the week begins on Sunday, the practice can take place at the specified time on any day up to Saturday. But if it hasn't occurred on Friday, it will be certain, in the terms of the notice, to occur on Saturday; and therefore, against the terms of the notice, it will not be a surprise. It therefore cannot occur on Saturday, and the last possible day for its occurrence must be Friday. If, therefore, the practice has not occurred by the specified time on Thursday, it must, then, be going to occur on Friday. But if this is so it cannot be a surprise; therefore Thursday is the last possible day on which the fire practice can occur. If it hasn't happened by Wednesday—but we needn't go on. Each day of the week is relentlessly excluded until there is no possible day on which the fire practice can occur so as to satisfy the terms of the notice. But this is absurd: for clearly the practice can happen on any day of the week and yet be a surprise. The paradox is ingenious, but the frequency theory of probability can match it.

We will imagine a randomizer producing a series of ones and noughts 100 units long. We now choose a criterion of significance which rejects all series in which the excess of either kind of unit exceeds 25. The machine is thus prevented from producing series which are outrageously biased. But, being a randomizer, it must in addition obey the rule of independence, i.e. we mustn't be able, from what has gone before, to tell what will come next. We set the machine going and find that at its 90th digit in the series of 100 it has already produced 62 noughts. Another nought would make

the series rejectable and therefore cannot occur. The next 10 digits must all, therefore, be ones. But this won't do at all, since we should now be able to bet on them with certainty. In this set-up, then, the 62nd nought can never be allowed to occur before the ultimate digit. Suppose, then, we have reached our 90th digit and find so far a count of only 61 noughts. We now know that all the remaining digits up to at least the penultimate digit must be ones. Again we can bet with certainty, and this is not allowed. Let us try, therefore, the supposition that by our 90th digit we have counted only 60 noughts. Now, if any of the next 6 digits is a nought, we are left in the same difficulty as before. Therefore, all the next 6 digits must be ones.

This argument, like the argument from the paradoxical announcement, can be extended back to a similar absurd conclusion. But though the conclusion is absurd, it is enough to throw at least doubt upon both frequency and limit concepts of probability. Clearly, at any rate, there can be no limit to the possible variation of a probability series. But this brings in its train further difficulties. It means, for example, that one of the conditions of the probability of event E being $\frac{1}{2}$ is that in a test to verify the statement of its probability we must sometimes get a ratio suggesting, for example, that its probability is $\frac{1}{20}$. The ratio mustn't amount to $\frac{1}{20}$ every time, but the important point is that if we *never* get a result suggesting $p = \frac{1}{20}$, then we cannot say that $p = \frac{1}{2}$. Similarly, if the result of our verification of the proposition that $p = \frac{1}{2}$ cannot give us a ratio suggesting $p = \frac{1}{1{,}000{,}000}$ or $\frac{999}{1{,}000}$ or anything else, then again we cannot say that $p = \frac{1}{2}$. Thus in discovering criteria for the truth of the proposition $p = \frac{1}{2}$ we find that in

most tests the ratio $E : E \,\&\, \bar{E}$ must be something other than $\frac{1}{2}$. But of course it is also necessary at some stage for the ratio of $E : E \,\&\, \bar{E}$ to be $\frac{1}{2}$.

This paradox has its converse. One of the criteria for the truth of the proposition that the probability of event E is not $\frac{1}{2}$ is that sometimes the ratio $E : E \,\&\, \bar{E}$ must be $\frac{1}{2}$.

If event E has probability p, then at some stage the ratio $\frac{E}{E \,\&\, \bar{E}} = p$ must hold, and at some other stage $\frac{E}{E \,\&\, \bar{E}} \neq p$ must also hold. But precisely similar conditions must hold for the verification of the proposition that the event E has the probability not-p. Nor is it any use trying to get over this paradox by saying, for example in the first case, that although these criteria are necessary, the former ratio must hold more often than the latter; for then we have a right to ask *how much more often*, and are served with the paradox as before at a different level. For if we answer the question 'How much more often?' we are at once ascribing a limit to the probability set-up which, as we have shown, thereby ceases to be a probability set-up.

We have seen that a tautology is a statement which is true but not helpful, and that a contradiction is a statement which is both useless and false. Neither form gives information, each being, in the strict empirical sense, meaningless. We have now shown that probability statements are meaningless in the contradictory sense. But there is worse to come.

Let us examine two important statements taken from Sir Ronald Fisher's *The Design of Experiments*. On page 13 we read that

> if we . . . agree that an event which would occur by chance only once in seventy trials is decidedly 'significant', in the statistical

> sense, we thereby admit that no isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon; for the 'one chance in a million' will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to *us*. In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result.

Sir Ronald Fisher goes on to say that the sensitivity of any one experiment can be increased by making it larger. This, he says, can be achieved by increasing the length of the individual experiment or by doing a number of similar experiments. Of the latter he says: 'This procedure may be regarded as merely a second way of enlarging the experiment and, thereby, increasing its sensitiveness, since in our final calculation we take account of the aggregate of the entire series of results, whether successful or unsuccessful.' We have now gathered together all the material for a new paradox, which we can formulate as follows.

First we make a syllogism.

Major premiss: No isolated experiment, however significant in itself, can suffice for the experimental demonstration of any natural phenomenon.

Minor premiss: Any series of experiments may be regarded, for the purposes of statistical argument, as a single or isolated experiment whose length is the aggregate of the separate experiments comprising it.

Conclusion: Therefore, no series of experiments, however

significant in themselves, can suffice for the experimental demonstration of any natural phenomenon.

This is a valid syllogism. Thus from the two premisses laid down by statistical procedure we arrive at a conclusion which flatly contradicts the basic assumption of scientific method. The purpose of series of experiments is to gain information about natural phenomena; it is therefore paradoxical to draw up rules of procedure which assert that no such information can be so gained.

We have found so far that the concept of probability used in statistical science is meaningless in its own terms; but we have found also that, however meaningful it might have been, its meaningfulness would nevertheless have remained fruitless because of the impossibility of gaining information from experimental results, however significant. This final paradox, in some ways the most beautiful, I shall call the Experimental Paradox.

X

CRITICAL SERIES

PROLEGOMENA

Let us consider a series of L decimal digits. These are to be used for biological or other experiments involving significance tests. Therefore, before the significance tests can be considered valid, the series must pass tests for randomness.

A test for randomness consists in the application of a criterion for the acceptance or rejection of a series as an experimental standard. Suppose we take the class of all possible series containing L decimal digits. The total number of such series is 10^L. A single test for randomness at the 1% and 99% levels would reject exactly 2% of these series. Two independent tests at the 5% and 95% levels would reject rather less than 20%, and so on. In other words, the greater the number of tests and the more stringent they are, the greater the proportion of series which must be rejected.

Exactly which series will be rejected and which accepted depends entirely upon how we choose the tests. Most tests for randomness are not specific, but reject large groups of series at once. There is no reason why we should not increase the precision of the tests so that each test rejects one and one only of the possible series, accepting all the others. In this case there would be as many tests for randomness as there were series which could be tested.

Let us suppose now that we are about to conduct a very important experiment involving a significance test, and we wish to make sure that the standard series we use is thoroughly random. There are now 10^L tests for randomness, and we decide to apply all of them except one. We go on producing series until we find one which is not rejected by this very stringent set of tests, and then we use it. This is equivalent to choosing whatever series we please. But textbooks on statistics warn us against the dangers of this procedure, which are obvious. Nevertheless, as we have now shown, any test for randomness to some extent applies it. Suppose, therefore, we decide to use only series which have not been tested for randomness.

The trouble with untested series is that they tend to be biased, and the dangers of using a biased series may be even worse than the dangers of making up the series oneself. Some compromise must be reached. But it is now clear that whatever we do we shall not in the end be able to apply the ordinary probability calculations with any confidence, since they are based on the assumption that we have done nothing. It is true that such calculations have been applied in the past and have often achieved their purpose by indicating where repeatable results might be found. But we have no evidence that the answers they have given are *accurate*; we have, on the contrary, considerable evidence, both practical and theoretical, that they are not. It is our task therefore to examine (*a*) how inaccurate they are, and (*b*) what, if anything, we can do to improve them.

Definitions

A *series* is a one-dimensional arrangement of events, units, or *terms*. The number of terms $a_1, a_2, a_3, \ldots, a_L$ comprising it is

called its *length* L. The number of different valued terms or *elements* $v_1, v_2 \ldots v_w$ possible within it is called its *scope* w. The series contains u_1 terms of value v_1, u_2 terms of value v_2, etc., where $0 \leq u_j \leq L$.

$$\sum_{j=1}^{w} u_j = L.$$

The length of a series divided by its scope is called its *absolute expectation* or *expectation e*.

$$e = \frac{L}{w}.$$

The amount by which a given element exceeds absolute expectation is called the *bias d* of the series with respect to that element, and may be negative if the number falls short of expectation.

$d_j = u_j - e$ where v_j is the element with respect to which the bias is assessed.

$$\sum_{j=1}^{w} d_j = 0.$$

Deviation $|d|$ is used to designate the magnitude of a bias independently of its direction.

$$|d_j| = |u_j - e|.$$

Phrases such as 'more (or less) biased than . . .' usually refer only to deviation. A further expression for bias will later be used to facilitate the algebraic expression of matching operations. It is called *reciprocal bias r*, and is the total number of

terms contained in the elements other than the one with respect to which the bias is assessed.

$$r_j = L - u_j = L - e - d_j.$$

$$\textit{Critical bias } \sigma = \sqrt{\left(\frac{L(w-1)}{w^2}\right)}.$$

A *critical series* is a series critically biased for a given element.

Series of scope $w = 2$ are called *binary*, of scope $w = 3$ are called *ternary*, etc. Series in which $|d_1| = |d_2| = \ldots = |d_w|$ are called *symmetrical*. All unbiased series and all binary series are symmetrical. Biased series whose scope is odd cannot be.

Random and Critical Series

Newton's Theorem states that *the probability* P(p, m, n) *of the occurrence of exactly* n *events of a given type on* m *occasions when the probability of such an event occurring on any one occasion is* p *is given by*

$$P(p, m, n) = p^n(1-p)^{m-n}\binom{m}{n} \text{ where } 0 \leq p \leq 1.$$

The ordinary interpretation of Newton's theorem is circular: we cannot discover $P(p, m, n)$ without knowing p; but our estimate of p is ultimately determined by the value taken by $P(p, m, n)$. If we interpret the theorem in terms of relative frequencies, we can avoid the circularity as follows.

In the set of all possible series of length m *and scope* w, *the proportion* P(w, m, n) *of series in the subset containing exactly* n *terms of a given value is*

$$P(w, m, n) = \frac{(w-1)^{m-n}\binom{m}{n}}{w^m}. \qquad (1)$$

This formula is obtainable by substituting $\frac{1}{w}$ for p in Newton's formula.

In experiments designed for significance tests, *standard* series of length m are selected from a *parent* series of length $L \geq m$. Selection is defined so that any m terms of the parent series may appear in any order in any one subseries. The standard series thus selected are then used as criteria against which experimental or *comparate* series are matched.

Because of the primary importance and symmetry of the binary choice, we shall first consider operations on binary series and then generalize for series whose scope may be greater. In binary terms the Newtonian ratio becomes

$$P(2, m, n) = \frac{\binom{m}{n}}{2^m}. \tag{1b}$$

Suppose we select all possible subseries of length m from an unbiased binary parent series of length $L = 2e$. The proportion of these subseries containing exactly n terms of a given value is

$$\frac{\binom{e}{n}\binom{e}{m-n}}{\binom{2e}{m}}.$$

Now as e increases the value of this ratio approaches that of the Newtonian ratio. Thus

$$\frac{\binom{e}{n}\binom{e}{m-n}}{\binom{2e}{m}} \to \frac{\binom{m}{n}}{2^m} \text{ as } e \to \infty. \tag{2b}$$

Generalizing for $w \geq 2$ we obtain

$$\frac{\binom{e}{n}\binom{e(w-1)}{m-n}}{\binom{ew}{m}} \to \frac{(w-1)^{m-n}\binom{m}{n}}{w^m} \text{ as } e \to \infty. \quad (2)$$

This formula covers also the case where deviation rather than bias is taken as the classifying criterion. But if the binary parent series is biased the proportions classified by deviation $|d|$ are given by the function

$$\frac{\binom{e+d}{n}\binom{e-d}{m-n} + \binom{e+d}{m-n}\binom{e-d}{n}}{2\binom{2e}{m}}$$

where d is the *bias* of the parent series. It will be seen that where d attains the critical value of $\sqrt{\frac{e}{2}}$ the value of this function approaches that of $\frac{\binom{m}{n}}{2^m}$ much more closely than that of the function for the unbiased parent represented in (2b).

Critical bias may be described as a bias of one standard deviation in a Bernoulli series. A Bernoulli series is defined as a series in which the probability of any particular kind of term at any place is $\frac{1}{w}$. Thus a Bernoulli series may be any one of all its possibilities; the expression implies a haphazard selection of one of a set of series of which length and scope are the only strictly defined attributes. It is true that the term 'haphazard' imposes some additional restriction upon the range of the selection, but it is impossible to say what this restriction is until the word is defined in a way which is both rigid and interpretable, which it seldom, if ever, is.

There can be no doubt that the term 'haphazard' connotes a regressive function of bias, and it is through this concept that we see the ultimate connexion between any assessment of probability and the mental attitude or set of the assessor. Assessment of probability is made subject to a previous assessment of bias; but bias, being assessable in respect not only of any element but also of any *compound* in a series, can be given any of a set of discrete values from 0% to 100% for the same series according to the element or compound in respect of which it is assessed. It is also a truism that, in respect of their larger compounds, all finite series are heavily biased.

We shall not, therefore, speak of Bernoulli series (or von Mises collections) with their vague and even misleading implication of haphazardry; instead we shall confine ourselves to the simple selections of whose biases haphazardry is a complex function. It will readily be seen that once we adopt this procedure the concept of probability itself becomes superfluous, at least in the stage of analysis. For it, too, is a child of haphazardry and therefore a function of bias.

