Chapters 1 and 2 from The Design of Experiments by Ronald A. Fisher. 8th Edition, 1966.

THE DESIGN OF EXPERIMENTS

I

INTRODUCTION

1. The Grounds on which Evidence is Disputed

When any scientific conclusion is supposed to be proved on experimental evidence, critics who still refuse to accept the conclusion are accustomed to take one of two lines of attack. They may claim that the *interpretation* of the experiment is faulty, that the results reported are not in fact those which should have been expected had the conclusion drawn been justified, or that they might equally well have arisen had the conclusion drawn been false. Such criticisms of interpretation are usually treated as falling within the domain of *statistics*. They are often made by professed statisticians against the work of others whom they regard as ignorant of or incompetent in statistical technique; and, since the interpretation of any considerable body of data is likely to involve computations, it is natural enough that questions involving the logical implications of the results of the arithmetical processes employed, should be relegated to the statistician. At least I make no complaint of this convention. The statistician cannot evade the responsibility for understanding the processes he applies or recommends. My immediate point is that the questions involved can be dissociated from all that is strictly technical in the statistician's craft, and, *when so detached*, are questions only of the right use of human reasoning powers, with which all intelligent people, who hope to be intelligible, are equally concerned, and on which the statistician, as such, speaks with no special authority. The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation.

The other type of criticism to which experimental results are exposed is that the experiment itself was ill designed, or, of course, badly executed. If we suppose that the experimenter did what he intended to do, both of these points come down to the question of the *design*, or the *logical structure* of the experiment. This type of criticism is usually made by what I might call a heavyweight *authority*. Prolonged experience, or at least the long possession of a scientific reputation, is almost a pre-requisite for developing successfully this line of attack. Technical details are seldom in evidence. The authoritative assertion "His *controls* are *totally* inadequate" must have temporarily discredited many a promising line of work; and such an authoritarian method of judgment must surely continue, human nature being what it is, so long as theoretical notions of the principles of experimental design are lacking—notions just as clear and explicit as we are accustomed to apply to technical details.

Now the essential point is that the two sorts of criticism I have mentioned are aimed only at different aspects of the same whole, although they are usually delivered by different sorts of people and in very different language. If the design of an experiment is faulty, any method of interpretation which makes it out to be decisive must be faulty too. It is true that there are a great many experimental procedures which are well designed in that they *may* lead to decisive conclusions, but on other occasions may fail to do so; in such cases, if decisive conclusions are in fact drawn when they are unjustified, we may say that the fault is wholly in the interpretation, not in the design. But the fault of interpretation, even in these cases, lies in overlooking the characteristic features of the design which lead to the result being sometimes inconclusive, or conclusive on some questions but not on all. To understand correctly the one aspect of the problem is to understand the other. Statistical procedure and experimental design are only two different aspects of the same whole, and that whole comprises all the logical requirements of the complete process of adding to natural knowledge by experimentation.

2. The Mathematical Attitude towards Induction

In the foregoing paragraphs the subject-matter of this book has been regarded from the point of view of an experimenter, who wishes to carry out his work competently, and having done so wishes to safeguard his results, so far as they are validly established, from ignorant criticism by different sorts of superior persons. I have assumed, as the experimenter always does assume, that it *is* possible to draw valid inferences from the results of experimentation; that it is possible to argue from consequences to causes, from observations to hypotheses; as a statistician would say, from a sample to the population from which the sample was drawn, or, as a logician might put it, from the particular to the general. It is, however, certain that many mathematicians, if pressed on the point, would say that it is not possible rigorously to argue from the particular to the general; that all such arguments must involve some sort of guesswork, which they might admit to be plausible guesswork, but the rationale of which, they would be unwilling, as mathematicians, to discuss. We may at once admit that any inference from the particular to the general must be attended with some degree of uncertainty, but this is not the same as to admit that such inference cannot be absolutely rigorous, for the nature and degree of the uncertainty may itself be capable of rigorous expression. In the theory of probability, as developed in its application to games of chance, we have the classic example proving this possibility. If the gamblers' apparatus are really *true* or unbiased, the probabilities of the different possible events, or combinations of events, can be inferred by a rigorous deductive argument, although the outcome of any particular game is recognised to be uncertain. The mere fact that inductive inferences are uncertain cannot, therefore, be accepted as precluding perfectly rigorous and unequivocal inference.

