282

SCIENTIFIC THOUGHT AND THE REFINEMENT OF HUMAN REASONING

SIR RONALD AYLMER FISHER

University of Adelaide(*Presented at the Special Meeting on May 28, 1960, sponsored by the Asahi Shimbun under the joint collaboration of the Mathematical Society of Japan, the Research Association of Statistical Sciences and the Operations Research Society of Japan.*)

It is a truism among philosophers that the external world is only perceived by us through the medium of our sense organs, and the brain, or central nervous system, associated with them. This is indeed no reason for doubting, as some philosophers have doubted, the objective reality of that real world which is accessible to study by the methods of science. It is a reason, on the other hand, for doubting whether our understanding of Nature can advance far, can become more reliable and more penetrating, without an accompanying re-assessment of the reasoning processes by which this understanding is accomplished. For, as science advances, as new concepts are brought into discussion, so do the tasks change which our reasoning minds are attempting to perform.

The physical basis of the human mind, has not apparenty been enlarged or improved since that remote era, about 50,000 years ago, when *Homo sapiens*, the species to which human beings throughout the world belong, re-entered the small continent of Europe, previously inhabited for thousands of years by a different species, *Homo neanderthalensis*, now extinct. Though the physical basis of mental life was much

the same, the contents of the mind must, for much of the intervening period, have been greatly different. Primitive men appreciate their world much more emotionally than we do ourselves ; the disentangling of intellection from emotion must have taken place gradually, and by several stages.

Unambiguous words for numbers, and a strict grammatical structure of the language seem to be required before the human mind can conceive that a statement can be rigorously correct, or a quantitative value perfectly accurate. On this basis can be raised the concept of a rigorous deductive proof, that is, a formally faultless transition from one or more statements accepted as perfectly correct and exact, to others, which, though derivative, possess precisely the same quality of certainty. The formal elaboration of the principles of deductive proof seems to have been the work in ancient Greece of the geometrical predecessors of Euclid, and of Euclid himself. Deductive logic was thus nursed in a medium, both aesthetically attractive *and* utilitarian, without which the aid to exact thought which it affords might not have survived the turbulent centuries which perceded the revival of learning in Europe.

INDUCTIVE LOGIC

Though deductive reasoning was thus familiar and tolerably well understood in the traditional intellectual heritage of the modern world, the same could not be said of inductive reasoning, based on observational material, with its errors of observation, and its errors of random sampling, by which, however, all that we know of the real world studied by the sciences must be inferred. Indeed, there seem to be in the United States many converts to the opinion of J. Neyman[4] who so recently as 1938 averred that inductive reason positively did not exist, and that no process deserving the name of *reasoning* could be applied to the data of science. In expressing this opinion, in the fourth decade of our century Neyman was, in my opinion, some hundred years, and probably much more, out of date, for the great Karl Gauss had developed, though he had not finally perfected, such a process for the interpretation of the data obtained in astronomy and geodesy ; moreover 100 years earlier still the theorem of Thomas Bayes[1] (1763) would have

to be regarded as a meaningless gesture were it not for its deliberate aim to make mathematically rigorous statements (though still uncertain statements), about an unknown parameter, on the basis of data susceptible to errors of random sampling. Neyman's doctrine challenged not only the rapidly developing statistical science of the 20th century, but its foundations in the 19th and 18th centuries. On the contrary, it will be my thesis that the continuous development of mathematical thought in Western Europe from the great French mathematicians of the 17th century onward, has come to fruition in our own time, by cross-fertilization with the Natural Sciences, in supplying just such a model of the correct use of inductive reasoning, as was supplied by Euclid for *deductive* logic. A model only, for the development of its manifold latent possibilities has remained almost untouched, as any reader of my book *Statistical Methods and Scientific Inference*[2] (1956, 1959) will easily appreciate.