Taking all 2^m binary series of m terms and summing the squares of their biases with respect to a given element, we know that the mean value of these squares is $\frac{m}{4}$. Thus

$$\sum_n \frac{\binom{m}{n}\left(n-\frac{m}{2}\right)^2}{2^m} = \frac{m}{4}. \tag{3b}$$

And for $w \geq 2$ this becomes

$$\sum_n \frac{(w-1)^{m-n}\binom{m}{n}\left(n-\frac{m}{w}\right)^2}{w^m} = \frac{m(w-1)}{w^2}. \tag{3}$$

But since

$$\sum_n \frac{(w-1)^{m-n}\binom{m}{n}n}{w^m} = \frac{m}{w},$$

the function $\frac{m(w-1)}{w^2}$ is in fact the variance of the numbers of terms of a given value in the set of all possible series of length m and scope w.

Now, if we perform the same operation on the set of subseries each of m terms selected from an unbiased binary parent of length $L = 2e$ terms, we find that the variance of the numbers of terms of the given value is in general different from $\frac{m}{4}$ but approaches it as e increases:

$$\sum_n \frac{\binom{e}{n}\binom{e}{m-n}\left(n-\frac{m}{2}\right)^2}{\binom{2e}{m}} \to \frac{m}{4} \text{ as } e \to \infty.$$

This approach in fact occurs for selections from any parent whose bias varies as the square root of its length. Thus if k is finite

$$\sum_n \frac{\binom{e+k\sqrt{e}}{n}\binom{e-k\sqrt{e}}{m-n}\left(n-\frac{m}{2}\right)^2}{\binom{2e}{m}} \to \frac{m}{4},$$

and

$$\sum_n \frac{\left\{\left[\binom{e+k\sqrt{e}}{n}\binom{e(w-1)-k\sqrt{e}}{m-n} + \binom{e-k\sqrt{e}}{n}\binom{e(w-1)+k\sqrt{e}}{m-n}\right]\left(n-\frac{m}{w}\right)^2\right\}}{2\binom{ew}{m}} \to \frac{m(w-1)}{w^2}$$

$$\text{as } e \to \infty. \qquad (4)$$

If we fix the value of m and vary k we find that the approach to $\frac{m(w-1)}{w^2}$ as e increases is either from above or from below, depending on k. But when k is given the critical value of $\sqrt{\left(\frac{w-1}{w}\right)}$ the approach sign is replaceable by the equality sign for all values of e.

Now, since

$$\sum_n \frac{\left\{\left[\binom{e+d}{n}\binom{e(w-1)-d}{m-n}+\binom{e-d}{n}\binom{e(w-1)+d}{m-n}\right]n\right\}}{2\binom{ew}{m}} = \frac{m}{w}, \qquad (5)$$

it becomes possible to formulate the following theorem of selections.

Special Theorem of Selections. *The variance of the numbers of terms of a given value in each of the set of subseries of* m *terms selected from a critically biased binary parent series is equal to the variance of the number of such terms in the set of all possible binary series of* m *terms.*

$$\sum_n \frac{\binom{2\sigma^2+\sigma}{n}\binom{2\sigma^2-\sigma}{m-n}\left(n-\frac{m}{2}\right)^2}{\binom{4\sigma^2}{m}} = \frac{m}{4}.$$

Generalizing for scope we obtain the

General Theorem of Selections. *The mean variance of the numbers of terms in any two given elements in the set of subseries of length* m *and scope* w *selected from a parent series in which these two elements are critically and oppositely biased is equal to the*

variance of the numbers of terms in any given element in the set of all possible series of m *terms whose scope is* w.

Proof. It can be shown that

$$\sum_n \frac{\left\{\left[\binom{\frac{w\sigma^2}{w-1}+\sigma}{n}\binom{w\sigma^2-\sigma}{m-n}+\binom{\frac{w\sigma^2}{w-1}-\sigma}{n}\binom{w\sigma^2+\sigma}{m-n}\right]\left(n-\frac{m}{w}\right)^2\right\}}{2\binom{\frac{w^2\sigma^2}{w-1}}{m}} = \frac{m(w-1)}{w^2}. \tag{6}$$

Now, we know from (5) that by substituting n for $\left(n-\frac{w}{m}\right)^2$ in (6), the LHS becomes equal to $\frac{m}{w}$. This represents the mean bias with respect to any given two elements taken separately of the set of subseries selected from a parent in which these two elements are equally and oppositely biased. But as $\frac{m}{w}$ is also the expectation for series of m terms whose scope is w, it is clear that, assuming the length of the parent series to be $L = \frac{w^2\sigma^2}{w-1}$, the theorem selects from (6) the cases where $\frac{w^2\sigma^2}{w-1}$ and σ are integers.

In classical probability the theorems of selection depend for their precision upon the concept of extending the parent series to infinity. But we see now that this is not necessary in the case of some important functions of deviations in subsets, where critical bias of the parent series is an effective substitute for infinite length.

We turn now to the consideration of matching operations. Here we are not concerned to select finite subseries from a parent of any length equal to or greater than that of the subseries, but with the operation of comparing one finite series with another of the same length. We define matching as the apposition of each term $a_1a_2a_3 \ldots a_L$ in a standard series to each term $b_1b_2b_3 \ldots b_L$ in an experimental or comparate series of similar length and scope, scoring one unit of comparison whenever what have been defined as similar terms coincide. To match a *set* of series against a series means to match each member of the set in turn.

In calculating the bias or deviation of matching scores, their expectation is taken to be that of the standard series. The mean square of the deviations (from absolute expectation) of a set of matching scores is called their *absolute variance*. 'Variance' unqualified has its ordinary meaning whereby the deviations are measured from the *mean* of the set of values observed.

Suppose now we have to match a 'random' series $S_R(L, w)$ against a standard series $S_t(L, w)$. We can say nothing about the particular result, which can be any score from 0 to L since a 'random' series is merely one of the possible series of the set (L, w) selected 'haphazardly'. But we can learn something about the results of matching a very large number of such 'random' series against the standard if we interpret the term 'random' as implying that, within certain limits of tolerance, every possible series of its length and scope will occur equally often in the comparate set. Idealizing this concept, we find that the variance of the scores obtained by matching such a set against *any* standard is $\frac{L(w-1)}{w^2}$ since this operation is

clearly subject to the same algebraic generalization as the calculation of the variance of the terms in any given element.

But in experimental work, where a 'random' series must serve as the standard, we do not normally have time to repeat the experiment until every one of the possible w^L standards has been exhausted, even though this is the only way to give a proper validity to classical significance tests. Instead, we choose a single standard *with little or no elementary bias* and hope for the best. Fortunately, as it turns out, there is some justification for this procedure.

Elementary bias is common to nearly all naturally produced series. Moreover, series sharing the same source tend to be biased for any given element in direct proportion to their lengths. If such series are matched against a standard which itself has elementary bias, the matching scores thereby produced will tend also to be biased. Indeed, if the bias of the standard is maximal, the matching scores will deviate from expectation by the exact amount of the deviation in the experimental series. But in significance tests bias in a matching score, if it is large enough, is taken as evidence for a causal relationship between standard and experimental series. Thus, if the standard is biased, any natural bias which is likely to be present in the experimental series will tend to give results which can be mistaken for evidence of a causal relationship which does not exist.

Ideally a 'random' series is one with which any set of experimental series may be compared to give a set of matching scores whose characteristics in no way reflect the bias of the experimental series unless there is a causal relationship between it and them. This convenient and contradictory property, although possible in a sense, as we have seen, to a whole set of

series at once, is beyond the limited attainments of which a solitary series, even at its best, is capable. Nevertheless, its best is not as bad as we might imagine.

Let us first experiment with an unbiased binary standard. Matching against it an unbiased set, we find that the variance of the matching scores is greater than $\frac{L}{4}$ but approaches it as L increases. Putting $2e = L$,

$$\sum_s \frac{\binom{e}{s}^2 (2s - e)^2}{\binom{2e}{e}} \to \frac{e}{2} \text{ as } e \to \infty.$$

But as we increase the bias of the comparate set both the variance and the absolute variance of the matching scores decrease until, when the bias is maximal, they are zero.

If we perform a similar operation on a binary standard which is heavily biased we find, on increasing the bias of the comparate set, that, though the variance of the scores again sinks to zero, their absolute variance rises to a maximum when the set is maximally biased. Between these extremes there is a critical bias, and when successive sets of increasing bias are matched against a critically biased binary standard the absolute variance of the scores, which is the relevant factor determining an assessment of their statistical significance, remains constant at the value $\frac{L}{4}$. As this is also the value for the variance, as well as the absolute variance, of the number of terms in a given element in the binary set $(2, L)$, we can formulate the following theorem of matching.

Special Theorem of Matching. *If the binary set* (2, L, r) *of length* L *and reciprocal bias* r *is matched against a standard series of*

critical bias $\sigma = \sqrt{\frac{L}{4}}$ *then the absolute variance of the scores for either element will be equal to the variance of the scores for the set of all possible matching arrangements of binary series of* L *terms.*

$$\sum_s \frac{\binom{2\sigma^2+\sigma}{r-s}\binom{2\sigma^2-\sigma}{s}(2s+\sigma-r)^2}{\binom{4\sigma^2}{r}} = \sigma^2$$

Generalizing for $w \geq 2$ we obtain the

General Theorem of Matching. *If the set of all the series of a given length, scope and bias with respect to at least one of two given elements is matched against a series in which these two elements are critically and oppositely biased then the absolute variance of the matching scores is equal to the variance of the scores for the set of all possible matching arrangements of series of this length and scope.*

Proof. Consider the expression

$$\sum_s \sum_t \frac{\left\{(w-1)^{r-s}(w-2)^{s-t}\binom{s}{t}\left[\left(\frac{w\sigma^2}{w-1}+\sigma\right)\binom{w\sigma^2-\sigma}{s}\times (s+t+\sigma-r)^2+\left(\frac{w\sigma^2}{w-1}-\sigma\right)\binom{w\sigma^2+\sigma}{s}(s+t-\sigma-r)^2\right]\right\}}{2(w-1)^r\binom{\frac{w^2\sigma^2}{w-1}}{r}} = \sigma^2 \qquad (7)$$

which can be shown, with a certain amount of donkey work, to be an identity. By interpreting r as reciprocal bias in the set of comparate series of $\frac{w^2\sigma^2}{w-1}$ terms, s as the number of r-terms

in any comparate series which coincide with r-terms in the standard and t as the number of such r-terms which match, it becomes clear that the theorem selects the instances of (7) in which both σ and $\frac{w^2\sigma^2}{w-1}$ are integers.

The matching theorem has a converse which can be generalized in terms of the following algebraic identity.

$$\sum_s \sum_t \left[\frac{(w-1)^{r-s}\,(w-2)^{s-t}\binom{r}{s}\binom{s}{t}}{2\,(w-1)^r} \right] \times \left[\frac{\binom{\frac{w^2\sigma^2}{w-1}-r}{w\sigma^2-\sigma-s}(s+t+\sigma-r)^2}{\binom{\frac{w^2\sigma^2}{w-1}}{w\sigma^2-\sigma}} + \frac{\binom{\frac{w^2\sigma^2}{w-1}-r}{w\sigma^2+\sigma-s}(s+t-\sigma-r)^2}{\binom{\frac{w^2\sigma^2}{w-1}}{w\sigma^2+\sigma}} \right] = \sigma^2$$

It appears, then, that critical series imitate two of the supposed properties of 'random' series; first, in selection, whereby variance in any set of selections from finite critical series is equal to variance in the set of selections which have hitherto been assumed to be taken from an infinite parent; and secondly, in the respect of insensitivity of absolute variance in matching scores to bias in the comparate set, hitherto believed possible only when the scores were averaged for each possible 'random' series in turn. Such insensitivity is seen to exist unrestricted with a *single* critically biased binary standard, and with certain qualifications when the scope of the standard is greater than 2.

XI

BIAS AND STRETCH

In its use as a criterion of statistical significance probability can be strictly interpreted, in the form we gave earlier (Chapter VII), as the ratio of the number of cases considered to the number of cases considered possible. For example, the probability of an elementary bias of exactly d in a 'random' binary series of L terms can be calculated from the modified Newtonian function (Chapter X). But we see now that any elementary or compound bias in a series can be eliminated by considering the series as an unbiased distribution of different but indistinguishable elements. For example, a series of 3 tails and 1 heads may be regarded as a biased binary series, in which case its probability is $\frac{\binom{4}{1}}{2^4} = 0{\cdot}25$; or it may be regarded as, say, an *unbiased* quaternary series with two suppressed elements, in which case its probability is $\frac{3^{4-1}\binom{4}{1}}{4^4} = 0{\cdot}42$.

Thus the nature of a probability *hypothesis* can be reduced to a speculation about scope. A goodness of fit test is really a test to see if the assumption of a particular scope is reasonable. In this respect, a probability estimate is an inductive guess.

It is often more convenient to consider a series as if it were

of scope $x + y$ with y suppressed elements than to consider it as being a very improbable example of a series of scope x. The concept of probability is really like Professor Wisdom's cow. Any single criterion for distinguishing this animal may be absent without interfering with its cowness. In a similar way, if we have a series of 100 ones and noughts and the probability of a nought is estimated to be $\frac{1}{2}$, we may turn any one of these noughts into a one without altering our estimate of the probability. But if we turn all the noughts into ones, the probability of a nought ceases thereby to be $\frac{1}{2}$. In the same way, when we take away all the distinguishing criteria for Professor Wisdom's cow, it loses its cowness absolutely.

Now classical theory says that, in spite of there being no noughts at all in our run of 100, the probability of a nought can still remain at $\frac{1}{2}$. This is like saying that even a creature with no horns, hooves, udder, tail or proper internal arrangements might yet just be a cow. But in fact we say it isn't. Similarly, a series of 100 ones does not have a fifty-fifty chance of producing a nought.

We can now distinguish between bias and stretch. A series is biased if we think that another series from the same source will be biased in the same sense by about the same amount. If we think otherwise, the series is *stretched*. Thus it is convenient to reduce bias in the former case to a negligible amount by introducing suppressed elements, since the biased values are in fact the *expected* ones. In the case of stretch this is not a convenient procedure, since the stretched values are *not* the expected ones. The statement 'with respect to their larger compounds, all series are heavily biased' would be more precise if after the words 'heavily biased' were added the words 'or highly stretched'.

Any unbiased binary series containing an odd number of terms or any unbiased series whose length is not a multiple of its expectation must be stretched. But if we consider a single series there is never stretch: only bias. Stretch is deviation from a *norm*. Bias is deviation from an *expectation*. If all 1,024 of the possible binary series of length 10 are choosable, and one with 9 ones and 1 nought is chosen, we consider it a stretched sample from the norm of 5 ones and 5 noughts. But considered by itself, it is a biased series. Thus stretch is a property of a series in its relation to other series of its kind; bias is a property in relation to another property (scope) of the series itself. We can have a group of series all of which are biased but none stretched—for example, a dozen series of 9 ones and 1 nought. But if we had a group of such series with an *average* of 9 ones, we should expect it to behave *as if* it were a sample from the class of series of scope 10 with 8 suppressed elements.

