290

THE PLACE OF THE DESIGN OF EXPERIMENTS IN THE LOGIC OF SCIENTIFIC INFERENCE

By SIR RONALD A. FISHER

Division of Mathematical Statistics, C.S.I.R.O. ***University of Adelaide***

I - EARLY AIMS.

When, a little more than 25 years ago, I first attempted a systematic exposition of the subject, known as the Design of Experiments, it is no very grave confession to avow that I did not fully understand the position among the statistical sciences of this new discipline. My approach at that time was frankly a technological one. As a statistician I had often set myself the task of analysing experimental data, and was much concerned with those improvements in statistical methods, which promised to make such analysis more thorough and more comprehensive. Technically, I could see that some methods were superior to others in the concrete sense of extracting from the data more "information" on the subjects under enquiry, and therefore of leading to estimates of higher precision, and to tests of significance of greater sensitivity. And so it was, in this atmosphere, borne in upon me that very often, when the most elaborate statistical refinements possible could increase the precision by only a few per cent, yet a different design involving little or no additional experimental labour might increase the precision two-fold, or five-fold or even more, and could often supply information in addition on relevant supplementary questions on which the original design was completely uninformative.

It was thus clear at an early stage that there were quantitatively large technological gains to be obtained through the deliberate study of Experimental Design, and that these gains were to be harvested by making the plan of experimentation and observation logically coherent with the aims of the experiment, or in other words with the kind of inference about the real world, which it was to be hoped the experimental results would permit. At this point we catch sight of some questions concerning the nature of the inferences which are possible from the data of the Natural Sciences, which seem still to be matters of dispute, and of which I hope to speak more fully at a later stage.

II - COMBINATORIAL PROGRESS.

For the present, it will be familiar to us all that in the numerous books which have appeared and are appearing on the Design of Experiments the chief attention has been given to developments in combinatorial mathematics often of quite an intricate character, which have undoubtedly served greatly to enlarge the experimenter's repertoire of designs effective in very varied fields of enquiry. Their efflorescence has been rapid, and it must have given pleasure to experimenters in many parts of the world to learn, as we learnt little more than a year ago, of the success of R. C. Bose and his American and Indian colleagues in settling a problem in pure mathematics, which has been a challenge since the time of Euler, by demonstrating the existence of Graeco-Latin squares of side 10, and some other numbers of the form $4r + 2$.* Euler had satisfied himself of the non-existence of such squares of side six, though he had not enumerated the Latin squares exhaustively, or recognised the existence of the twelve types into which they may be classified. Even for the 6×6 squares his paper left room for doubt,

* **For some other numbers, read all larger numbers.**

Colloques Internationaux du Centre National de la Recherche Scientifique, **(Paris), No. 110:** 13-19, (1962); translated into Italian in *La Scuola in Azione*, 9: 33-42.

though his conclusion was, in fact, correct. He also suggested that the same might be true for all even numbers not divisible by four. For my own part I think such a conjectural suggestion, which he never suggested he had demonstrated, desirable for the advancement of mathematical thought. Bose's discovery seems to do no injury to Euler's reputation, but it is, of course, a decisive refutation of several claims made in the *twentieth* century to have proved the truth of Euler's conjecture, and of even more inclusive theorems, by topological methods. The case is not that of a contribution of pure mathematics to experimental design so much as a by-product of the intensification of interest in combinatorial mathematics due to the modern recognition of their importance in this applied field. It is certainly good to have a supposed restriction decisively overcome.

III - SCIENTIFIC EXPERIENCE.

Now, if we are to encourage the experimental scientist to seek expert advice *before* the experiment is executed, and not merely, as too often in the past, to call in a statistician for the *post mortem*, I think we must do a little more than many of the writers of these books seem to consider necessary, and this in two respects. The literature as it has grown up seems to be unbalanced in its comparative neglect of the *Scientific* aspects of the problem, and of its *Logical* aspects. This perhaps might have been expected, since many of the authors, albeit talented mathematicians, have evidently never submitted their minds to the specifically educational discipline of any one of the Natural Sciences, have never taken personal responsibility for experimentation at ground level, and have no direct experience of the kind of material involved, but only of the reactions of mathematically minded students, exhibiting acuity perhaps, but without depth of focus. Seminars discussing and rediscussing questions, whether substantial or trivial, but always at this shallow level, taking time that might have been given to laboratory experience, must bear their share of responsibility, for the sameness, and lack of constructive thought, of the many new books which come on to our desks in their glossy covers.