MISAPPREHENSIONS

Some of the obstacles which have stood in the way of the rational exploitation of these opportunities are perhaps worthy of special consideration. I spoke a moment ago of *rigorous*, though still *uncertain*, statements. This was not a paradox. The word rigorous referred to the process of reasoning, the inductive logic, by which correct inferences may be drawn from observations, imperfect in the various ways characteristic of scientific observations. Because they are imperfect, because, as one might put it, the observational basis of our reasoning might equally have been somewhat different from the data which we have, it would be a failure of rigour to draw inferences purporting to be statements of certainty. To be rigorously complete our inferences *must* incorporate a mathematically correct specification of the nature and extent of the uncertainty by which they are affected. There is then no contradiction, such as might superficially be imagined, between the rigour of the argument and the uncertainty of its conclusions. The argument must indeed be more subtle than those required in merely deductive reasoning, and means must be found for expressing correctly the uncertainty necessarily entailed. Fortunately, since the 17th century, mathematicians have become familiar with the concept of Mathematical Probability,

introduced in its origin to specify exactly that kind of uncertainty which confronts a gambler in a game of chance, played fairly and with perfect appratus. It is, indeed not always the case that the uncertainty of scientific conclusions can be accurately specified in this way. Other kinds of uncertainty exist[3], and some of these can be accurately specified, but the cases in which Mathematical Probability meets the need are numerous and important. Rigorous inferences expressible in terms of Mathematical Probability are, indeed, the strongest type of uncertain inferences that can be made. Experiments may properly be designed so as to lead to inferences of this type. Weaker statements are, none the less, very often serviceable.

A curious misapprehension has arisen in the present century about the applicability of probability statements to such numerical values of the natural world as the distance of the Sun. For it has been claimed[4] that if x stand for this distance then no statement of the mathematical form

$$Pr(x<x_p)=P$$

can be derived from the imperfect data of astronomy, when x_p is calculated from these data. This claim is made on the remarkable ground that if the true distance of the Sun were greater than x_p, then the probability should be zero, while if it were less, the probability should be unity. The objection is a surprising one, for it seems to show a complete misunderstanding of the correct usage of the word probability, as it has been recognized for hundreds of years. For the probability statement implies that it is not known, with mathematical certainty, whether the true distance be greater or less than the comparison value x_p calculated from the data. If, indeed, a new datum were *added* to that on which the reasoning has been based, asserting an exact value for the unknown, the probability statement would cease to be the correct inference. It would be not only futile, but erroneous, on the new data supposed. In fact inductive logic resembles that of deduction in that from different premises different conclusions are properly inferred. It is truly astonishing to find this elementary error, incorporated in the teaching of many mathematical departments in the United States. Especially, when we remember that during the inter-war period, the U. S. seemed likely to become one of the worlds leaders in Statistics. The

rebuttal of this error does not of course imply that the data of astronomy are necessarily competent to supply probability statements on just this subject. Astronomers do, however, use "Probable errors", and it is doubtful if any competent astronomer seriously thinks that probability statements about astronomical parameters are, in principle, inadmissible.

MATHEMATICAL PROBABILITY

For the validity of probability statements about the real world there are I believe only *three* necessary and sufficient requirements. (i) As Kolmogoroff rightly insisted now many years ago every statement of mathematical probability implies a mathematically well-defined Reference Set of possibilities, which must be measurable at least so far that members of the Set, comprising a known fraction P of the whole, possess some characteristic which is absent from the remainder. (ii) The subject, or particular entity about which the probability statement is asserted, must be a member of this Set. (iii) No sub-set may be recognizable having a fraction possessing the characteristic differing from the fraction P of the whole. Such a precise specification, or semantic analysis, of the meaning of the word is necessary if it is to be used in a mathematically unambiguous manner. If the three conditions are satisfied then a correct statement of mathematical probability is possible ; if any one of them is not satisfied then manifestly such a statement fails. There is thus only one *kind* of mathematical probability, and the distinction introduced by some writers between " fiducial probability " and " ordinary probability " is a good example of what is called a " distinction without a difference " All genuine probability statements are necessarily of the same kind, whether the premises from which they are derived are *observational,* using the fiducial argument, or *axiomatic.*

The three stipulations I have made for the validity of a probability statement serve different purposes. The first is specifically Mathematical, it requires an abstraction, a well defined Set, measurable at least in some respects. The quantitative element of our statement is then mathematically precise. The second requirement, that the subject of the statement shall belong to the Set, introduces realism. It puts the statement into the real world, by requiring particularity, and that kind of *recognition* and *identification,* which is characteristic of work in the

Natural Sciences, in Chemistry or Physics, or any branch of Biology, but scarcely so in Mathematics. The third stipulation is more interesting *logically*. It is a postulate of ignorance, such as is quite unfamiliar in purely deductive reasoning, but is obviously necessary when a state of uncertainty needs to be specified with rigorous exactitude, and, therefore, when the extent of our ignorance must be just as explicitely recognised as the state of our knowledge. Three distinct stipulations are the least that is necessary for defining a type of statement which must be mathematically exact, which must be valid in the real world, and which must incorporate a well defined degree of uncertainty. The logical nature of the concept of mathematical probability, as it has been understood by mathematicians for centuries, thus makes it peculiarly appropriate to the needs of scientific inference, whenever the observational data are sufficient to supply inferences of this type.