Thus the probabilities when considering the number of elements of a given kind in a series depend on (*a*) the absolute scope chosen, and (*b*) the length and bias of the parent from which they are selected. The case when the class of all series of the given length and scope are considered possible can be taken as the limiting case of the possibilities when the parent from which they are selected is unbiased (or critically biased) and infinite.

It will be remembered that the probabilities of the occurrence of any one series of given stretch is given by Newton's formula, to which the results for selection of a critically biased parent closely approximate. But now suppose we are selecting standard series for experiment. Stretched series will be rejected. If we are making up standards out of our heads, a stretched

series, at least for the first experiment, will not only be unlikely; it will be impossible. Similarly if a randomizing machine is used to make the selection, it will be regarded as unsatisfactory if it selects a stretched series; and machines will be made which tend to suppress such series. We shall indicate in Chapter XIV the mechanism by which this is achieved. It is true at least to say that the probability of a chance machine's producing a very stretched series is reduced to less than that given by Newton's formula. D'Alembert, as we have mentioned, was right to regard such very small probabilities as being really zero. For it is certainly true that the so-called very small probability that an unbiased binary chance machine will produce 100 zeros in a row is not really 'very small' but zero. We simply consider 100 zeros from an *unbiased* machine to be a contradiction in terms. We should probably say the machine had got stuck. Of course, if we were creatures who habitually observed enough cases for the whole gamut of the 2^{100} possibilities to occur to us several times, then 100 zeros together here and there would cause us no alarm. But at present we neither live long enough nor count fast enough to realize this laudable aim. So whenever we see 100 zeros together, although we are at liberty to assume that the series is enormously stretched, we shall never in practice (though we could in principle) verify or falsify this assumption. We assume, therefore, that the series is grossly biased but *only moderately stretched*, since this assumption is verifiable in practice by taking more series from the same source.

In other words, we do not assume more stretch than could be substantiated as such within countable instances. That is why $p = \frac{1}{20}$ is such a popular criterion of significance. A

stretch giving, say, $p = \frac{1}{19}$ is easily verifiable *as such*, since it has a good chance of appearing again in 100 experiments. But it would be silly to take $\frac{1}{10,000}$ as a criterion of significance, assuming that, say, $p = \frac{1}{9,000}$ was only stretch; for here, especially if the experiment is cumbersome, we have probably gone beyond the practical limits within which the concept of stretch can have a verifiable meaning.

Thus, within the logical framework of the experimental set-up, though such 'very small' probabilities, such as 100 zeros in an unbiased binary series of 100 terms, might exist if we were made differently, they are, with our present capacities, quite indistinguishable and therefore meaningless. And whenever we do observe 100 zeros in a row, we assign to them not a particular type of totally unverifiable probability (which would be a counsel of despair), but a probability of a different kind which we can afford to verify. Whether 100 zeros occur more or less often than 99 zeros in an unbiased binary probability series is a question which we cannot hope, with present methods, to answer. It is therefore unprofitable to consider it and so, in applied probability, we should take both probabilities to be equal and zero.

In terms of significance, if our assumption of bias without stretch in a series leads if verified to a large gain and if falsified to a small loss, we use a loose criterion of significance. For example, to save the life of a loved one we will try almost any remedy, however unsubstantiated by previous tests, if only it appears to be the best available. If, on the contrary, verification of bias without stretch leads to a small gain while its

falsification leads to a large loss, we need a stringent criterion of significance before we act on the assumption. For example, if the successful farming of a piece of land depends on a yearly rainfall within certain limits, and if a single year without the proper rainfall spells ruin, then we make the most stringent tests on the meteorological records of past years before we decide to farm the land. In the former case we are hasty because we cannot afford to wait: in the latter we are slow because we may not be able to afford to hurry. The criterion of significance chosen depends partly upon how much counting-time we have. This is why longer-lived animals can adopt more stringent criteria.

But in no case can it be to our advantage to adopt a criterion of significance so stringent that even the maximum counting-time we could spare would not be sufficient to verify its p-value in a chance or null set-up. p-values like 10^{-35}, which are so much prized in a certain type of investigation, cannot possibly give us greater certainty than, at the outside, a p-value of 10^{-3}; and a p-value of no less than 10^{-2} serves in nearly all cases as the most stringent practicable criterion, since if an experiment is at all difficult we are not likely to do it more than 100 times.

XII

BERNOULLI'S THEOREM

IT is an interesting fact that the probabilities referred to in Bernoulli's Theorem are of an unverifiable order. The Theorem, it will be remembered, runs as follows.

If the chance of an event occurring in a single trial is p, *then the probability that the ratio of the number of times the event occurs to the number of trials differs from* p *by less than any preassigned quantity, however small, can be made as near certainty as desired by increasing the number of trials.*

The Theorem suffers from a circularity similar to that of Newton's Theorem: the original value of p is ascertainable only from the series of trials. Any original estimate of p has therefore only an infinitesimal chance of being correct, and we may assume that, in practice, the value we decide upon will be wrong.

Let us call the small preassigned difference in Bernoulli's Theorem ϵ. Let us also assume that our original estimate of the probability of an event in a single trial differs from its true probability by small quantity δ. Now if ϵ is larger than δ, the approach to certainty asserted by Bernoulli's Theorem still occurs. But if ϵ is smaller than δ, then the approach is not towards certainty, but towards impossibility. So in a practical interpretation of Bernoulli's Theorem, the contrary holds. In other words, it is true to say

If the chance of an event occurring in a single trial is estimated to be p′, *then the probability that the ratio of the number of times the event occurs to the number of trials differs from* p′ *by less than a small enough preassigned quantity can be made as near zero as desired by increasing the number of trials.*

It follows that the probability ratio over a number of trials will tend to differ from the estimated probability by an increasingly significant amount.

It is commonly supposed that if we take a penny and begin tossing it, the more we toss it the less significantly the ratio of heads to tosses will tend to depart from $\frac{1}{2}$. But, as we now see, the contrary is true: the ratio will tend to deviate more and more significantly from $\frac{1}{2}$ as the number of tosses increases. Practitioners of probability should not be surprised, as they invariably seem to be, on finding a long series produced by one of their carefully made randomizers to be significantly biased. This surprise seems to continue in spite of the fact that almost all long series, from Weldon's dice data to Rand's random digits, have been found to be significantly biased when tested. Those which have been subsequently published, such as Fisher's tables and Rand's digits, have had the bias removed before publication.

An interesting exception occurs in the published part of the series generated by the Kendall and Babington Smith machine. Here the lack of bias is almost certainly due to feedback from the operator which provided an adequate check to any bias which looked like becoming excessive. This conjecture is strongly supported by the fact that the 10,000 digits run off by an operator who was unaware of what tests were to be applied turned out to be so biased that they could not be published.

It is thus obvious that any series in which the proportions of the elements are not at once apparent (which is one of the criteria for randomness) will have, as its length grows, an increasing chance of containing them in proportions which differ from the original estimate (or expectation) by an amount of any significance, however great.

Thus

If the chance of an event occurring in a single trial is estimated to be p′, *then the probability that the ratio of the number of times the event occurs to the number of trials differs from* p′ *by an amount of any significance, however great, can be made as near certainty as desired by increasing the number of trials.*

Bernoulli's Theorem is seen to be the most symmetrical but least practical of a number of parallel theorems about the behaviour of ambiguous series. Moreover, it and its parallels depend for their meaning on there being a 'true probability' of an event's occurring at any given trial. But since the 'true probability' may be estimated only as the ratios in very long series of trials, and since it is admitted that any such estimate is almost certainly wrong if the series is continued further (and the further it is continued, the greater the certainty), it seems odd that, having been wrong with increasing certainty as the series lengthens, it should at infinity suddenly come absolutely right.

But whether it comes right or not at infinity is literally a metaphysical question: it cannot even in principle be answered by a practical test. The 'true probability' so necessary for any interpretation of the Bernoulli Theorem and its parallels is itself a metaphysical and therefore uninterpretable concept. Thus totters the whole edifice of Bernoullian probability.

Bernoulli's Theorem is about as relevant to some of the

practical problems to which it is applied as is the fact that the chance of choosing a six at random, given the choice of all integers, is zero is relevant to the fact that six is frequently chosen. And the fact that even if six were always chosen the proposition of its infinitesimal probability would not be falsified, is sufficient to show that any limit concept of probability is devoid of either scientific meaning or usefulness. In short, Bernoulli's Theorem, with its stronger forms invented by subsequent investigators, its parallels stated here and its curious conversion by von Mises, is one of the most flagrant metaphysical excesses ever to pass as an adjunct to modern science.

XIII

SOME PRACTICAL CONSIDERATIONS

STATISTICAL significance is generally concerned with matching probabilities, and it is to these we now turn our attention. The probability of excessive stretch in a matching score is not smaller than classical theory calculates (as it is in a single series), but greater.

We have shown that a single excessively stretched series is for practical reasons classed (by ascription to it of suppressed elements) with series of similar bias but only moderate stretch. Thus, large probabilities are made even larger by their cannibalization of the small probabilities of the very unlikely events which are classified as having zero probability. Thus everything that happens is made out to be at least moderately probable.

But this means that no standard series of given scope may be excessively stretched. And there is a further refinement in tests for randomness which reject also series which are excessively near to expectation. Thus fashioned by intuitive genius, the tests tend to exclude *all but critical and nearly critical series*. The approach to criticality is not accurate since by equal chopping of excessively biased series from the ends and excessively unbiased series from the middle we reach not the point of standard deviation, but the point of probable error. A closer approach could be made by removing series

in the respective proportions 8 : 17. A near enough practical approximation would be to remove twice as many series from the middle unbiased region as from the outer biased regions.

Thus, for example, the Kendall and Babington Smith Tables, which were tested for randomness at the 1% and 99% levels of significance, would, if my thesis is correct, serve their purpose better if they had been tested at 1% and 98% levels. If they had been tested even more stringently, using 4 tests at the 5% and 90% levels, nearly 48% of the possible standards would have been rejected. This means at least that probabilities computed for the scores obtained by matching two such series would be considerably less on the classical theory than in actuality.

Experiments, of course, don't always involve matching two such series. Rather, they consist in the use of one tested standard and one natural or experimental series. But if the natural series is a series of guesses (as in the Fisher teacup experiment) then the set-up is in some ways one in which two tested series are matched, since a guesser tends to avoid the same sorts of bias in guessing as are excluded in tests for randomness. It is true, then, that, in all guessing experiments at least, the significance of the results is over-estimated by classical theory. And, by the mathematics of the matching set-up, a large significance is over-estimated more than a small one.

It is possible to construct series which are nearly critical in several respects. For example, the binary series of 64 units is also a series of 16 elements 16 units long. This length almost, but not quite, coincides with the calculated length of such a 16-element series with a critical bias of 1 with respect to any of its elements. Thus by putting in two each of half of our

possible compounds of 4 and by leaving out the rest we can produce a series which is exactly critical for units and nearly critical for all compounds of 4. For compounds of 2 and 3 its approach to criticality is not so close. An interesting thing about these series is that they tend to look random however we arrange them. But the factor by which we must multiply the probability calculated on the binomial hypothesis to correct it for the matching hypothesis now becomes much larger.

It is an interesting fact that although it is possible to construct series which are *unbiased* for all compounds up to those of length n in a series of minimum length $L = nw^n$, it seems to be impossible to construct any series which are critically biased for more than one type of compound. If the inequality

$$\frac{w^{2m} ma^2}{w^m - 1} \neq \frac{w^{2n} nb^2}{w^n - 1}$$

holds in all cases where all variables are integers and $a \neq b$, then we can say it is impossible in all cases. Of course, it must always be possible for a series to approach criticality in as many respects as we like as nearly as we like. The intervals by which we miss it can be made as small as we please by choosing series of appropriate length. They are like the commas of Pythagoras which must exist in harmonically constructed musical scales because of the inequality $2^m \neq 3^n$ for integral values of m and n.

We have already remarked that not all experiments are analysable in terms of the matching set-up we have devised above. Suppose we wish to test the hypothesis that a particular fertilizer has the effect of increasing the height of the plants to which it is given. We might try scattering it around a field in particular places, and then compare the heights of the resultant

crops at some specified later date. There is now, of course, no reason whatever for 'randomizing' the places where the fertilizer is spread—no reason, that is, if we are entirely honest. But suppose, in order to advertise the fertilizer, we put it in the spots where we know the crops tend to grow tallest in any case. Now of course our results are vitiated. In this sort of experiment a randomizer is no more than an aid to honesty. But if, by chance, the use of an approved randomizing procedure led to the scattering of fertilizer in all the best places, we should not be justified in accepting the result of the experiment. Knowing what had happened, we should have to reject it just as firmly as we should reject a dishonest experiment.

We thus find two distinct uses for randomizers in science. In the first place they can be used to validate experiments in communication, such as the Fisher teacup experiment. The validation here is of a special kind, demanding a multiplication of the probabilities as calculated by classical theory by a specific and calculable factor. The other and quite different use of the randomizer is merely to save the experimenter from his own potential dishonesty. And by potential dishonesty I do not mean anything that is necessarily corrigible. For, given a very complicated set-up, a completely honest procedure may be very difficult to determine. The randomizer, because of its special mode of working in relation to the experimenter which we shall discuss in the next chapter, can save us from types of dishonesty otherwise too complex for us to be able to correct.

It may be a source of annoyance to some readers that, having shown the concept of probability to be self-contradictory, or at least paradoxical, we yet continue to use the word

'probability' and, in general, the concept of chance, without comment in subsequent chapters. Perhaps, therefore, we should now analyse more deeply the foundations of the paradox in order to resolve it.

A paradox is like a conjuring trick with words. It depends always upon a subtle transformation in the meaning of one of its terms in the course of the argument. The 'most ingenious paradox' in *The Pirates of Penzance*, in which a man of twenty-one who was born on the 29th of February is reckoned by birthdays as being only five, depends for its effectiveness upon the simple confusion of birthdays with years. Similarly the probability paradoxes produced in Chapter IX depend for their effectiveness upon two hitherto indistinguished meanings inherent in the frequency concept of probability.