Naturally, writers on probability have made determined efforts to include the problem of inductive inference within the ambit of the theory of mathematical probability, developed in discussing deductive problems arising in games of chance. To illustrate how much was at one time thought to have been achieved in this way, I may quote a very lucid statement by Augustus de Morgan, published in 1838, in the preface to his essay on probabilities in *The Cabinet Cyclopædia.* At this period confidence in the theory of inverse probability, as it was called, had reached, under the influence of Laplace, its highest point. Boole's criticisms had not yet been made, nor the more decided rejection of the theory by Venn, Chrystal, and later writers. De Morgan is speaking of the advances in the theory which were leading to its wider application to practical problems.

" There was also another circumstance which stood in the way of the first investigators, namely, the not

having considered, or, at least, not having discovered the method of reasoning from the happening of an event to the probability of one or another cause. The questions treated in the third chapter of this work could not therefore be attempted by them. Given an hypothesis presenting the necessity of one or another out of a certain, and not very large, number of consequences, they could determine the chance that any given one or other of those consequences should arrive; but given an event as having happened, and which might have been the consequence of either of several different causes, or explicable by either of several different hypotheses, they could not infer the probability with which the happening of the event should cause the different hypotheses to be viewed. But, just as in natural philosophy the selection of an hypothesis by means of observed facts is always preliminary to any attempt at deductive discovery; so in the application of the notion of probability to the actual affairs of life, the process of reasoning from observed events to their most probable antecedents must go before the direct use of any such antecedent, cause, hypothesis, or whatever it may be correctly termed. These two obstacles, therefore, the mathematical difficulty, and the want of an inverse method, prevented the science from extending its views beyond problems of that simple nature which games of chance present."

Referring to the inverse method, he later adds: " This was first used by the Rev. T. Bayes, and the author, though now almost forgotten, deserves the most honourable remembrance from all who treat the history of this science."

3. The Rejection of Inverse Probability

Whatever may have been true in 1838, it is certainly not true to-day that Thomas Bayes is almost forgotten. That he seems to have been the first man in Europe to have seen the importance of developing an exact and quantitative theory of inductive reasoning; of arguing from observational facts to the theories which might explain them, is surely a sufficient claim to a place in the history of science. But he deserves honourable remembrance for one fact, also, in addition to those mentioned by de Morgan. Having perceived the problem and devised an axiom which, if its truth were granted, would bring inverse inferences within the scope of the theory of mathematical probability, he was sufficiently critical of its validity to try to avoid the axiomatic approach, and, perhaps for the same reason, to withhold his entire treatise from publication until his doubts should have been satisfied. In the event, the work was published after his death by his friend, Price, and we cannot say what views he ultimately held on the subject.

The discrepancy of opinion among historical writers on probability is so great that to mention the subject is unavoidable. It would, however, be out of place here to argue the point in detail. I will only state three considerations which will explain why, in the practical applications of the subject, I shall not assume the truth of Bayes' axiom. Two of these reasons would, I think, be generally admitted, but the first, I can well imagine, might be indignantly repudiated in some quarters. The first is this: The axiom leads to apparent mathematical contradictions. In explaining these contradictions away, advocates of inverse probability seem forced to regard mathematical probability, not as an objective quantity measured by observable frequencies, but as measuring merely psychological tendencies, theorems respecting which are useless for scientific purposes.

My second reason is that it is the nature of an axiom that its truth should be apparent to any rational mind which fully apprehends its meaning. The axiom of Bayes has certainly been fully apprehended by a good many rational minds, including that of its author, without carrying this conviction of necessary truth. This, alone, shows that it cannot be accepted as the axiomatic basis of a rigorous argument.

My third reason is that inverse probability has been only very rarely used in the justification of conclusions from experimental facts, although the theory has been widely taught, and is widespread in the literature of probability. Whatever the reasons are which give experimenters confidence that they can draw valid conclusions from their results, they seem to act just as powerfully whether the experimenter has heard of the theory of inverse probability or not.