TESTS OF SIGNIFICANCE

All tests of significance involve probability statements, though these are conditioned on the truth of the hypothesis to be tested. Carelessness as to the appropriate reference set by which the hypothesis can supply probability statements verifiable from the kind of observations available, has led to numerous disagreements. These have originated largely from an unfortunate phrase formerly used by Neyman and Pearson[6] in expounding their "Theory of Testing Hypotheses", namely that the level of significance could be defined by "repeated sampling from the same population." Of course, the only way in which a probability statement can be verified by sampling is to find some means of sampling the appropriate Reference Set, and this is not often accomplished by repeating mechanically the operations by which the original data came into existence. The phrase was in fact an unhelpful one, and so long as it was didactically repeated, it inhibited exponents of this school from seeing or understanding any further. A whole series of erroneous tests of significance were incorporated into statistical teaching, and although one by one they have been exposed and discredited, and have seldom gained a place among the tests used in genuine research, they have ensured that many young men now entering employment in research, or industry, or administration, have been partly incapacitated by

the crooked reasoning with which they have been indoctrinated. Familiar examples include the test of proportionality in a two by two table, perhaps the most frequently used of all tests of significance, in which the recognizable sub-set of possibilities having the same marginal totals as the sample observed, has been more than once over-looked, while stress has been laid on the futile question whether the data were obtained by a sampling process in which one pair, or the other, or both pairs of the marginal totals were "fixed" or invariant. It is, on the contrary, the fact that the marginal totals in the sample are *known* or *recognizable*, which defines the sub-set of possibilities appropriate and available for the test of significance. The test would be the same even if the process of sampling used could not have led to any other member of the sub-set, as if a count had been made of ducks and geese, and of drakes and ganders, using the convention that the count should be closed when a fixed number, such as 20, of the occupants of same one cell had been enumerated. The relative frequencies of the series of possible samples

	♀	♂	♀	♂	♀	♂
Ducks	19	$b+1$	20	b	21	$b-1$
Geese	$c+1$	$d-1$	c	d	$c-1$	$d+1$

extended to the vanishing point in both directions, is just as relevant to the test of proportionality of the sample observed, whether or not, in the matrix of causation by which it came into existence, the margins, or any one of the individual entries, were conventionally "fixed". The rules by which it was decided to close the count may be quite unknown, and are irrelevant to testing the significance of the evidence. The example is a good one as showing the difference between "random sampling of the same population", and the genuine verification of a probability statement by sampling its reference set.

The series of abortive attempts to solve the rather simple problem of testing the significance of the difference between the observed means of two Normal samples, when both of the true variances are independently unknown, also involves the ignoration of a critical and recognizable sub-set, namely that defined by the ratio s_1^2/s_2^2 of the two available estimates of variance. But the same unfortunate phrase has been misleading in another way. For some authors such as Pearson, have defended the view that in testing a *composite hypothesis* the level

of significance must be equated to the frequency with which the test criterion is satisfied for any particular case within the composite.

This doctrine, which has been very dogmatically asserted, makes a truly marvellous mystery of the tests of significance. On the earlier view, held by all those to whom we owe the first examples of these tests, such a test was logically elementary. It presented the logical disjunction: Either the hypothesis is not true, or an exceptionally rare outcome has occurred. If we are speaking of a composite family of hypotheses the position might be: either no one of these hypotheses is true, or an event has occurred the probability of which is less than or equal to P for any hypothesis of the family. If the matter had not been confused by half-understood slogans it would be universally accepted that the level of significance of any test is set by the *greatest* frequency, among the family of hypothesis under consideration, with which the criterion is surpassed. The test put forward by Behrens[6] in 1929 has been becoming increasingly available as fuller tables for this test have been published. The only criticism to which this test has been exposed, and which in the more benighted circles has been regarded as fatal and final, has arisen from the fact that repeated sampling from population having different variance ratios (σ_1^2/σ_2^2) often surpass the criterion with a lower frequency than that for which the criterion was calculated or tabulated. I do not know what else the critics would expect: Some hypotheses when true, give less help than others towards testing the aggregate in which they are combined; it may be that some give no help at all.