When we say that to verify the proposition that the probability of event E is p, the ratio of E to $E \,\&\, \overline{E}$ must be p, we are referring to probability as measurable in terms of single events or elements. But when we say that in order for the set-up to remain a chance set-up, the ratio of E to $E \,\&\, \overline{E}$ must at some stage take any other of its possible values, we are referring to the concept of probability applied to molecular events or compounds. It is assumed in classical theory that certain simple rules for transforming the one type of probability into the other are practicable. For example, it is assumed that if the probability of event E in one trial is $\frac{1}{x}$, then the probability of getting nothing but E's in 100 trials is $\frac{1}{x^{100}}$. But we have seen that this is not so and that, rather than being complementary, these two probabilities are in the limit contradictory. We have here merely an extension of the

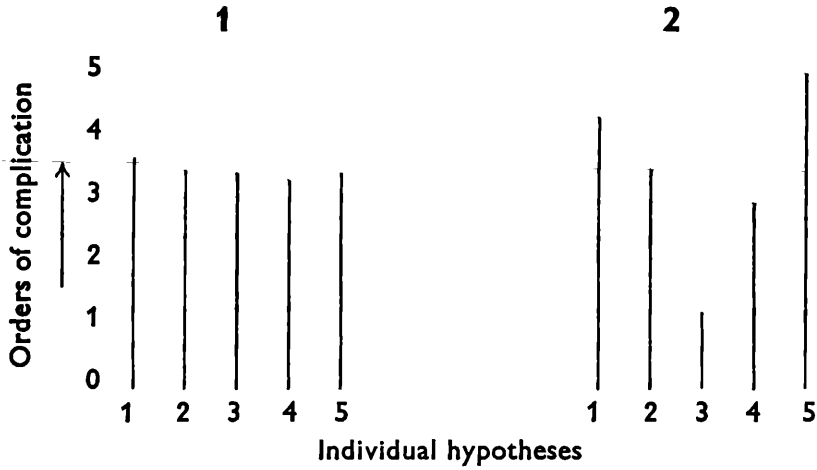
concepts of primary and secondary randomness into the concepts of the elementary or compound bias with reference to which the probability estimates are made. Once we have made this distinction, we are again allowed to use the concept of probability, and simple inspection is sufficient to show in what sense we have subsequently used it. But in resolving the paradox we have not reached the end of the road. For, besides being paradoxical, the frequency concept of probability is also vague.

There are two forms of vagueness: weak and strong. The term 'alcohol' is vague, but only weakly so. Weak vagueness can be eliminated by further definition. We can take away the vagueness of the word 'alcohol' by qualification—for example by addition of the term 'ethyl'.

Strong vagueness cannot be removed in this way. We are vague about the position and velocity of an electron because our methods of observing them react upon these variables to produce an uncertainty which is absolute. We are similarly vague in our sociological propositions: for if we publish predictions about what people are going to do, the people might confound the predictions out of perversity. They might be like the ruler who commanded a seer to foretell him the future so that he could take steps to prevent it from happening.

We see that vagueness is strong when we try to describe a quality or variable which changes because it is observed. The concept of probability is a variable of this kind. A random series is like a drawing of a stairway with no perspective. First we seem to see it from above, then suddenly we see it from below. In terms of a stairway, the drawing is ambiguous. The ambiguity can be resolved by considering the drawing as a geometrical pattern.

Our impressions of a random series should suffer from this kind of indecision. A random series is one which ought to appear ambiguous; we try hypotheses of its formation in turn, only to reject them one by one. As with a Rorschach ink-blot, we cannot decide upon any one interpretation; and any suggestion of a particular interpretation is likely to lead to its rejection in favour of another. It is in this latter quality that the strong vagueness of probability resides. A series consisting of nothing but ones is not ambiguous because, of all the hypotheses we can make about its rules of formation, the formula 'put one after one' appears by far the simplest. But a long series of coin tossings is generally a series about which we may have many hypotheses, none of which we can say is definitely simpler than some of the others. In the former example, Occam's razor served to give us a definite rule; in the latter, the razor is not sharp enough.



The shape of the graph—whether the gradient between simplicity of hypotheses is flat or steep—is dependent upon the observer to whom the data are presented. For example, a hundred places in the decimal expansion of π starting at the

500th digit would appear random to most people; it could be accommodated on any line of the first diagram; but to others who knew π to 600 decimal places, these 100 digits would have a simple interpretation. They could be represented by hypothesis 3 in the second diagram. Here only can Occam's razor be used effectively.

The advance of knowledge consists in the making of steep gradients between hypotheses where only flat ones existed before. But this is a process dependent upon the ultimate nature of observers; for what is simple to one may be complicated to another. Common knowledge is possible only through a measure of agreement in the nervous structures of communicating animals.

XIV

THE CHANCE MACHINE

I AM told to make up a series of random digits 0 to 9. My first concern will be to make sure that none of the digits occurs very much more often than any of the others. If I notice, say, sixes appearing more often than the other digits, I give them a rest. After going on like this for some time, I shall begin to notice the frequency of other combinations. I might discover that sixes were more frequently followed in my series by sevens than by any other digit, and I should therefore strive in future to reduce this tendency. And as the series got longer and longer, I should notice in it more and more complicated types of repetition which I should have to avoid. If I was very sharp and noticed all these repetitions before they happened too often, I should have produced a series which passed all the tests for randomness which the tendencies I noticed suggested. My colleague Mr Babington Smith is at present engaged upon producing such a series by this means, so we are at liberty to reflect upon its validity.

Now each event in a series that is to be called random is supposed to be independent of the events which have gone before it. But in the method of producing the series which we have described, our success is dependent upon a very careful watch on what has gone before, with subsequent modification of what is to come. If we were not already suspicious of the concept of randomness itself, we might regard this as a

paradox. The only alternative is to believe that the digits in random series are not in fact independent.

The French philosopher Poincaré described a chance machine or randomizer as a machine in which a very small cause produces a very large effect. Thus when we cast a die, it needs only a very small alteration in the initial conditions to produce a very big difference in the result. He could have said also that a different cause can produce the same result. For example, two throws, although made differently, could each result in the turning up of the same face. This is an important condition in the randomizing set-up which I propose now to examine in some detail.

Poincaré's notion can be clarified as follows. Let us imagine a lever which we can push back and forth. This lever is connected by a system of gears to a pointer which moves across a scale. The gearing is such that by moving the lever a millimetre we move the pointer a metre. Suppose now we can place the lever at any position we like within plus or minus a tenth of a millimetre. This will mean that, using only the lever, we shall not be able to place the pointer with a minimum accuracy of less than a decimetre. Suppose the pointer scale is divided up into centimetres which are in turn divided into millimetres and that the first division in each centimetre is marked 0, the second 1, the third 2, etc. up to 9. We now have, in essence, a randomizer. All the experimenter has to do is place the lever in some pre-arranged position, and then note the figure to which the pointer points. He cannot, by definition, point it at any figure he likes; that is because his error of manipulation could be anything up to 100 divisions away from the figure he aims at.

It will be seen at once that only a very slight modification

is necessary to turn this machine into the machine used by Mr Babington Smith in the production of the well-known Kendall and Babington Smith random sampling numbers. In the latter machine a rapidly revolving wheel whose circumference was painted with the figures 0 to 9 was illuminated by electronic flashes under hand control. The number appearing at a particular point at each flash was noted and published as random. The error which this machine magnified was the error from regularity of the flashes. Given perfectly regular flashes and a wheel revolving at a constant speed, however fast, there is no randomization. The sequence produced will itself be perfectly regular. But given a certain error from regularity in the flashes, provided that that error covers a maximum of several revolutions of the wheel, then randomization occurs. Let us suppose, then, that, in the Babington Smith machine, these conditions are fulfilled. Let us suppose that, in attempting perfectly regular flashes, the experimenter is subject to a maximum error involving, say, 10 complete revolutions of the wheel. Using this machine, he sets about to produce a table of random sampling numbers. What are the logical implications of his behaviour?

One of the first things we can say about the set-up is this: if the experimenter wants to produce a particular number, he won't be able to except by chance. He would be silly to take on an evens bet that he could produce at the next flash, say, a 9. He might be lucky once, but if he did it often he would probably lose. In fact, if he did it often enough and didn't lose, we should eventually say that he had more control over the machine than we had hitherto suspected and therefore that the machine was not a randomizer. The degree of his control is defined in terms of the results he produces. Thus we

can be sure that, provided the experimenter wants to produce a particular number, that number will appear with only its proper frequency in the random series.

Let us now suppose our experimenter goes on working his machine for a long period. After a time, the flashes he produces get into a rhythm. There is no question here of control, since the experimenter is working automatically with no particular goal in view. It is known that behaviour patterns of great regularity can take place although it is impossible for the subject to choose what sort of regularity should occur or for him to repeat it at will. Such regular behaviour will eventually become apparent as a regularity in the sequence of numbers produced. The moment the experimenter notices this regularity he will begin to suspect the machinery; he will also begin to test the machinery by noting whether or not the regularity continues. His behaviour from this point can be interpreted as an experimental attempt to continue the regularity. This, we have seen, is bound to fail. It is bound to fail because it is a form of regular procedure by definition outside the subject's control. The subject here is like the centipede who was asked how he remembered which leg to move next. We all know what happened to that poor beast, and it is the same with the man at the randomizer. Acquiring a reputation for behaviour produced automatically, he tries to reproduce it consciously but it disintegrates. When it disintegrates it makes way for some new regularity to replace it; and this new regularity will continue until it in its turn is noticed and disintegrated.

Let us now see what sort of series will be produced by this set-up. It will be a series closely resembling that which Mr Babington Smith is producing out of his head: for it continues unchecked until some regularity or bias is noticed

by the experimenter. But whereas in the case where no machinery is used the experimenter must subsequently be careful to avoid the patterns he has noticed, all he has to do when using the machine is to attempt to repeat them. An attempt to repeat a pattern is less of a strain than an attempt to avoid it; therefore it is to be suspected that the use of a randomizing machine will give slightly better results, or the same results more easily, than an attempt to produce a series to pass the tests without it. But we must not be lulled into supposing that a series produced with the machine will *necessarily* be any better than a series produced without it, *provided the latter is done carefully enough.* The two series finally should pass exactly similar tests for randomness. The machine is a useful adjunct to help us combat certain animal tendencies; it is by no means essential to the randomizing process. Its usefulness derives from the psychological fact that it is easier to repeat than to avoid doing what we have done before. And we must not repeat ourselves if we are writing out a random series.

Using a randomizing machine, all the experimenter needs to do is to keep in mind the patterns in the series he wishes to avoid producing. His natural tendency to reproduce them will do the rest. The randomizer employs information feed-back as an inhibitor, and thus relieves the experimenter of the responsibility of using his own less reliable power of inhibition.

If this is a true description of the randomizing set-up, then we should have evidence of it in the behaviour of series produced by these means. I shall refer to two examples as illustrating this most clearly.

The Babington Smith randomizer was used almost entirely by Mr Babington Smith himself and one assistant. When Mr Babington Smith used the machine he had in mind all the

tests which he and Kendall proposed to do upon the series he produced, and the series, when he had produced it, did in fact pass the tests. But for some of the time Mr Babington Smith's assistant, who did not know what tests were proposed, worked the machine. He produced 10,000 digits which were originally intended for publication with the rest. But upon examination they failed by a very big margin to pass even a simple frequency test and had to be rejected.[1]

The other sort of evidence comes not from the probability worker but from the psychical research worker. In the last twenty years psychical research workers have turned their attention to dice and other randomizing machines. One of their more constant observations has been that patterns which have built up a high degree of significance in long series of randomizing or matching data disappear when the series are continued after the patterns have been noticed. The explanations given by the experimenters of such matching patterns and their disappearance have been somewhat complicated and do not concern us here; the chief interest such patterns hold for us lies in the fact that the model of randomization suggested in this chapter explains them.

The essence of randomness has been taken to be absence of pattern. But what has not hitherto been faced is that the absence of one pattern logically demands the presence of another. It is a mathematical contradiction to say that a series has no pattern; the most we can say is that it has no pattern that anyone is likely to look for. The concept of randomness bears meaning only in relation to the observer; if two observers habitually look for different kinds of pattern they are bound to disagree upon the series which they call random.

[1] *J. Royal Statistical Soc. Supplement*, **6**, 58 *sq.* (1939).

XV

THE DIMINISHING FIELD

We have seen that the degree of probability we attach to any matching score is dependent upon the kind of bias which we consider relevant in the standard series. Our final p-value depends quite arbitrarily upon what order of compound, or which combination of orders, we are considering. We have seen further that if we have reason to suppose that there is relevant bias in compounds where nw^n exceeds L, our test of significance is vitiated in all cases and the p-values obtained mean nothing.

Let us consider a case where we have used a standard series upon which we have performed four independent tests from randomness at the 5% and 90% levels. We re-estimate our probabilities on the basis of having excluded some 48% of the possible Bernoulli series. We will now suppose that, having done this and arrived at certain p-values, we make further tests for randomness upon the original series. If the series passes these tests also, we are put in an awkward predicament. For by making the subsequent tests we have now excluded even more than 48% of the possible standards available, and must revise our p-values accordingly. We see at once the obvious but alarming possibility of continuing this procedure further.

In converse, what we have shown is that any experimental result involving a significance test can be vitiated by the

discovery of a pattern in the original standard series. If there are in existence at the present moment a certain number of statistically significant results, and if no further experiments are made, then the subsequent history of these results will be as follows. In the first place, all the results will have the form of validity which statistical results in general may possess: that is, each probability will be assessed on the grounds that we have no suspicions of any relevant bias in the standard series which were used. As time goes on and we examine the data more closely, we are bound eventually to find some kinds of patterning in the standard series which before had passed unnoticed. As we do this with each standard in turn, the *p*-value for the experiment in which the standard was used becomes suspect. Thus, one by one, each result is vitiated.

This process can be described as a tendency to diminution of scientific knowledge in the absence of further experimentation or confirmation. Left to itself, the world of science slowly diminishes as each result classed as scientific has to be reclassed as anecdotal or historical. Thus, in the absence of further research, all science eventually becomes history. Science is a continuous living process; it is made up of activities rather than records; and if the activities cease it dies. Science differs from mere records in much the same way as a teacher differs from a library.