4. The Logic of the Laboratory

In fact, in the course of this book, I propose to consider a number of different types of experimentation, with especial reference to their logical structure, and to show that when the appropriate precautions are taken to make this structure complete, entirely valid inferences may be drawn from them, without using the disputed axiom. *If* this can be done, we shall, in the course of studies having directly practical aims, have overcome the theoretical difficulty of inductive inferences.

Inductive inference is the only process known to us by which essentially new knowledge comes into the world. To make clear the authentic conditions of its validity is the kind of contribution to the intellectual development of mankind which we should expect experimental science would ultimately supply. Men have always been capable of some mental processes of the kind we call "learning by experience." Doubtless this experience was often a very imperfect basis, and the reasoning processes used in interpreting it were very insecure; but there must have been in these processes a sort of embryology of knowledge, by which new knowledge was gradually produced. Experimental observations are only experience carefully planned in advance, and designed to form a secure basis of new knowledge; that is, they are systematically related to the body of knowledge already acquired, and the results are deliberately observed, and put on record accurately. As the art of experimentation advances the principles should become clear by virtue of which this planning and designing achieve their purpose.

It is as well to remember in this connection that the principles and method of even *deductive* reasoning were probably unknown for several thousand years after the establishment of prosperous and cultured civilisations. We take a knowledge of these principles for granted, only because geometry is universally taught in schools. The method and material taught is essentially that of Euclid's text-book of the third century B.C., and no one can make any progress in that subject without thoroughly familiarising his mind with the requirements of a precise deductive argument. Assuming the axioms, the body of their logical consequences is built up systematically and without ambiguity. Yet it is certainly something of an accident historically that this particular discipline should have become fashionable in the Greek Universities, and later embodied in the curricula of secondary education. It would be difficult to overstate how much the liberty of human thought has owed to this fortunate circumstance. Since Euclid's time there have been very long periods during which the right of unfettered individual judgment has been successfully denied in legal, moral, and historical questions, but in which it has, none the less, survived, so far as purely deductive reasoning is concerned, within the shelter of apparently harmless mathematical studies.

The liberation of the human intellect must, however, remain incomplete so long as it is free only to work out the consequences of a prescribed body of dogmatic data, and is denied the access to unsuspected truths, which only direct observation can give. The development of experimental science has therefore done much more than to multiply the technical competence of mankind; and if, in these introductory lines, I have seemed to wander far from the immediate purpose of this book, it is only because the two topics with which we shall be concerned, the arts of experimental design and of the valid interpretation of experimental results, in so far as they can be technically perfected, must constitute the core of this claim to the exercise of full intellectual liberty.

The chapters which follow are designed to illustrate the principles which are common to all experimentation, by means of examples chosen for the simplicity with which these principles are brought out. Next, to exhibit the principal designs which have been found successful in that field of experimentation, namely agriculture, in which questions of design have been most thoroughly studied, and to illustrate their applicability to other fields of work. Many of the most useful designs are extremely simple, and these deserve the greatest attention, as showing in what ways, and on what occasions, greater elaboration may be advantageous. The careful reader should be able to satisfy himself not only, in detail, *why* some experiments have a complex structure, but also *how* a complex observational record may be handled with intelligibility and precision.

The subject is a new one, and in many ways the most that the author can hope is to suggest possible lines of attack on the problems with which others are confronted. Progress in recent years has been rapid, and the few sections devoted to the subject in the author's *Statistical Methods for Research Workers*, first published in 1925, have, with each succeeding edition, come to appear more and more inadequate. On purely statistical questions the reader must be referred to that book; on logic, and the analysis of meaning, to *Statistical Methods and Scientific Inference*. The present volume is an attempt to do more thorough justice to the problems of planning and foresight with which the experimenter is confronted.

REFERENCES AND OTHER READING

T. Bayes (1763). An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society, liii. 370.

A. de Morgan (1838). An essay on probabilities and on their application to life contingencies and insurance offices. Preface, vi. Longman & Co.

R. A. Fisher (1930). Inverse probability. Proceedings of the Cambridge Philosophical Society, xxvi. 528-535.

R. A. Fisher (1932). Inverse probability and the use of likelihood. Proceedings of the Cambridge Philosophical Society, xxviii. 257-261.