There is, frankly, no easy substitute for the educational discipline of whole-time personal responsibility for the planning and conduct of experiments designed for the ascertainment of fact, or the improvement of Natural Knowledge, I say "educational discipline" because such experience trains the mind and deepens the judgement for innumerable ancillary decisions, on which the value or cogency of an experimental programme depends. A man with five, or ten, or fifteen years experience given to such discipline has been himself profoundly modified in his capacity for the direction of such work. He has, as we say, learnt by experience, and this effect will be the more profound the more deeply his thought has been immersed in his problems. Such men, if they have the taste and gift for exposition, should be the authors of our text-books on Experimental Design, and the teachers and directors of our advanced schools of statistics.

For, in truth, the Design of Experiments is not, as it might have been thought but a few years ago, a casual extension of statistical studies, but is central to the whole process of the Natural Sciences. In this process we proceed from observational data, imperfect indeed in that they are qualified by errors of measurement, and by errors of random sampling, to conclusions, which by reason of these imperfections are subject to some uncertainty, which may, none the less, be an uncertainty of a rigorously well-defined kind. The strongest possible type of statement of uncertainty to which our investigations can lead are statements of Mathematical Probability, as understood by the great French mathematicians of the 17th century. The possibility of inferring such statements depends on the quality of the data which provide the premises of the inference. To ensure that the data shall provide a logical foundation for inferring exact statements of mathematical probability is one of the tasks to be considered in experimental design.

IV - MATHEMATICAL PROBABILITY.

I have in recent years made use of a definition, or semantic analysis, of the concept of mathematical probability intended to bring out more thoroughly both the logical and the naturalistic aspects of this concept, than is done by such bare phrases as "the frequency theory". I conceive that there are three requirements for a probability statement which shall be precisely defined and valid in the real world. First, it must imply a Reference Set, a well defined mathematical construct, which must be measurable in the sense that members of a precise fraction, P, of the whole set possess some characteristic absent from the others. This defines the strictly mathematical aspect of the concept, and suffices to ensure that the statement of probability shall be exact. Secondly, we require that the subject of the probability statement shall be a member of the Set; this is the specifically scientific or naturalistic aspect, which must be taken care of by experimental design. For it entails the characteristically scientific processes of recognition and identification. Finally, there is the specifically logical requirement that no subset having a fraction different from P, can be recognised. Such subsets always exist. Their non-recognisability is a postulate of ignorance, for if the nature and extent of our uncertainty is to be well defined, it is necessary to specify our ignorance just as exactly as our knowledge. Manifestly, if the subject did belong to such a recognisable subset the latter would replace the original set, as the appropriate basis for the probability statement.

It will be noticed that this specification may be called a frequency theory of probability in the sense that the value of the probability asserted may theoretically by verified to any chosen degree of approximation by sampling the Reference Set, and this, I believe, is the only meaning that can be attached to the phrase "frequency theory". The three distinct stipulations I have made have, however, distinct purposes in satisfying the requirements, first, that the statement shall be mathematically exact; secondly, that it shall be valid in the real world of the Natural Sciences; and thirdly that it shall incorporate a rigorous specification of the nature and extent of the uncertainty.

V - MEASUREMENT IN NATURE.

If, then, it be proposed to obtain knowledge of some quantitative property of the real world, such as Planck's constant of angular momentum, the atomic weight of some chemical element, the average number of vertebrae in the European eel, or the time constant of the expanding universe, though we cannot attain to absolute exactness, yet we may aim to base our knowledge on data good enough to lead by a rigorous argument to precise statements of probability, rather than to those weaker levels of uncertainty represented by Mathematical Likelihood, or only by tests of significance. The conclusion of our induction will be of the form:

$$\Pr(x < x_P) = P,$$

asserting that the Mathematical Probability that the unknown x shall be less than a value x_P calculable with exactitude from the observations, is equal to P for all values of P from 0 to 1. In other words we shall express the unknown as a *Random Variable*, not imagining that it has more than one value, but incorporating in our conclusion the fact that within a restricted region we do not know what value it has, just as the gambler of the 17th century did not imagine that the die he was about to throw would turn up simultaneously all six faces, but in stating that the probability of an ace was just 1/6, asserted by implication that he did not know which face it would be.

The task we have thrown upon experimental design of providing data good enough to support conclusions of this definite kind is not *prima facie* a difficult one, although technical knowledge of the material would be required for confidence in whatever method suggests itself. If in any problem of this sort the experimenter's mastery of his material is such that he can obtain a Normal sample of, say, ten or a dozen independent measurements which, though inexact, are collectively unbiased, and of equal precision, then he can provide the empirical basis for such a continuum of probability statements, strict as are the mathematical conditions for such an outcome. Though the mathematics are familiar, the logic, which is equally rigorous, seems not to be so widely appreciated.