It is by constant repetition of experiments, using different standards, that scientific knowledge is kept alive. And here now we see the force of the fundamental principle enunciated by Sir Ronald Fisher which led to the Experimental Paradox. That principle, we remember, was that no single experimental result, however significant in itself, could suffice to demonstrate any natural law. The paradox resulted from

saying that any number of experiments could be counted as a single experiment. But on the diminishing field hypothesis it can at once be resolved.

The difference between a single experiment and a series of experiments is only this: a single experiment is likely to become anecdotal more quickly. The series of experiments is not exempt. To find a pattern common to all the standards so far used is only a matter of time; and once we have found it, the whole series, like the single experiment, becomes anecdotal. What is asserted is the living quality of scientific method. Scientific knowledge, like negative entropy, tends constantly to diminish. It is prevented from dwindling completely into anecdote only by the attitude which seeks to repeat experiments and confirm results without end.

APPENDIX I

ON MIRACLES

When we pray for a miracle, we ask for something very improbable. Left drowning in the middle of the Pacific, we say: 'Only a miracle can save us now.' The use of the word 'miracle' is valid because we have in mind the extreme smallness of the probability of a ship or a helicopter coming by. If we happened to know that a helicopter with the proper rescue apparatus was on its way, we should not despondently talk of miracles.

We thus note a resemblance between miracles and the results of scientific experiments: for the result of a scientific experiment is highly significant (and therefore much prized) when the probability of its occurrence on a null hypothesis is very small. What prevents most scientific results from being miracles is the fact that they keep on occurring so often. We begin to know that they will happen, and this vitiates their validity as miracles. Highly significant scientific results would turn into miracles only if they never (or hardly ever) happened again.

The phenomena of psychical research, as we have already observed, behave much after this fashion. Here results of great significance have been obtained; such significance has sometimes, although more rarely than is usually supposed, built up over moderately long periods of several weeks or months. But the end is always the same; at some stage in the

experiment the results fall off to insignificance, never to recover. The falling-off may be gradual or sudden, but once it has happened it usually marks the end of the particular kind of result, although the same set-up might later produce significant results of a different kind which undergo the same history.

Now the trouble with results like these is that eventually they all cease to be valid as science as, one by one, each becomes transformed into anecdote. It is no use Professor Rhine or Dr Soal protesting that in 1934 or 1943 they obtained results of significance of the order of such-and-such (giving, in most cases, a *p*-value wildly beyond the limits of empirical verifiability); for if neither they nor anyone else can repeat these results, then their findings must at length succumb to the inexorable closing-in of the boundaries of the diminishing field. Of course, this is not an end to the matter; results of great statistical significance must excite comment whatever their final outcome. And the comment, where it is not directed against the soundness of the experimental methods themselves, can react, after the law of the diminishing field has taken its toll, only against the usefulness of the statistical procedures by which their significance was calculated.

On classical theory the results have remained a mystery. Small probabilities used as significance criteria are supposed to indicate demonstrable repeatability. But the trouble with psychical research results is that their repeatability never turns out to be a function of their significance. They comprise, in fact, the most prominent empirical reason for beginning to doubt the universal applicability of classical frequency probability.

But, in the light of the thesis here presented, and in particular in its relation to the workings of chance machines, these results are not so mysterious. Indeed, the thesis demands them. We see from Chapters XII and XIV that a chance machine left running without disturbance will produce results which are, by any standard we choose, as significantly biased as we please. But a good chance machine is not set to work freely without intervention. It employs feed-back from its operator or observer to prevent an excess of the patterns he has in mind. In this latter kind of machine, significant patterns can build up only as long as they remain unnoticed. The moment the operator sees them, a negative feed-back circuit is closed which at once inhibits further repetition of the particular compound considered.

The validity of a test of statistical significance is always dependent upon our being able to consider the experimental set-up as a pure chance machine. The final result, consisting of a series of matching scores, can be taken as the output of a simple randomizer. We call this output 'significant' or 'insignificant' according to whether we think it is biased or stretched.

Now, as we have seen, a randomizer left entirely to itself will produce a series whose proportions differ as significantly as we like from the original probability-estimate. But we do not expect its tendencies to remain exactly constant. Even without intervention its bias is bound to change slightly as, for example, it wears out. We may thus be presented with the spectacle of a particular kind of bias first building up to great significance, then gradually diminishing. This is indeed frequently observed in psychical research. But what is much more dramatic is the sort of significance which has built up

over a period and which is suddenly noticed by the experimenter, after which it disappears completely. This sort of occurrence has become so common that ardent psychical researchers like Dr Thouless have attempted to devise means of preventing it in the planning of their experiments. These means consist mainly of never looking to see if anything peculiar is happening until the end of the experiment. In other words, the whole experimental set-up is made as like as possible to Mr Babington Smith's randomizer when his assistant was working it. Every effort is made to block the feed-back circuit which would effectively inhibit the peculiarity.

Aristotle's remark that we cannot demonstrate the fortuitous is particularly relevant here. For, if my thesis is correct, the whole organization of psychical research is bent on demonstrating what cannot be demonstrated. No one can demonstrate that he is a lucky person without bringing upon himself the suspicion of sharp practice. We have sometimes the brief realization of being in a state where nothing seems to go wrong. This realization is followed all too frequently by a break-down of the uncontrollable mechanism by which it is maintained.

Some people do in fact seem to be able to maintain their runs of luck better than most of us. To them the knowledge that they are in luck does not always bring about its catastrophic cessation. Interestingly enough, these people tend (as we have now discovered from psychical research) to be extraverts rather than introverts. And, as we know, the extravert can more easily and effectively set up blocks to information exchange within his own nervous system than can the introvert. Thus, although an average card guesser who is having

a run of success in, say, a telepathy experiment nearly always begins to fail the moment she is told of her success, this may not happen if she is hysterical. The information-block is present even here; but it is now internal. These remarks, of course, apply also to the experimenters themselves, some of whom can get results where others can't. They also explain Professor Rhine's finding that the more people observe a psychical research experiment, the less likely it is to succeed. The inference here is obvious; the more people there are about, the more likely becomes the establishment of a negative feed-back circuit.

The average psychical researcher is interested in mysteries, but has not always faced the fact that once a mystery is explained it can no longer be mysterious. If you are interested in mysteries as such, then a scientific attitude is the last thing you should indulge in. Mysteries are observations we do not know how to classify. Random series are mysterious in precisely this sense. The fundamental error of some psychical researchers is to pose as scientists who are interested in mysteries *as such*. This only brands them as sheep in wolves' clothing.

APPENDIX II

ON PRACTICE

A medium claims that she can tell whether we are thinking of an odd or an even digit. But she maintains her faculty is reliable only in the mornings. We decide to test her claim.

We design an experiment to last one five-day week, giving her 100 guesses each day, 50 in the morning and 50 in the afternoon. We arrange that after each guess she shall be told whether she was right or wrong. As a safeguard against parallel thinking or special ability at the heads and tails game we decide to use published random numbers. The first 500 random digits from the statistical tables of Sir Ronald Fisher and Dr Yates are admirably suited to our purpose. Each column contains 50 digits, and the columns are arranged in pairs. We will use a pair of columns for each day. The five pairs of columns in the first block of digits are thus sufficient for the whole experiment.

The first column from the tables transcribed into odds and evens is given below, together with the information whether the subject's guess was successful (S) or unsuccessful (F).

01111	10011	10011	11111	01000
FFSSS	SFSFS	SFSFS	SSSSS	FFFSS
11111	01000	00001	11001	10001
FSSSS	FFFSS	SSSSF	SSFSF	SFSSF
11001	00111	00101	00101	00101
SSFSF	FSFSS	FSFFF	FSFFF	FSFFF
01101	00000	11000	10101	01100
FFSFF	FSSSS	FSFSS	FFFFF	FFSFS

Thus in the morning of the first day the medium scores 32 successes—7 above chance expectation. In the afternoon she scores only 19—6 below expectation. The morning score is quite suggestive, but the difference between the scores in the morning and the afternoon, 13, is even more impressive. The critical ratio of the difference is more than 2·5, which is quite significant. In subsequent days, therefore, we shall look out for both prevalence of high scores in the morning, as

suggested by the original claim of the medium, and also the prevalence of large differences in the same direction between the morning and the afternoon scores.

Table I summarizes the results of the complete experiment. It will be seen that both the high morning scores and the differences continue, though in each case the scores tend to fall off towards the end of the experiment. But in each case, especially the latter, quite highly significant scores are built up. The medium's claim would appear, therefore, to be established.

TABLE I

Runs of odds and evens in the Fisher Tables.

Scores above (+) or below (—) chance expectation.

Day	1	2	3	4	5	Total	C.R.	p
Morning	+ 7	+ 2	+ 5	+ 5	+ 5	+ 24	3·03	<0·004
Afternoon	— 6	— 6	— 4	0	+ 1	— 15	1·90	
Differences (expected direction)	13	8	9	5	4	39	3·48	<0·0006

Let us now suppose we try to confirm this result by doing a similar experiment using the next column of 500 digits. And let us also suppose that none of these tendencies is confirmed. What are we to say? If we follow the fashion of psychical research, we shall say that the medium displayed a telepathic or clairvoyant power which she subsequently lost. But if we follow the fashion of the stricter sciences, we might be made suspicious by the lack of confirmation.

Let us go back and examine our data more closely. In the list of failures and successes in the first two runs we note that a success occurs when and only when the symbol presented is the same as the immediately preceding symbol. This suggests that

the medium determines every guess by reference to the success or failure of her previous guess. Examination of the subsequent scores brings nothing to conflict with this hypothesis. There is therefore overwhelming evidence that the results have been determined mechanically and not by clairvoyance. All they reflect is a significant tendency for long runs to occur in the odd columns.

There are now several pertinent observations we can make. First, we see how easy it is to obtain highly significant results which reflect nothing more than an artifact in the randomizing device used as a standard. Secondly, we note that a tendency once discovered, although building up to considerable significance, slowly dies away throughout the experiment. And, lastly, we note that a subsequent search for the effect in a continuation of the series is fruitless.

The main purpose of this mock experiment has been to show that significant results like these can occur in any attempt to make a random series. I have made it into a plausible psychical research experiment mostly for the purposes of amusement. The design is based on an experiment by Dr Thouless, who thinks[1] that he obtained evidence showing that psychical research experiments give better results in the mornings. The results given here are, by the way, more significant than his. But my argument is not, nor ever has been, that such artifacts contained in the currently used tables of random numbers are responsible for the significant scores in psychical research. My thesis would be poor indeed if this was all it said. What I have to say is in fact much more fundamental: it is that any attempt to randomize, *of which tables of random numbers and psychical research experiments are*

[1] Thouless, Robert H., *Proc. Soc. Psych. Res.*, **49**, 107 (1951).

both typical examples, will lead all too frequently to the curious results which have been thought in the past by psychical researchers to be evidence of telepathy and whatnot. My suggestion in *Nature* in 1953[1] was simply that if many of the psychical research scores were, as I suspected, merely examples of the 'failures' likely to occur in our attempts to randomize, then similar examples should be found in such sources of randomization as published random numbers themselves. This has now been shown beyond any doubt; but as some of the evidence is not yet well known, I propose to give it here with some new evidence of my own. The mock experiment with which this Appendix begins is an elementary piece of such evidence.

At the time when I wrote my *Nature* paper I had done a series of matching counts on the Fisher and Yates tables. The results of this experiment are summarized in Table II.

TABLE II

Matching Scores with the Fisher tables

1st Series: Sheets and columns determined by dice.

Run	*Sheet* : *D-set,*		*col* : *T-set.*		*Matching Scores*		
					Straight	Displacement *A*	Displacement *B*
1.	I	2	1	4	+ 15	+ 2	+ 4
2.	V	2	1	5	0	+ 9	+ 9
3.	VI	3	1	5	− 3	− 1	− 2
					+ 12	+ 10	+ 11

[1] Spencer Brown, G., *Nature, Lond.*, **172**, 154 (1953).

3rd Series: T-columns taken in sequence; choice of 2 D-columns determined by random numbers.

1st part

1.	I	1	9	2	+ 16	+ 21	+ 11
2.	I	3	9	4	+ 16	0	+ 4
3.	II	2	9	3	− 3	0	+ 3
4.	II	4	9	5	− 4	− 4	+ 3
5.	III	1	9	2	− 3	+ 9	− 17
6.	III	3	9	4	+ 2	+ 3	− 3
7.	IV	2	10	3	+ 7	− 5	− 3
8.	IV	4	10	5	+ 2	+ 12	− 6
					+ 33	+ 36	− 8

3rd Series

2nd part

1.	I	2	9	3	− 3	+ 16	+ 1
2.	I	4	10	5	0	+ 6	+ 6
3.	II	1	10	2	− 2	− 3	− 8
4.	II	3	10	4	+ 2	+ 4	− 1
5.	III	2	10	3	+ 2	+ 6	− 2
6.	III	4	10	5	− 5	+ 21	− 9
7.	IV	1	10	2	− 6	− 16	− 1
8.	IV	4	10	5	− 7	0	+ 10
					− 19	+ 34	− 4

4th Series: T-columns taken in sequence; choice of 2 D-columns determined by random numbers.

Run	*Sheet* : *D-set,*		*col* : *T-set.*		*Matching Scores*		
					Straight	Displacement *A*	*B*
1st part							
1.	V	1	10	2	− 8	− 4	+ 3
2.	V	3	10	4	− 14	+ 5	− 7
3.	VI	2	10	3	+ 4	− 8	+ 3
4.	VI	4	10	5	+ 3	+ 9	+ 1
					− 15	+ 2	0

2nd part

1.	V	2	10	3	− 8	+ 1	− 5
2.	V	4	9	5	− 4	− 1	− 4
3.	VI	1	9	2	+ 3	+ 7	− 7
4.	VI	3	10	4	+ 12	+ 6	+ 5
					+ 3	+ 13	− 11

SUMMARY

	Matching Scores		
	Straight	Displacement	
1st Series		*A*	*B*
3 runs	+ 12	+ 10	+ 11
3rd Series, 1st part			
8 runs	+ 33	+ 36	− 8
3rd Series, 2nd part			
8 runs	− 19	+ 34	− 4
4th Series, 1st part			
4 runs	− 15	+ 2	0
4th Series, 2nd part			
4 runs	+ 3	+ 13	− 11
Total:	+ 14	+ 95	− 12

$$\text{C.R.}_A = \frac{95}{34{\cdot}5} = 2{\cdot}75 : p < 0{\cdot}007$$

As best of 3 counts, $p_A < 0{\cdot}02$

A column of digits, called D digits, were randomly chosen. Matched against it was a block of 10 columns randomly chosen on the same page. The block was called the T-set. The block from which the D-column was chosen was called the D-set. Straight matching scores were made as follows.