R. A. Fisher (1935). The logic of inductive inference. Journal Royal Statistical Society, xcviii. 39-54.

R. A. Fisher (1936). Uncertain inference. Proceedings of the American Academy of Arts and Sciences, 71. 245-258.

R. A. Fisher (1925-1963). Statistical methods for research workers. Oliver and Boyd Ltd., Edinburgh.

R. A. Fisher (1956, 1959) Statistical methods and scientific inference. Oliver and Boyd Ltd., Edinburgh.

II

THE PRINCIPLES OF EXPERIMENTATION, ILLUSTRATED BY A PSYCHO-PHYSICAL EXPERIMENT

5. Statement of Experiment

A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested. For this purpose let us first lay down a simple form of experiment with a view to studying its limitations and its characteristics, both those which appear to be essential to the experimental method, when well developed, and those which are not essential but auxiliary.

Our experiment consists in mixing eight cups of tea, four in one way and four in the other, and presenting them to the subject for judgment in a random order. The subject has been told in advance of what the test will consist, namely that she will be asked to taste eight cups, that these shall be four of each kind, and that they shall be presented to her in a random order, that is in an order not determined arbitrarily by human choice, but by the actual manipulation of the physical apparatus used in games of chance, cards, dice, roulettes, etc., or, more expeditiously, from a published collection of random sampling numbers purporting to give the actual results of such manipulation. Her task is to divide the 8 cups into two sets of 4, agreeing, if possible, with the treatments received.

6. Interpretation and its Reasoned Basis

In considering the appropriateness of any proposed experimental design, it is always needful to forecast all possible results of the experiment, and to have decided without ambiguity what interpretation shall be placed upon each one of them. Further, we must know by what argument this interpretation is to be sustained. In the present instance we may argue as follows. There are 70 ways of choosing a group of 4 objects out of 8. This may be demonstrated by an argument familiar to students of "permutations and combinations," namely, that if we were to choose the 4 objects in succession we should have successively 8, 7, 6, 5 objects to choose from, and could make our succession of choices in $8 \times 7 \times 6 \times 5$, or 1680 ways. But in doing this we have not only chosen every possible set of 4, but every possible set in every possible order; and since 4 objects can be arranged in order in $4 \times 3 \times 2 \times 1$, or 24 ways, we may find the number of possible choices by dividing 1680 by 24. The result, 70, is essential to our interpretation of the experiment. At best the subject can judge rightly with every cup and, knowing that 4 are of each kind, this amounts to choosing, out of the 70 sets of 4 which might be chosen, that particular one which is correct. A subject without any faculty of discrimination would in fact divide the 8 cups correctly into two sets of 4 in one trial out of 70, or, more properly, with a frequency which would approach 1 in 70 more and more nearly the more often the test were repeated. Evidently this frequency, with which unfailing success would be achieved by a person lacking altogether the faculty under test, is calculable from the number of cups used. The odds could be made much higher by enlarging the experiment, while if the experiment were much smaller even the greatest possible success would give odds so low that the result might, with considerable probability, be ascribed to chance.

7. The Test of Significance

It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result. It is obvious that an experiment would be useless of which no possible result would satisfy him. Thus, if he wishes to ignore results having probabilities as high as 1 in 20—the probabilities being of course reckoned from the hypothesis that the phenomenon to be demonstrated is in fact absent—then it would be useless for him to experiment with only 3 cups of tea of each kind. For 3 objects can be chosen out of 6 in only 20 ways, and therefore complete success in the test would be achieved without sensory discrimination, *i.e.* by "pure chance," in an average of 5 trials out of 100. It is usual and convenient for experimenters to take 5 per cent. as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means, to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. No such selection can eliminate the whole of the possible effects of chance coincidence, and if we accept this convenient convention, and agree that an event which would occur by chance only once in 70 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.