TEACHING AND RESEARCH

I have given some instances in which the teaching of statistics has fallen back into grave confusion even while their applications have been becoming more widespread and more important. If reform were impossible in the theoretical ideas instilled in teaching departments, it would be impossible to prevent the spread of erroneous methods in the applied fields. I have no doubt, however, that throughout the world many statisticians in applied fields see the matter as I do, and that considerable resistance already exists to the use of such misleading numerical tables as No. 11 of Biometrica Tables, which is the first example of the the kind which has come to my notice.

I believe sanity and realism can be restored to the teaching of Mathematical Statistics most easily and directly by entrusting such teaching largely to men and women who have had personal experience of research in the Natural Sciences. At least there should be a nucleus of teachers with practical experience in all departments teaching statistical methods.

In regard to advances and extensions in mathematical methods two fields are likely to be fruitful. Many kinds of data do not seem capable of yielding exact probability statements about the appropriate aspects of the real world. It is easy usually to make inferential statements at a weaker level, as when we specify the *Likelihood Function*[2] of an unknown parameter, without being able to make any *probability* statement about it. Or, as in many important cases, we can make tests of significance of uniquely appropriate kinds, without their entailing any defensible probability statements. In this field the attention of mathematicians should be drawn to the task of specifying more carefully the various kinds of uncertainty which we encounter in ordinary life, and especially in applied mathematics. It is certain that the historical concept of Mathematical Probability only defines uncertainty of a particular kind, and the appropriate specification of other kinds, with an examination of their mathematical consequences, is a field widely open to exploration[3]

In the second place, although statisticians are familiar with a considerable range of examples in which exact probability statements are derivable from the observations, no one imagines that the class of cases in which this can be done has been exhausted. When, with two parameters, θ_1 and θ_2, it is possible to find an unconditional or marginal distribution for θ_2, valid irrespective of θ_1, and also to find a conditional distribution for θ_1, for any given θ_2, it is obvious that the simultaneous distribution of the two parameters has been obtained. This implies the existence of a Sufficient or exhaustive estimate of θ_2, with distribution independent of θ_1, in addition to such an estimate of θ_1, when θ_2 is given.

In other cases[2] it may be possible by a direct argument to establish the probability of a simultaneous inequality

$$Pr\{\theta_1<\alpha,\ \theta_2<\beta\}=P(\alpha,\ \beta,\ T_1,\ T_2),$$

where T_1 and T_2 are jointly Sufficient for the estimation of θ_1 and θ_2.

In general the probability found in this way must involve all the independent elements of an exhaustive set of statistics, for if any were omitted its numerical value would serve to identify a sub-set of possible observations not irrelevant to the probability statements about θ_1 and θ_2. It would be too strict, however, to say that exhaustive simultaneous estimation of θ_1 and θ_2 was a *necessary* condition. In fact even if no exhaustive *estimates* exist, but if the set of statistics T_1, T_2 and T_3 is *jointly* exhaustive, there is no cogent reason to exclude the possiblity of such probability statements as

$$P_r(\theta_1<\alpha_1,\ \theta_2<\beta)=F(\alpha,\ \beta,\ T_1,\ T_2,\ T_3),$$

even when no function of these three statistics has a sampling distribution independent of θ_1, θ_2, as with Ancillary Statististics Consequently, we may anticipate that the possibility of establishing the simultaneous distribution of two or more parameters, extends beyond the limitation to simultaneously exhaustive *estimation.*

As I have already spoken at some length, perhaps you will forgive me if I do not on the present occasion enter into the detailed mathematics any such example. I shall hope to expand these matters further in Conferences with my mathematical colleagues.

REFERENCES

(1) T. Bayes (1763) An essay towards solving a problem in the doctrine of chances. *Phil Trans.* v. 53, p. 379

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

(3) R. A. Fisher (1957) The underworld of probability. *Sankhya* v. 18, pp. 201—210.

(4) J. Neyman (1938) L'estimation statistique traité comme un problème classique de probabilité. Actualities sci. indust. v. 739, pp. 54—57.

(5) E. S. Pearson (1947) The choice of statistical tests illustrated on the interpretation of data in a 2×2 table. Biom. v. 47, pp. 139—163.

(6) W. U. Behrens (1929) Ein Beitrag zur Fehlen-Berechnung bei wenigen Beobachtungen. Lander. Yb. v. 68 pp. 807—837.