D-column	*T-set*					*Score*
3	33	26	16	80	45	(2)
4	27	07	37	07	51	(0)
5	13	55	38	58	59	(4)
9	57	12	10	14	21	(0)
3	06	18	44	32	53	(2)

Over the whole series, no particular significance was obtained with these scores. But, like Dr Soal, I persisted, and rechecked the whole series for displacement scores. These were made as follows.

						Displacement scores	
D-column			*T-set*			*A*	*B*
3	33	26	16	80	45	—	(1)
4	27	07	36	07	51	(1)	(0)
5	13	55	38	58	59	(1)	(1)
9	57	12	10	14	21	(1)	(0)
3	06	18	44	32	53	(0)	—

It is notable that throughout the series the excess of positive scoring on the *A* Displacement builds up to considerable significance.

A series (the 2nd) of 16 runs is not included here because it was not rechecked by the author and is thought to contain inaccuracies. The unchecked A-scores for this series were + 47, a figure which, if it is reasonably correct, adds considerably to the significance of the whole.

As my third example of non-inductive significance discovered in randomized data I take Mr Oram's interesting experiment with the Babington Smith random numbers. I record Mr Oram's paper, with my reply to it, in full.

AN EXPERIMENT WITH RANDOM NUMBERS[1]

By A. T. Oram

A remarkable statement appeared in an article in *Nature* for 25 July, 1953 (Vol 172, p. 154); the author was Mr G. Spencer Brown and the statement read as follows: '. . . I have evidence, also to be published shortly, that statistically significant results similar to those of psychical research are obtainable simply by making selections in published tables of random numbers as if the tables were themselves the data of a psychical research experiment.'

[1] *Journal of the Society for Psychical Research*, **37**, 369 (1954).

It seems that the evidence referred to has not been published but the author has, nevertheless, made further critical observations on somewhat similar lines from time to time since July 1953, in broadcast talks and in lectures. It has therefore become desirable to have available the detailed results of a reasonably comprehensive test of the 'scores' to be obtained from published random numbers, and with the help of some 55 members of the Society such a test has now been carried out.

The longest readily available series of random numbers is that prepared by M. G. Kendall and B. Babington Smith.[1] It consists of 100,000 random digits, printed in 100 blocks of 1,000, prepared and tested for randomness by the authors in rows, across the pages. Their test relating to the frequencies of occurrence of the digits applies, however, equally well in whatever order they may be taken. As each block of 1,000 is set out in 20 pairs of columns, each column containing 25 digits, and this layout lends itself to the type of experiment referred to by Spencer Brown, the digits in the first column of a pair have been taken to represent the guesses in a card-guessing experiment, and those in the second column the actual values of the cards. The 'scores' have been recorded on two separate bases, in Series A on a comparison between the 'guess' and the 'target', and in Series B on a comparison between the 'guess' and the 'target' next to come, being the equivalent of plus-one displacements in ESP experiments. The whole book has been used in each case, so that there has been no selection and as far as possible no personal factors have been allowed to affect the results.

Recording the scores from the printed pages of the book is comparatively simple, but it becomes very tedious if continued for long periods. For this reason the task involved in checking the scores for 50,000 pairs in Series A and a further 48,000 pairs in Series B (one is lost in each column of 25 in scoring for Series B) was divided into 25 sections, each covering four blocks of 1,000 digits. Fifty members of the Society were asked, by post, if they would each complete two record sheets (A and B) prepared for the purpose. In this way the ground was covered twice, so as to provide a check on the results by comparing the corresponding record sheets and by checking and correcting any scores that were recorded differently by the two persons.

The results are fully in agreement with the generally accepted theory of probability, and they show no tendency to follow the

[1] *Tables of Random Sampling Numbers*, Cambridge University Press, 1939.

pattern of scoring that arises in successful ESP experiments, except to the extent that there is a slight decline effect, shown in tables below. The overall scores for the whole experiment are as follows:

	Series A	*Series B* (With plus-one displacement)
1. Number of Tests	50,000	48,000
2. Mean Chance Expectation	5,000	4,800
3. Actual Scores	5,029	4,735
4. Deviations from Mean Chance Expectation	+ 29	− 65
5. Standard Deviations	67	66

A more detailed statement of the scoring is set out below. Mean Chance Expectation for the score in one column of 25 (or, in Series B, 24) is 2·5 (or 2·4), the actual scores ranging from 0 to 10; the frequency of occurrence of the various possible scores for the 2,000 columns in each series has been compared with the mean chance expectations for those frequencies and it is shown below that the actual frequencies obtained differ but little, and certainly not significantly, from the calculated or theoretical expectations.

	Series A		*Series B*	
Score	Actual Frequency	Mean Chance Expectation	Actual Frequency	Mean Chance Expectation
0	141	143·6	163	159·5
1	380	398·8	456	425·4
2	555	531·8	523	543·6
3	463	453·0	430	442·9
4	264	276·8	264	258·4
5	120	129·2	120	114·8
6	53	47·9	27	40·4
7	19	14·4	12	11·6
8	3	3·6	4	2·7
9	1		1	
10	1	0·9	0	0·7
over 10	0		0	
	2,000	2,000·0	2,000	2,000·0
Value of chi-squared	5·3		8·5	
Probability for such a value	0·6		0·3	

NOTE. In calculating chi-squared the values for scores of 7 and over have been pooled in each case.

A further test of the figures is obtained by noting the cumulative scores at a number of stages from the beginning to the end of each series. This experiment, for instance, might have been designed so as to use only the first half of the table or some other portion of it, and it is of interest to know at least whether any sequence of columns, starting from the beginning, gives a significant deviation from a chance score. For this purpose the scores in each series have been recorded, in the first place in groups of 500 tests (480 for Series B) up to 10,000 (9,600), and then in groups of 1,000 (960) to the end, giving the score at 60 points in each series. At no point in either series does the cumulative score even approach a significant deviation from mean chance expectation. The highest values of *t* (or the 'critical ratio') are as follows:

Highest Values of t *after* n *Tests*

	n	*t*
Series A	44,000	1·03 (+)
Series B	12,480	1·19 (+)
	13,440	1·03 (+)
	34,560	1·18 (−)

At each of the other 115 points in the cumulative scoring of the two series the deviation from mean chance expectation is less than one standard deviation.

In some experiments in card-guessing there has apparently been found a significant decline effect in the scoring as an individual continues to guess the cards. In some cases it is claimed that there are decline effects within columns or sheets, as well as within the whole series, while different parts of columns or sheets may show results that differ significantly between themselves. The question as to whether there are substantial overall declines within either of the two series of the present experiment has been sufficiently covered in the paragraph above, where it has been shown that the cumulative scores have kept within close limits of the mean chance expectations throughout, but declines and other special effects within the record sheets remain to be dealt with.

TEST FOR DECLINE EFFECTS

Series A

Sub-Groups	Columns 1	2	3	4	*Totals*	*Mean Chance Expectations*	*Deviations*	*Standard Dev'ns*
1	259	246	271	265	1,041	1,000	+ 41	30
2	269	251	261	224	1,005		+ 5	
3	257	263	257	243	1,020		+ 20	
4	231	264	251	226	972		− 28	
5	251	251	246	243	991		− 9	
Totals	1,267	1,275	1,286	1,201	5,029	5,000	+ 29	67
M.C.E.	1,250				5,000			
Dev'ns	+ 17	+ 25	+ 36	− 49	+ 29			
S.D.	33·5				67			

Series B

1	251	283	225	240	999	960	+ 39	29·4
2	246	247	249	229	971		+ 11	
3	226	219	237	220	902		− 58	
4	242	262	203	218	925		− 35	
5	253	233	222	230	938		− 22	
Totals	1,218	1,244	1,136	1,137	4,735	4,800	− 65	66
M.C.E.	1,200				4,800			
Dev'ns	+ 18	+ 44	− 64	− 63	− 65			
S.D.	32·9				66			

Series A and B together

1	510	529	496	505	2,040	1,960	+ 80	42
2	515	498	510	453	1,976		+ 16	
3	483	482	494	463	1,922		− 38	
4	473	526	454	444	1,897		− 63	
5	504	484	468	473	1,929		− 31	
Totals	2,485	2,519	2,422	2,338	9,764	9,800	− 36	94
M.C.E.	2,450				9,800			
Dev'ns	+ 35	+ 69	− 28	− 112	− 36			
S.D.	47				94			

CHI-SQUARED TESTS ON THE DATA SET OUT ABOVE

	Degrees of Freedom		*Series A*	*Series B*	*Series A and B Together*
The 20 figures in the body of each table	19	χ^2	13·6	26·6	23·4
		P	0·8	0·1	0·2
The Column totals	3	χ^2	3·5	7·8	7·8
		P	0·3	0·05	0·05
The Row totals	4	χ^2	2·8	6·2	6·5
		P	0·6	0·19	0·16

The data have therefore been examined for aggregate decline (or other) effects (i) within, and (ii) between, each of the four record sheet columns. Three tables are set out above, showing the results for Series A and B, separately and together.

Each column in the tables summarises the scores in the corresponding column of all the relevant record sheets, and each row shows the scores for one-fifth (i.e., 4 entries out of 20) of the columns, in the corresponding position, from top to bottom. For instance, the figure 263 in the first table is the aggregate of the 9th, 10th, 11th, and 12th scores (for the 17th and 18th to the 23rd and 24th columns of the printed random numbers) in the second column of each of the 25 record sheets in Series A.

It will be observed that the value of chi-squared for the column totals of Series B is rather high, and that $P = 0{\cdot}05$. This chi-squared test is, however, not altogether fair, for two reasons. In the first place it does not take account of the progression of the signs—the very essence of a decline effect—which in this case must be taken as increasing the weight of evidence for a decline effect. In the second place, on the other hand, we must not lose sight of the fact that four independent tests have been applied for rows and columns of the first and second tables, and that if we take sufficient independent tests we shall almost certainly come across results which, considered alone, would appear to be significant.

The order of the 100 columns in the record sheets does not correspond to that of the 100 blocks of 1,000 digits in the printed book of random numbers, because the unbacked sheets in the book (blocks 1 and 2, 11 and 12, etc.) were sent in pairs with the first five record sheets and the backed sheets (blocks 3, 4, 5, 6 7, 8, 9, 10 13, 14, 15, 16, etc.) consecutively, one with each of the remaining twenty. This order was followed because of the layout of the book, although other

orders could have been adopted, but once we depart from that in the book there are a great many orders which are at least theoretically possible, and some of these would presumably produce unusual results on comparing the aggregate column scores. It so happens that if the data are re-cast in the 'book order' the apparent declines between columns disappear. The details are as follows:

	COLUMNS				TOTALS
	1	2	3	4	
Series A	1,243	1,243	1,310	1,233	5,029
Series B	1,185	1,219	1,169	1,162	4,735

The values for chi-squared are 3 and 1·6 and the values of P 0·4 and 0·6, with three degrees of freedom.

The analysis of the column data into five rows is not affected by the order of the columns.

It is unfortunate that this investigation of the possibility of finding decline effects within the data was only started after the completion of the first draft of this report on the rest of the data, and there has not been time to look further into the background of the results obtained in this section. It is hoped to do this, although there seems to be little reason to suspect that there is any substance in these apparent but hardly significant declines.

The Possibility of Undetected Errors

It will be noted that although all record sheets have been prepared, independently, in duplicate and checked one against the other, there remains the possibility that, here and there, each person may have made the same error. The most likely frequency of such undetected errors can be estimated for each pair of record sheets, taking into account the actual number of observed errors in each, and making an allowance for the fact that of all errors only those of a similar nature, occurring together (e.g., a score that is too small by 1, the commonest type of error), can give rise to an undetected error. From such estimates it is clear that the undetected errors are of but little importance.

In Series A, however, the three sets of record sheets (i.e., 3 out of 25) with the highest numbers of detected errors were checked throughout, in order to eliminate the most likely sources of undetected errors in that series; the result was to add only 3 to the total score. Further

checking on these lines would presumably produce relatively smaller adjustments, because the numbers of detected errors on the remaining record sheets are smaller, indicating a greater accuracy in their preparation.

In Series B only one set was checked right through, no previously undetected errors being found in that set. As, however, the total score in Series B is below mean chance expectation, and as it is demonstrated below that the effect of any undetected errors would almost certainly be to add to the score, they would merely reduce, slightly, the existing negative deviation. For this reason it has not been considered worth while at this stage to put any further check on the scores in Series B.

Brief Details as to Errors made in the Scoring

Out of 100 record sheets, each with 80 separate column-scores, there were three that contained so many errors that it appeared that special circumstances must have applied. In several other cases scores were wrong for an obvious reason (e.g., a transposition) that did not amount to a mere error in scoring. In the details that follow the three record sheets and the special errors have been left out and the apparently genuine errors in scoring have been analysed.

The errors consisted predominantly of those understating the true score by 1. Out of a total of 371 errors only 16 involved an overstatement, and of these 16, 4 were by one person, on one record sheet in Series B (2 entered, when it should have been 1, four times). A further 3 of the 16 were by another person, the balance of 9 being spread between 9 participants.

The types of error may be summarised as follows:

True Score	*Series A*		*Series B*	
Understated by 1	147	86%	168	84%
" 2	18	11%	19	9%
" 3	0		2	
" 4	1		0	
		3%		7%
Overstated by 1	2		10	
" 2	2		2	
	170	100%	201	100%

The incidence of errors at different levels of true column-scores has been calculated; details are set out below.

Error Rates (a) Per 1,000 *Occurrences of each True Column-Score and (b) Per* 1,000 *Individual Scores*

True Column-Score	*Series A*		*Series B*	
	(*a*)	(*b*)	(*a*)	(*b*)
0	0	—	6	—
1	10	10	32	32
2	32	16	47	23
3	46	15	54	18
4	105	26	83	21
5	96	19	113	23
6 and over	81	13	92	14
All scores	44	18	51	23

From personal experience in scoring the figures it appears that as the score per column becomes larger one is confronted with (i) more to remember, per column, but, (ii) on the average shorter intervals between reviving the memory, by adding to the score. In addition to these memory problems there is straightforward missing, and there seems to be no trend in columns (*b*) above to suggest that there is any powerful factor at work beyond this simple missing of items.