Returning to the possible results of the psychophysical experiment, having decided that if every cup were rightly classified a significant positive result would be recorded, or, in other words, that we should admit that the lady had made good her claim, what should be our conclusion if, for each kind of cup, her judgments are 3 right and 1 wrong? We may take it, in the present discussion, that any error in one set of judgments will be compensated by an error in the other, since it is known to the subject that there are 4 cups of each kind. In enumerating the number of ways of choosing 4 things out of 8, such that 3 are right and 1 wrong, we may note that the 3 right may be chosen, out of the 4 available, in 4 ways and, independently of this choice, that the 1 wrong may be chosen, out of the 4 available, also in 4 ways. So that in all we could make a selection of the kind supposed in 16 different ways. A similar argument shows that, in each kind of judgment, 2 may be right and 2 wrong in 36 ways, 1 right and 3 wrong in 16 ways and none right and 4 wrong in 1 way only. It should be noted that the frequencies of these five possible results of the experiment make up together, as it is obvious they should, the 70 cases out of 70.

It is obvious, too, that 3 successes to 1 failure, although showing a bias, or deviation, in the right direction, could not be judged as statistically significant evidence of a real sensory discrimination. For its frequency of chance occurrence is 16 in 70, or more than 20 per cent. Moreover, it is not the best possible result, and in judging of its significance we must take account not only of its own frequency, but also of the frequency of any better result. In the present instance "3 right and 1 wrong" occurs 16 times, and "4 right" occurs once in 70 trials, making 17 cases out of 70 as good as or better than that observed. The reason for including cases better than that observed becomes obvious on considering what our conclusions would have been had the case of 3 right and 1 wrong only 1 chance, and the case of 4 right 16 chances of occurrence out of 70. The rare case of 3 right and 1 wrong could not be judged significant merely because it was rare, seeing that a higher degree of success would frequently have been scored by mere chance.

8. The Null Hypothesis

Our examination of the possible results of the experiment has therefore led us to a statistical test of significance, by which these results are divided into two classes with opposed interpretations. Tests of significance are of many different kinds, which need not be considered here. Here we are only concerned with the fact that the easy calculation in permutations which we encountered, and which gave us our test of significance, stands for something present in every possible experimental arrangement; or, at least, for something required in its interpretation. The two classes of results which are distinguished by our test of significance are, on the one hand, those which show a significant discrepancy from a certain hypothesis; namely, in this case, the hypothesis that the judgments given are in no way influenced by the order in which the ingredients have been added; and on the other hand, results which show no significant discrepancy from this hypothesis. This hypothesis, which may or may not be impugned by the result of an experiment, is again characteristic of all experimentation. Much confusion would often be avoided if it were explicitly formulated when the experiment is designed. In relation to any experiment we may speak of this hypothesis as the "null hypothesis," and it should be noted that the null hypothesis is never proved or established, but is possibly disproved, in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.

It might be argued that if an experiment can disprove the hypothesis that the subject possesses no sensory discrimination between two different sorts of object, it must therefore be able to prove the opposite hypothesis, that she can make some such discrimination. But this last hypothesis, however reasonable or true it may be, is ineligible as a null hypothesis to be tested by experiment, because it is inexact. If it were asserted that the subject would never be wrong in her judgments we should again have an exact hypothesis, and it is easy to see that this hypothesis could be disproved by a single failure, but could never be proved by any finite amount of experimentation. It is evident that the null hypothesis must be exact, that is free from vagueness and ambiguity, because it must supply the basis of the "problem of distribution," of which the test of significance is the solution. A null hypothesis may, indeed, contain arbitrary elements, and in more complicated cases often does so: as, for example, if it should assert that the death-rates of two groups of animals are equal, without specifying what these death-rates actually are. In such cases it is evidently the equality rather than any particular values of the death-rates that the experiment is designed to test, and possibly to disprove.

In cases involving statistical "estimation" these ideas may be extended to the simultaneous consideration of a series of hypothetical possibilities. The notion of an error of the so-called "second kind," due to accepting the null hypothesis "when it is false" may then be given a meaning in reference to the quantity to be estimated. It has no meaning with respect to simple tests of significance, in which the only available expectations are those which flow from the null hypothesis being true. Problems of the more elaborate type involving estimation are discussed in Chapter IX.