A random sample of N from and Normal population may be characterised by three quantities, the population mean μ, the sample mean $\bar{x}$, and the estimated standard deviation s, calculated from:

$$N(N-1)s^2 = \Sigma(x-\bar{x})^2 .$$

Such triads of values, appropriate to all random samples from all possible Normal populations, constitute our Reference Set, which is thus rigorously defined. Some members of this set, but not all, will satisfy the inequality

$$\mu < \bar{x} + st_P ,$$

and if t_P is calculated so that

$$\Pr(t < t_P) = \mathrm{P} ,$$

whatever may be the parent population, for any chosen value of P, then, since the definition of t is

$$t = \frac{\mu - \bar{x}}{s} ,$$

the frequency with which the inequality is satisfied in the Reference Set must be equal to P. Knowing $\bar{x}$ and s for his particular data the experimenter may thus easily infer the specification of the unknown μ as a Random Variable. The mathematical part of the inference is simple in this case; it is provided by Student's distribution of t, which has been adequately tabulated. In its logical aspect two further stipulations are required.

VI - LOGICAL AND OPERATIVE CONDITIONS.

If Bayesian probability *a priori* had been available, that is to say, if the constant to be determined were expressible prior to the observations exactly as a random variable, we should properly use the method of Bayes to arrive at the distribution *a posteriori* appropriate to the observations. This is indeed a dream from fairyland for we do not in fact, prior to observation, possess such a continuum of exact probability statements. However, if indeed we did, it would be very wrong to ignore them, and Bayes' argument is available for just this case, and for the more realistic one in which an auxiliary experiment can be used to supply this ingredient of Bayes' argument.

Secondly, if the statistics $\bar{x}$ and s used in our inequality had not been jointly exhaustive for the extimation of the two parameters of the Normal distribution, it would have been possible to find some third statistic such that its distribution, subject to constrained values of $\bar{x}$ and s, still depended to some extent on the unknown μ. Such a statistic would, in fact, define a recognisable subset within the general class of random probability statements from those we have inferred.

The validity of the inductive argument has, then, the logical requirements that prior to observation the unknown is not already expressed exactly as a random variable, and that the statistics used in the probability statements inferred should jointly constitute an exhaustive set. To lead to estimation problems admitting of exhaustive extimates is therefore a reasonable aim for an experimental programme, which is fully realized if it can obtain such a Normal sample as I have supposed.

Many experimenters will consider, at first sight, that the task which, in this example, we have assigned to experimental design is an easy one. It does seem, none the less, to deserve some thought, and some precautions. The mere fact that N values are presented to the statistician does not justify him in accepting them as a random sample from a Normal population. Measurements, each with its own components of error, but collectively unbiased are the sort of observational material from which such a situation was first inferred. When steps are taken to improve the accuracy of a measurement, the largest and simplest sources of error are first eliminated; there remain smaller, but more numerous, components having collectively a smaller variance. The limit of this process, closely approached in all measurements of precision, is the Gaussian or Normal form.

The experimenter may now consider that he wants his observations to be *collectively unbiased*, *independent*, and of *equal precision*. The policy adopted by many pharmaceutical firms in making assays of the potency of pharmaceutical preparations affords the best answer to which I can point to the experimenters problem under these heads. Recognizing that systematic errors inevitably affect the determinations in any one laboratory or department, so that repeated determinations by the same department are not truly independent, the collaboration of several laboratories is obtained, and though it is often useful for each laboratory to make a number of parallel determinations, as independently as it can, yet the variance among these will not be used in the estimation of error, but rather as an internal confirmation of that estimate, which will itself be based exclusively on the variation among laboratories. So, if there were eight laboratories each performing five parallel measurements, the thirty-two degrees of freedom within laboratories are confirmatory only, and should give a mean square somewhat less, and sometimes strikingly less, than the seven degrees of freedom among laboratories.

For equality of precision it is necessary that the number of repetitions shall be the same in each laboratory. The apparatus, balances, glassware, standard reagents, etc. will be different, though they may be supplied by the same firms, or approved by the same standarizing authority. Scientists and technicians will, of course, be different, but sensibly equivalent in capacity and experience. They should not be required to undertake tasks involving exceptional personal skill. The different departments should accept in advance, after due discussion, and carry out without variation, an agreed programme including schedules of interim computations. An erroneous mathematical theory connecting the crude readings with the quantities to be measured will still introduce a bias affecting all departments equally, but this can be removed easily at a later stage, if the mathematics are corrected.

Without particularizing the special object in view, and speaking to our audience with wide theoretical interests, it would be impossible to descend to the technical details of any such research. I have, however, thought it worthwhile to emphasize that the design of experiments is a subject with not only mathematical, but also scientific and logical aspects. And that the thoroughness with which the logic and the mathematics have been explored with respect to the Gaussian sample should not blind us to the fact that the production of *good* data of this kind is a technological accomplishment worthy to sustain the perfected form of inference which such data make possible.