The average error rate represented by the 371 errors referred to above, taking into account the fact that the scoring was done twice, but that three sheets were left out for the data in this section on errors, works out at 371/18,932 or just below 2 per cent.

The foregoing details relating to errors have been set out in case they may be of some value as a guide to the probable error patterns in any other experiments; for the present exercise all these errors have been corrected.

Further Details Relating to Procedure

The following notes on procedure are in amplification of the outline in the opening paragraphs.

(i) The names of the participants were selected by Miss E. M. Horsell, in consultation with Mrs K. M. Goldney.

(ii) Participants were not informed of the object of the experiment.

(iii) A letter, signed by Mrs Goldney and by A.T.O., together with a sheet or sheets cut from a book of the random numbers, and two record sheets, were sent to each participant.

(iv) In sending out the sets of digits, the addressed envelopes were sorted into London and Provincial groups and the same set of figures was as far as possible sent to two persons, one from each group, or at least living at a distance from each other.

(v) There had been no prior enquiry and it was to be expected that some of the recipients of the letter and other documents would, for various reasons, find it inconvenient or impossible to help. In fact the response was excellent; without any 'reminders' the results were as follows:

	Sets of Record Sheets
Returned satisfactorily completed	42
Returned completed, but on a wrong basis	1
Returned not completed	2
Not returned	5
	50

(vi) Further members were asked to help with the 8 sets (the 1, 2, and 5 above) that remained to be scored, and all in this second circulation were 'Returned Satisfactorily Completed'.

(vii) All checking and correcting was done by A.T.O.

(viii) The original data can be seen, by arrangement, at the offices of the Society.

(ix) In the letter that was sent out it was explained that the scoring and the summary should not take more than about an hour. Several participants wrote to the effect that it took considerably longer. The original estimate had been based on two short timed runs, but subsequently, in view of these comments, several sheets were completed by A.T.O. in connection with the checking and each section of the task was timed accurately. In each case they were completed within the rate of one hour per set and with no more than average errors. In fairness, however, to those members who made this comment, the estimate of one hour did not make any allowance for checking the scores. One set of forms was returned with no errors and six had only one error, out of the first circulation of 50 sets; the additional trouble taken by the members who completed them has made considerably easier the task of checking.

Acknowledgements and Conclusions

I am most grateful to Mrs Goldney for her help in the planning stages of this experiment. It could not, however, have been carried

out without the kind help of so many members of the Society. It has been quite impractical to write and thank each one, but I would ask them to accept this expression of appreciation of all that they have done. The work was dull, but now and then we need, particularly in view of certain criticisms, to have a simple factual reminder that our statistical methods, when tried out in the absence of any possible influence from psi phenomena, do give reliable 'chance' results. Here we have 98,000 trials and except for the possibility of a slight decline effect, which is probably not statistically significant, they show chance results in each way that they have been tested. All the figures have been included and all tests made upon the results have been reported here.

Correspondence [1]

AN EXPERIMENT WITH RANDOM NUMBERS

SIR,—I am delighted to see that someone else has thought it worth while to match some random numbers. It is also very proper that the new test has been made on the tables compiled by M. G. Kendall and B. Babington Smith. Some of my tests were made on the digits prepared by Sir R. A. Fisher, and the rest on those prepared by Mr Tippett, so when these are published we shall have matching data from the three most widely used random number tables.

I suspect that we owe Mr Oram's work (*Jnl. S.P.R.*, 1954, **37**, 369) partly to the delay in the publication of details of my own records. If so I should like to be the first to congratulate Mr Oram on his well-designed experiment, and on the rapidity with which he has made his data publicly available. It is unfortunate that, like Coover, he appears to have missed one of the main features of his results.

Writing of the Duke PK test data, J. G. Pratt (*J. Parapsychol.*, 1949, **13**, 11) said that 'Because of the declines both *down* and *across* the page, the greatest difference in scoring was expected on the upper-left to lower-right diagonal. Accordingly, this difference was the one always evaluated in the QD analysis'. On p. 12 in the same paper he gives the ranking order of the four quarters of the page in the recorded PK experiments, from highest to lowest scoring sections, as follows.

1. Upper left.
2. Lower left.
3. Upper right.
4. Lower right.

[1] *J. Soc. Psych. Res.*, **38**, 38 (1955).

In the matching scores given by Mr Oram it is impossible to know the exact dividing line across the page, but by missing out rank 3 we can make an approximation. In the combined scores of Series A and B together we find deviations from the mean expectation in the four divisions as follows.

1. Upper left, + 92
2. Lower left, + 27
3. Upper right, + 4
4. Lower right, −121

Thus the ranking of the quarters by scoring tendencies is exactly similar to that which was found in the Duke University PK data. This in itself is remarkable, since the probability of its occurring by chance is $1/P_4^4 = 1/24$. But the result is even more remarkable if we follow the practice established for the PK data and test the significance of the difference between the scores in the upper left and lower right divisions. Here there is a drop in score of 213 in the expected direction, which gives a critical ratio of more than 3·58. The value of p for this critical ratio, using a single tail, is less than 1/5,000. A QD test of series B alone gives an even more significant result. These results are in fact rather better (or worse) than any which I obtained from my own counts. What is more we know that much smaller critical ratios resulting from QD comparisons have been cited as good evidence for PK (e.g. McConnell, Utrecht Conference Report No. 7). Mr Oram's results therefore hardly support his contention that 'our statistical methods, when tried out in the absence of any possible influence from psi phenomena, do give reliable "chance" results'.

Apart from its confirmation of my original suggestion, I should like to draw one further moral from Mr Oram's interesting experiment. This moral lies in the striking illustration of how, without cheating, an experimenter can contrive, by the way he presents his data, to maximize or minimize the significance of any particular trend according to whether it suits his purpose or not. Mr Oram, who presumably did not wish to find significant declines, used a method of estimation which gave their significance at $p = 1/20$, with the qualification that this might be a slight underestimation. Had he been hoping to find the declines, it seems hardly likely that he would have missed the legitimate opportunity of increasing his estimate of their significance more than 250 times.

Finally, a word about forthcoming explanations of Mr Oram's

data. Mr Oram himself denies the presence of 'any possible influence from psi phenomena', but others may be less strong-minded. Those who have an emotional need of or are otherwise committed to an ESP–PK hypothesis might want to say that the columns giving rise to significant declines were selected in the right order by ESP; or, alternatively, that Mr Babington Smith's original randomizing machine was influenced during the course of Mr Oram's experiment by retroactive PK. Such hypotheses are not controverted by the data. Indeed, they cover the findings completely. But explanations of this kind are becoming increasingly complicated; moreover, they stand in the way of the simpler alternatives which have now become plausible.

G. SPENCER BROWN

Christ Church,
Oxford.

As Mr Fraser Nicol has since pointed out,[1] this remarkable QD decline is more significant than *any* such decline in the psychokinesis experiments done by Professor Rhine and his confederates. But in spite of its obvious interest in relation to the work he has done, Professor Rhine has so far given it no mention in his *Journal of Parapsychology*.

The gradual decline of a particular scoring tendency throughout an experiment can be explained by a slowly changing bias in the randomizer (which might be the experimental set-up itself). (E.g. the mock experiment on the Fisher Tables.) But the repeated decline in scoring *within an experimental unit*, which is so common in psychical research and which has now been found in randomized data which were prepared with no thought of psychical research in mind, is not so easy to explain. It would seem that the randomizer tends to a bias which changes periodically, and that the period is somehow pulled into step with an arbitrary experimental

[1] Fraser Nicol, J., *J. Soc. Psych. Res.*, **38**, 80 (1955).

unit. That the tendency is something inherent in the randomizing set-up and not a result of psychokinesis or other occult phenomena now lies, I think, beyond reasonable doubt. We seem to lack only a detailed explanation.

The last example I shall give of a significant bias occurring in randomized data is from the standard series used by Dr Soal in the guessing experiments with Mrs Stewart. Here, as he reported,[1] there was a highly significant deficiency of *ABA* patterns. He omits to mention the actual figure in his published works, and has so far declined my written request for the information. I am thus able to go only by hearsay, but I am told that the significance of the deficiency is greater than the very large significance of his most significant result with Mrs Stewart. He protests that the absence of *ABA* patterns does not affect the validity of the significance of her guesses. But this is not the point. The fact is that here is a further unexplained piece of significance, of a very high order of magnitude, in randomized data.

The guessing scores, *because of their high significance*, are said to be very good evidence for telepathy; but results of *equal or greater significance* in the randomized data are *glossed over*. Is this because *they* are not interpretable in terms of telepathy? Whatever their explanation (and it is only Dr Soal's *conjecture* that the gentleman who prepared the digits did not properly act upon his instructions) it cannot be denied that they form another link in the chain of evidence substantiating my thesis that, whenever one tries to randomize, significant biases are bound to occur and can build up to a large significance before they are noticed. Why say that when they occur in certain

[1] Soal, S. G. and Bateman, F., *Modern Experiments in Telepathy* (London, 1954).

circumstances they are evidence of marvellous telepathy, but when they occur in other circumstances they are just statistical artifacts or mistakes on the part of the operators? It is much simpler to suppose that they are in each case the same thing.

It may occur to us to ask why, if the probability set-up can play such tricks, is psychical research the chief sufferer? How is it that other sciences are not equally hard hit? The answer is simple if the psychical research experiment is, as I suggest it is, a degenerate experiment. For any valid experiment designed to look for a tendency which does not exist must degenerate by its own experimental logic into a pure probability experiment; this, indeed, is the meaning of the null hypothesis. Furthermore, there has never been any practical doubt about the difficulty of obtaining a null result of any kind in a long probability experiment; and I have given the theoretical reasons for this in Chapters XII and XIV. Now, an ordinary scientific experiment in which no real (i.e. stable) tendency is discoverable is comparatively rare; in any natural set-up there is likely to be *some* sort of inherent bias *greater than* any bias, inherent or transitory, in the randomizing agent used. And if this is so, the inherent bias of the natural set-up will rapidly show itself in an increasingly significant and *demonstrably repeatable* deviation in matching scores. In the rare cases where the inherent bias of the natural set-up is less than or equal to that of the randomizer, a null hypothesis will be indicated in the following manner. If the experiment is short, the matching scores should, with luck, be insignificant or only slightly significant; alternatively, a watchful experimenter using a feed-back randomizer should get an insignificant score with a longer experiment. But otherwise, if the experiment is long, a statistically significant matching score is likely

in any case, and this will be distinguishable from significance due to a bias in the natural set-up only by its failure to repeat with a new randomizer.

The common appearance of statistical oddities in psychical research I take to be due, then, to the comparative frequency here of a natural bias failing to swamp the bias of the randomizer. This is quite plausible in the light of the fact that psychical research is perhaps the only present-day science which has looked for something (not already known to exist) for sixty years and failed to find it; and if it happened that what it was looking for did not exist, we should have in effect sixty years of pure probability experiments which there is no reason to suppose should have fared, in terms of significance, better than the best (and the worst) of all the pure probability experiments down the ages. It would thus be its remarkable additions to our experimental picture of pure probability for which we owe the most thanks to modern psychical research.

COMMENTARY

Chapter II

We can speculate about why our universe is what it is until we are blue in the face; the mystery remains. I shall give later a logical reason for its prevalence.

I am indebted to Lord Weymouth for the idea of a solid universe. I think the nothing-universe is my own idea. It is of course very wrong for us to try to describe the whole universe, and can only end in disappointment. The descriptions should therefore be taken no more seriously than a piece of poetic licence or a musical cadenza. The people who do it professionally and call themselves cosmologists are the modern fable-makers. Their task is to poeticize astronomy and stellar physics.

We seem to be able to answer the question, When is a thing conscious? by simple analysis. A thing is conscious only when it is conscious *of something*; but it is not conscious of something when its behaviour can be interpreted as showing merely that it reacts either to or from that something. This is simple feeling or sentience. A thing may be said to be *conscious* only if its behaviour can be interpreted as having constructed an imitation or *model* of an object, its subsequent reactions being determined at least partly with reference to the model rather than the object. For a thing to become self-conscious or introspective, it needs only to be able to model its own modelling capacity. There is no reason why machines should not be made to do all of these things.

Chapter III

We must note an oversimplification when we say that argument (2) reflects inductive procedure. More accurately, the argument runs

If all swans so far are white

Then all further swans will be white.

The form of the extrapolation at once shows itself. Generalizing, we obtain the Law of Uniformity.

If all somethings so far are so-and-so

Then all further somethings are so-and-so too.

But since this is plainly false there can be no such law.

Chapter IV

I am not suggesting that either the scientific or the mystical is the *better* way of talking. It depends on the purpose at hand. If I wish to navigate to Nova Scotia, I use the hypothesis of a spherical earth to choose a great circle track. But if I wish merely to mark out a lawn tennis court or even to sail across the English Channel, I use the flat earth or plane hypothesis. Hence the (usually misspelt) expression 'plane sailing'.

It is misleading to consider either of these hypotheses as absolutely right or absolutely wrong. Plane sailing is the more convenient kind of sailing for short distances, but its complexity increases and surpasses that of spherical sailing as the distances become greater.

Similarly the scientific way of talking may be very unhelpful if we wish to teach a person self-control. Here it may be useful for him to believe that he is more stable than he seems in order that eventually he may become so. Eastern philosophy, which uses a mystical rather than a scientific language, sets great store on individual self-control. It is

powerless in the face of objective phenomena, over which its control is minimal. The scientist, by assuming the constancy of objective phenomena, controls them rather than himself. Why exactly the assumption of constancy in anything should help us control it I am not sure; if we only imagine we control things, then explanation by definition is easy; but I think we could say more than that.

By showing the problem of induction to be interpretable in terms of recognition, we have not solved it. What we have done is merely to dissolve it by producing a language in which it can't occur. Whenever we get into difficulties, we can always invent a language to get us out of them. But each time we do this, we also create difficulties where none were before. The problem now centres around the process of recognition. This sort of thing is bound to happen, for we know that it is impossible to produce a language suitable to describe everything at once. There is no end to the process of description; the better we know something, the more ways we can describe it.

Chapter VII

My knowledge of the intensity of Mr X's desire to meet my other visitor is sometimes itself described in terms of probability; as, for example, in the works of Sir Harold Jeffreys and Professor Carnap. I side with Professor Braithwaite in deprecating this use of the term 'probability' in science, preferring instead to speak of 'reasonableness'. The fact that we allow our judgment of the reasonableness of an hypothesis to influence the way we calculate the significance of the observations which support it is another matter.