9. Randomisation; the Physical Basis of the Validity of the Test

We have spoken of the experiment as testing a certain null hypothesis, namely, in this case, that the subject possesses no sensory discrimination whatever of the kind claimed; we have, too, assigned as appropriate to this hypothesis a certain frequency distribution of occurrences, based on the equal frequency of the 70 possible ways of assigning 8 objects to two classes of 4 each; in other words, the frequency distribution appropriate to a classification by pure chance. We have now to examine the physical conditions of the experimental technique needed to justify the assumption that, if discrimination of the kind under test is absent, the result of the experiment will be wholly governed by the laws of chance. It is easy to see that it might well be otherwise. If all those cups made with the milk first had sugar added, while those made with the tea first had none, a very obvious difference in flavour would have been introduced which might well ensure that all those made with sugar should be classed alike. These groups might either be classified all right or all wrong, but in such a case the frequency of the critical event in which all cups are classified correctly would not be 1 in 70, but 35 in 70 trials, and the test of significance would be wholly vitiated. Errors equivalent in principle to this are very frequently incorporated in otherwise well-designed experiments.

It is no sufficient remedy to insist that " all the cups must be exactly alike " in every respect except that to be tested. For this is a totally impossible requirement in our example, and equally in all other forms of experimentation. In practice it is probable that the cups will differ perceptibly in the thickness or smoothness of their material, that the quantities of milk added to the different cups will not be exactly equal, that the strength of the infusion of tea may change between pouring the first and the last cup, and that the temperature also at which the tea is tasted will change during the course of the experiment. These are only examples of the differences probably present ; it would be impossible to present an exhaustive list of such possible differences appropriate to any one kind of experiment, because the uncontrolled causes which may influence the result are always strictly innumerable. When any such cause is named, it is usually perceived that, by increased labour and expense, it could be largely eliminated. Too frequently it is assumed that such refinements constitute improvements to the experiment. Our view, which will be much more fully exemplified in later sections, is that it is an essential characteristic of experimentation that it is carried out with limited resources, and an essential part of the subject of experimental design to ascertain how these should be best applied ; or, in particular, to which causes of disturbance care should be given, and which *ought* to be deliberately ignored. To ascertain, too, for those which are not to be ignored, to what *extent* it is worth while to take the trouble to diminish their magnitude. For our present purpose, however, it is only necessary to recognise that, whatever degree of care and experimental skill is expended in equalising the conditions, other than the one under test, which are liable to affect the result, this equalisation must always be to a greater or less extent incomplete, and in many important practical cases will certainly be grossly defective. We are concerned, therefore, that this inequality, whether it be great or small, shall not impugn the exactitude of the frequency distribution, on the basis of which the result of the experiment is to be appraised.

10. The Effectiveness of Randomisation

The element in the experimental procedure which contains the essential safeguard is that 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 randomisation, by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated. To demonstrate that, with satisfactory randomisation, its validity is, indeed, wholly unimpaired, let us imagine all causes of disturbance—the strength of the infusion, the quantity of milk, the temperature at which it is tasted, etc.—to be predetermined for each cup; then since these, on the null hypothesis, are the only causes influencing classification, we may say that the probabilities of each of the 70 possible choices or classifications which the subject can make are also predetermined. If, now, after the disturbing causes are fixed, we assign, strictly at random, 4 out of the 8 cups to each of our experimental treatments, then every set of 4, whatever its probability of being so classified, will certainly have a probability of exactly 1 in 70 of *being* the 4, for example, to which the milk is added first. However important the causes of disturbance may be, even if they were to make it certain that one particular set of 4 should receive this classification, the probability that the 4 so classified and the 4 which ought to have been so classified should be the same, must be rigorously in accordance with our test of significance.

It is apparent, therefore, that the random choice of the objects to be treated in different ways would be a complete guarantee of the validity of the test of significance, if these treatments were the last in time of the stages in the physical history of the objects which might affect their experimental reaction. The circumstance that the experimental treatments cannot always be applied last, and may come relatively early in their history, causes no practical inconvenience; for subsequent causes of differentiation, if under the experimenter's control, as, for example, the choice of different pipettes to be used with different flasks, can either be predetermined before the treatments have been randomised, or, if this has not been done, can be randomised on their own account; and other causes of differentiation will be either (*a*) consequences of differences already randomised, or (*b*) natural consequences of the difference in treatment to be tested, of which on the null hypothesis there will be none, by definition, or (*c*) effects supervening by chance independently from the treatments applied. Apart, therefore, from the avoidable error of the experimenter himself introducing with his test treatments, or subsequently, other differences in treatment, the effects of which the experiment is not intended to study, it may be said that the simple precaution of randomisation will suffice to guarantee the validity of the test of significance, by which the result of the experiment is to be judged.

11. The Sensitiveness of an Experiment. Effects of Enlargement and Repetition

A probable objection, which the subject might well make to the experiment so far described, is that only if every cup is classified correctly will she be judged successful. A single mistake will reduce her performance below the level of significance. Her claim, however, might be, not that she could draw the distinction with invariable certainty, but that, though sometimes mistaken, she would be right more often than not; and that the experiment should be enlarged sufficiently, or repeated sufficiently often, for her to be able to demonstrate the predominance of correct classifications in spite of occasional errors.

An extension of the calculation upon which the test of significance was based shows that an experiment with 12 cups, six of each kind, gives, on the null hypothesis, 1 chance in 924 for complete success, and 36 chances for 5 of each kind classified right and 1 wrong. As 37 is less than a twentieth of 924, such a test could be counted as significant, although a pair of cups have been wrongly classified; and it is easy to verify that, using larger numbers still, a significant result could be obtained with a still higher proportion of errors. By increasing the size of the experiment, we can render it more sensitive, meaning by this that it will allow of the detection of a lower degree of sensory discrimination, or, in other words, of a quantitatively smaller departure from the null hypothesis. Since in every case the experiment is capable of disproving, but never of proving this hypothesis, we may say that the value of the experiment is increased whenever it permits the null hypothesis to be more readily disproved.

The same result could be achieved by repeating the experiment, as originally designed, upon a number of different occasions, counting as a success all those occasions on which 8 cups are correctly classified. The chance of success on each occasion being 1 in 70, a simple application of the theory of probability shows that 2 or more successes in 10 trials would occur, by chance, with a frequency below the standard chosen for testing significance; so that the sensory discrimination would be demonstrated, although, in 8 attempts out of 10, the subject made one or more mistakes. 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. It would clearly be illegitimate, and would rob our calculation of its basis, if the unsuccessful results were not all brought into the account.

12. Qualitative Methods of increasing Sensitiveness

Instead of enlarging the experiment we may attempt to increase its sensitiveness by qualitative improvements; and these are, generally speaking, of two kinds: (*a*) the reorganisation of its structure, and (*b*) refinements of technique. To illustrate a change of structure we might consider that, instead of fixing in advance that 4 cups should be of each kind, determining by a random process how the subdivision should be effected, we might have allowed the treatment of each cup to be determined independently by chance, as by the toss of a coin, so that each treatment has an equal chance of being chosen. The chance of classifying correctly 8 cups randomised in this way, without the aid of sensory discrimination, is 1 in 2^8, or 1 in 256 chances, and there are only 8 chances of classifying 7 right and 1 wrong; consequently the sensitiveness of the experiment has been increased, while still using only 8 cups, and it is possible to score a significant success, even if one is classified wrongly. In many types of experiment, therefore, the suggested change in structure would be evidently advantageous. For the special requirements of a psycho-physical experiment, however, we should probably prefer to forego this advantage, since it would occasionally occur that all the cups would be treated alike, and this, besides bewildering the subject by an unexpected occurrence, would deny her the real advantage of judging by comparison.

Another possible alteration to the structure of the experiment, which would, however, decrease its sensitiveness, would be to present determined, but unequal, numbers of the two treatments. Thus we might arrange that 5 cups should be of the one kind and 3 of the other, choosing them properly by chance, and informing the subject how many of each to expect. But since the number of ways of choosing 3 things out of 8 is only 56, there is now, on the null hypothesis, a probability of a completely correct classification of 1 in 56. It appears in fact that we cannot by these means do better than by presenting the two treatments in equal numbers, and the choice of this equality is now seen to be justified by its giving to the experiment its maximal sensitiveness.