Chapter VIII

We have three relevant concepts: order, randomness, and chaos. The series '1111111111' is ordered; the series '1000110101' is random; and the series '1oxTπ6♂g★@' is chaotic.

An ordered series is a series with an obvious rule for predicting what comes next from what has gone before. A random series is a series in which we can predict what comes next only within certain limits; for example, in our second series we can predict that the next figure will not be a 2, but we cannot predict whether it will be a 1 or a 0. In a chaotic series we are unable to say anything about the next event. We see that random series, if they go on for ever, resemble chaotic series in that, by taking larger and larger compounds, there is no end to the possibilities of different events. They differ from chaotic series in the important respect of showing certain types of unavoidable order. For example, if single events are to appear disordered, there is bound to be order among the compounds; if the digits of a binary random series are to appear disordered, then the compounds of, say, ten such digits will appear ordered when considered as combinations. Combinations of 5 ones and 5 noughts will be much commoner than combinations of 9 ones and 1 nought. Conversely, if we make these latter compounds equally probable, then we could successfully predict single events.

The laws of chance are not really about chance; they are about the non-chance implications of chance events. J. M. Robertson in his brilliant letter [1] on chance, though he allows his indignation to run away with him and misstates his case over the question of long runs, was, if I am not mistaken, the

[1] Robertson, John M., *Letters on Reasoning* (London, 1902).

only philosopher in the last hundred years to see clearly that there were no laws of chance.

Chapter X

When I say in my proof of the matching theorem that it selects the instances of formula (7) in which both σ and $\frac{w^2\sigma^2}{w-1}$ are integers, my remark is unnecessarily strong. It was pointed out to me by Mr Antony Hoare that σ may take fractional values. For example, the binary series consisting of only one term happens to have an exactly critical bias of $\frac{1}{2}$.

Rereading this chapter prompts me to consider in what way mathematical propositions are elucidatory. What use is it to say that one expression is equal to another different-looking expression? I think the answer lies in their different forms stressing different aspects of what we want to consider. Even when Burns says 'A man's a man for a' that', the proposition is elucidatory only in so far as 'man' stresses something slightly different in its successive appearances. If we say that a screwdriver is an instrument containing a piece of metal fashioned into a narrow blade, we may be saying no more than that a screwdriver is a screwdriver, but the former may be more useful if we happen to be without one. If work is held up for want of a screwdriver, it is no help to say a screwdriver is a screwdriver. If we say a screwdriver is a bladed instrument, we can then begin to consider whether we have any other sort of bladed instrument which would do its job. The formula might now suggest we use a knife or a chisel. In other words, the different ways of saying the same thing emphasize different aspects of it. When we use a mathematical formula containing the equality sign, we do so in order to

couple together certain ideas we wish to *emphasize* on either side of it.

Chapter XII

In deprecating the uses to which Bernoulli's Theorem is put I do not wish to detract from the original brilliance and insight of James Bernoulli himself. Faced with a problem in mathematical logic he produces the correct solution in terms of the calculus he uses. That his solution is pertinent to some problems in probability I do not question. It is the uncritical application of his methods that leads to trouble.

We do not castigate Newton for not having noticed how oddly metaphysical was his concept of simultaneity. Nor do we throw away Newtonian mechanics for all purposes because Einstein invented mechanics which were for some purposes less trouble. Similarly, we do not throw away the whole of Bernoullian theory because of his oddly metaphysical concept of true probability upon which it rests. No system-builder can anticipate every future need, and the time must come for each of us when, in some activities at least, his system is rejected as misleading. In our vanity we hope it will be later rather than sooner, but this is often a matter of luck.

Chapter XIII

I am not the first to have deprecated the uncritical application of Bernoulli's Theorem. Keynes [1] listed similar objections which led him to conclude that the Theorem was strictly applicable in only a minority of cases.

Keynes's warning seems to have been ignored by subsequent practitioners, and he himself did not appear to take it very seriously. For, in the same chapter, he went on to list the

[1] Keynes, John Maynard, *A Treatise on Probability* (London, 1921).

various attempts which had been made to demonstrate empirically the laws of chance. In spite of the fact that nearly all of the results he cites fall very far outside of what these laws lead us to expect, he fails to notice the interesting connexion between such results and his earlier criticisms of Bernoulli's Theorem. Odd results with dice, roulettes, etc. are relevant, he says, 'not to the theory or philosophy of Chance, but to the material shapes of the tools of the experiment'. But if no series of chance events is relevant to probability theory, then how are we to suppose that probability theory is relevant to chance events? That Keynes's remark, where classical theory is concerned, is largely true I do not dispute. But what Keynes himself and those who came after him failed to see was the full seriousness of his admission.

In making corrections from the calculations of classical theory we must keep in mind two interacting processes. The first I shall call the Law of Bias: that no series may be excessively stretched. Following this comes the Law of Coincidence: that we must multiply the probabilities of coincidences by a factor determined by how strictly we apply the Law of Bias. The stricter the Law of Bias, the commoner the coincidences. But a coincidence is merely an example of stretch which looks like bias. And it will, by the Law of Bias, be classed as bias. When subsequently it does not repeat itself, it constitutes an impediment to the strict application of the Law of Bias. The set-up is therefore self-adjusting, and hunting may occur around an optimum strictness in the criterion for converting stretch into bias.

An interesting practical fallacy has been to assume that if a series is not rejected by one test for randomness, it is less likely to be rejected by another. And, the argument

continues, if it remains unrejected by several independent tests, then it is even less likely to be rejected by a further test. A moment's reflection shows this argument, based upon analogy with other inductive arguments, to be totally fallacious. It applies in practice only because the common tests for randomness are in fact *not independent*. For example, the four separate tests used by Professor Kendall and Mr Babington Smith were not independent in that they were all sensitive to elementary bias. If tests for randomness are really independent, then it is obvious that the more such tests a series passes, the less likely it is to pass the next one; for the more ways we disallow it to arrange itself, the more likely we are to find that the series, to which only a finite number of arrangements are possible, has arranged itself in one of the forbidden ways.

In practice, looking at series, we cannot distinguish always between the aperiodic and the incomplete period. We tend therefore to class them both as random. But perfect aperiodicity, if it occurs at all in natural series, is very rare. We should expect, therefore, the incidence of series we could call random to fall as we increase their length. We observe here an aspect of the diminishing field described in Chapter XV.

Chapter XIV

Mr Babington Smith is probably as familiar as anyone in this country with the making of random sampling numbers; and I can only record my indebtedness to the many discussions I have had with him.

Chapter XV

He has drawn my attention to a particularly virulent cause of diminishing field contraction. It is in the multiplication of

significance-probabilities by the number of different tests we have made on the data. A case in point is where, in the Appendix on Practice, I multiplied the *p*-value of the 'A' scores by 3 because I made two other tests on the same data.

As a rough practical rule a significance-probability is multiplied by the number of questions we have asked. Moreover, it is of no matter whether they are asked before or after the question which gives us the relevant significant answer. This procedure is not strictly correct, because the number of possible questions is such as could make the probabilities of all but the least probable answers greater than 1. But it can be put right by a relatively minor mathematical adjustment, and when we have made the adjustment what emerges is that, by asking enough questions, any result can be reduced to total insignificance by repeated multiplication of its *p*-value until it becomes unity.

We see evidence of attempts to stem the insidious contraction of the boundaries of the diminishing field (beyond which lies the rubble-heap called the past) by severely rationing the questions which may be asked. Unrepeatable results are 'saved' by disallowing further scrutiny. But this is no way to do science. Science is a significance game: one player tries to reduce significance to insignificance by asking more questions, while another seeks to counter his activities by doing more experiments. The scientist, like the chess-enthusiast, often plays both sides himself.

Repetitions of scientific results serve two purposes. First, they inhibit alternative questions which would tend to reduce their significance; and secondly, each successful repetition tends to increase the significance which such questions might reduce. We thus have a race between the questions and the

results. The *valid* results are the ones which always beat the questions. The constant and somewhat boring repetition of famous experiments in schools and universities is justifiable in terms of this significance race.

It is now clear that we can make no practical distinction between certain cases of phenomena which are controllable or demonstrably repeatable only very rarely and phenomena which are entirely fortuitous. The significance-boosting capacity of either of these types of occurrence is on an average so small that they must, in the face of repeated questioning, dwindle into anecdote.

Appendix on Miracles

We are told that the Second Law of Thermodynamics, being a statistical law, might go wrong; that a kettle, put on the fire, has a chance (very small) of freezing. But, as Eddington pointed out,[1] we should never believe such a thing. Application of the Law of Bias would always give us an alternative hypothesis.

If the kettle did freeze when we put it on the fire we might regard this as a miracle. It would also be a mystery. Mysteries are solved only by recurrence. Any event which can never (or only very rarely) be made to recur is a mystery. Thus the lumping together of all events at once automatically creates a mystery. This is the mystery of the universe, and cannot logically be solved.

The idea that two chance machines might get periodically in and out of step, thus producing high and low scores alternately in the matching series, was originally Mr Babington Smith's. My own development of it has been to

[1] Eddington, Sir Arthur, *New Pathways in Science* (Cambridge, 1935).

assume that all randomizers are fundamentally similar, but that those whose series come from matching scores might be more reliable because of their greater instability. But this does not mean that they may not give startling results if an information-block occurs somewhere.

One of the functions of superstition is to encourage an information-block during a run of luck. We say, 'Nothing has gone wrong so far, *touch wood.*' The ritual enshrined in the phrase 'touch wood' is designed to take our minds off uncontrollable success when thinking of it might break it. The other use of superstition is as a rule of thumb for making decisions between equally attractive or repulsive alternatives. Confronted with a ladder straddling the footpath, we must decide between the possibilities of going under it and getting slopped with paint or going round it and getting run over. We need a simple rule to save us the bother of choosing between these alternatives on their merits. Certain forms of etiquette are adopted for similar reasons; they leave us free to consider more interesting difficulties.

Appendix on Practice

Mr Oram, who has so kindly allowed me to reproduce his paper, has asked me to quote from his comments on my reply to it. He says:[1]

SIR,—I am most grateful to Mr Spencer Brown for letting me see his letter. In answering it, I must first refer to the background of my 'Experiment with Random Numbers'.

. . . I set out to conduct an experiment with a view to applying some tests of randomness to the random numbers in a well-known published set, on the lines indicated by Mr Spencer Brown in the passage quoted from his article in *Nature.*

[1] *J. Soc. Psych. Res.*, **38**, 40 (1955).

Now it is obvious that for such an exercise three essential elements are required:

(i) an experimenter;
(ii) the random numbers to be tested;
(iii) an experimental design.

The design is normally chosen by the experimenter and while there are a number of limiting factors there remain within these limits a great many possible designs from which one must be selected. (I am purposely leaving out of this analysis the model, as it is not involved in the argument that follows.)

When the experiment has been completed the results can be of several kinds; for instance, they can reflect:

(*a*) the nature of the random numbers;
(*b*) the nature of the experimental design;
(*c*) the result of selection by the experimenter (i.e., in effect, modifications to the design at the time of presenting the results).

The influence of (*b*) and (*c*) is normally kept as small as possible because the very object of the exercise is in connection with (*a*). If it should so happen that one is dealing not with a set of random numbers or any other clear-cut object but with, for instance, a strange phenomenon that may well have invaded the field of (*b*), then the position becomes more difficult, but in the case of 'An Experiment with Random Numbers' this complication did not arise.

In fairness to Mr Spencer Brown, the quotation with which I opened my report could have been taken equally well as posing a problem in category (*b*), but as he had so stressed the random nature of the random numbers I took that to be the subject for my experiment.

If we ignore for now the possibility of undetected errors, a matter that was dealt with briefly in my report and is referred to again below, it seems clear that the total scores (5,029 and 4,735) and the score distributions are definitely in category (*a*). They were regarded as the focal point of the experiment. The 'cumulative *t* test' belongs almost wholly to (*a*); the data were taken in book order but the experimental design included decisions as to the position of the 60 points in each series at which *t* should be measured, introducing a slight touch of (*b*) but probably not of any importance.

This was the limit of the results in the first draft of the report and it was at this stage that Dr Wassermann made reference to the experiment in a broadcast talk in August 1954. Subsequently I was pressed by one who had read the draft to extend the tests to cover the possibility of a decline effect, and this was done in September, just in time for the printers.

Some declines were shown by the results, as tabulated in the report, but it was suggested that they were probably not significant; I had considered, briefly, the overall declines and those between column and row totals but had overlooked the very significant effect to which Mr Spencer Brown has now drawn my attention. As the QD figures are not directly available from the tables as published I have now prepared them, in two forms, for Series B, which is the more interesting set from this point of view:

QUARTER DISTRIBUTIONS—SERIES B

Record Sheet Order			*Book Order*		
1254	1188	2442	1225	1217	2442
1208	1085	2293	1179	1114	2293
2462	2273	4735	2404	2331	4735

1254 − 1085 =	169	1225 − 1114 =	111
S.D.	46·48	S.D.	46·48
t	3·64	t	2·39
p	1/7340	p	1/118

Note. The four quarters do not represent a simple time sequence; the position is explained below.

It should be noted that while a QD decline can represent a simple time sequence if there are just two columns on a record sheet, where there are four or more columns it becomes a selection of items depending wholly on the layout. To take the first and last quarters of the task presented in each of the Record Sheets of my experiment we should have to consider the first and fourth columns, and while they show a fairly substantial decline effect in Series B the highest and lowest scores were in the second and third columns.

It seems to be clear from the figures set out above that the effect to which Mr Spencer Brown has referred arises mainly from the field

of (*b*), the nature of the experimental design. Had I not (i) arranged the blocks in a particular *order*, other than the book order, the remarkable decline of 169 would presumably not have arisen, and had I not (ii) arranged to set out the data in *columns* in the way that was adopted, this particular pattern would not have come to light.

In conclusion, whereas the declines obtained are very strange, I am wondering whether they are not of more importance in connection with the problem of the influence of (*b*) phenomena than as part of the study of the random numbers. Although I would like to try to track down the nature of the within-column declines in my Record Sheets (within-block declines in the book) there is some question as to whether it will be worth while to spend the time that this would entail. We need to know what value will come out of it beyond the satisfaction of our curiosity. I have not overlooked the possibility of undetected errors accounting for the effect but it seems most improbable that they could account for more than a small part of the total declines. . . .

A. T. ORAM

Purley, Surrey

INDEX