With respect to the refinements of technique, we have seen above that these contribute nothing to the validity of the experiment, and of the test of significance by which we determine its result. They may, however, be important, and even essential, in permitting the phenomenon under test to manifest itself. Though the test of significance remains valid, it may be that without special precautions even a definite sensory discrimination would have little chance of scoring a significant success. If some cups were made with India and some with China tea, even though the treatments were properly randomised, the subject might not be able to discriminate the relatively small difference in flavour under investigation, when it was confused with the greater differences between leaves of different origin. Obviously, a similar difficulty could be introduced by using in some cups raw milk and in others boiled, or even condensed milk, or by adding sugar in unequal quantities. The subject has a right to claim, and it is in the interests of the sensitiveness of the experiment, that gross differences of these kinds should be excluded, and that the cups should, not as far as *possible*, but as far as is practically convenient, be made alike in all respects except that under test.

How far such experimental refinements should be carried is entirely a matter of judgment, based on experience. The validity of the experiment is not affected by them. Their sole purpose is to increase its sensitiveness, and this object can usually be achieved in many other ways, and particularly by increasing the size of the experiment. If, therefore, it is decided that the sensitiveness of the experiment should be increased, the experimenter has the choice between different

methods of obtaining equivalent results; and will be wise to choose whichever method is easiest to him, irrespective of the fact that previous experimenters may have tried, and recommended as very important, or even essential, various ingenious and troublesome precautions.

12·1. Scientific Inference and Acceptance Procedures

In "The Improvement of Natural Knowledge", that is, in learning by experience, or by planned chains of experimentation, conclusions are always provisional and in the nature of progress reports, interpreting and embodying the evidence so far accrued. Convenient as it is to note that a hypothesis is contradicted at some familiar level of significance such as 5% or 2% or 1% we do not, in Inductive Inference, ever need to lose sight of the exact strength which the evidence has in fact reached, or to ignore the fact that with further trial it might come to be stronger, or weaker. The situation is entirely different in the field of Acceptance Procedures, in which irreversible action may have to be taken, and in which, whichever decision is arrived at, it is quite immaterial whether it is arrived at on strong evidence or on weak. All that is needed is a Rule of Action which is to be taken automatically, and without thought devoted to the individual decision. The procedure as a whole is arrived at by minimising the losses due to wrong decisions, or to unnecessary testing, and to frame such a procedure successfully the cost of such faulty decisions must be assessed in advance; equally, also, prior knowledge is required of the expected distribution of the material in supply. In the field of pure research no assessment of the cost of wrong conclusions, or of delay in arriving at more correct conclusions can conceivably be more than a pretence, and in any case such an assessment would be inadmissible and irrelevant in judging the state of the scientific evidence; moreover, accurately assessable prior information is ordinarily known to be lacking. Such differences between the logical situations should be borne in mind whenever we see tests of significance spoken of as " Rules of Action ". A good deal of confusion has certainly been caused by the attempt to formalise the exposition of tests of significance in a logical framework different from that for which they were in fact first developed.

REFERENCES AND OTHER READING

R. A. Fisher (1925-1963). Statistical methods for research workers. Chap. III., §§ 15-19

R. A. Fisher (1926). The arrangement of field experiments. Journal of Ministry of Agriculture, xxxiii. 503-513.

III

A HISTORICAL EXPERIMENT ON GROWTH RATE

13. We have illustrated a psycho-physical experiment, the result of which depends upon judgments, scored " right " or " wrong," and may be appropriately interpreted by the method of the classical theory of probability. This method rests on the enumeration of the frequencies with which different combinations of right or wrong judgments will occur, on the hypothesis to be tested. We may now illustrate an experiment in which the results are expressed in quantitative measures, and which is appropriately interpreted by means of the theory of errors.

In the introductory remarks to his book on " The effects of cross and self-fertilisation in the vegetable kingdom," Charles Darwin gives an account of the considerations which guided him in the design of his experiments and in the presentation of his data, which will serve well to illustrate the principles on which biological experiments may be made conclusive. The passage is of especial interest in illustrating the extremely crude and unsatisfactory statistical methods available at the time, and the manner in which careful attention to commonsense considerations led to the adoption of an experimental design, in itself greatly superior to these methods of interpretation.

14. Darwin's Discussion of the Data

" I long doubted whether it was worth while to give the measurements of each separate plant, but have