109

Probability Likelihood and Quantity of Information in the Logic of Uncertain Inference

By R. A. Fisher, F.R.S.

(Received December 14, 1933)

In a previous paper H. Jeffreys* put forward a method of obtaining the distribution *a priori* of the precision constant of a hypothetical normal distribution, by means of the principle that if three independent observations are made in succession, from a continuous distribution of any form, the probability that the third observation shall fall between the first two must be one-third (p. 48): "Two measures are made. What is the probability that the third observation will lie between them? The answer is easily seen to be one-third."

This proposition, in the form in which Jeffreys states it as the foundation for his deductions, is ambiguous, and may bear one of two distinct meanings, one true and the other demonstrably false. The proposition may mean :—

(*a*) If sets of three independent observations are taken from any continuous distribution, the probability that the third observation of any set shall lie between the first two of the same set is one-third.

It is obvious that this proposition is true, since the six orders in which any three specified observations may occur in a set will be realised in different sets in the long run with equal frequency, and in two of these six possibilities the observation with median value will occur last. The probability of any two observations coinciding in value is, of course, zero.

(*b*) If the first two observations are the same for all sets, and a third observation be chosen at random independently for each set, the probability that the third observation shall lie between the first two is one-third, for all values of the first two observations.

* 'Proc. Roy. Soc.' A, vol. 138, p. 48 (1932).

This proposition (*b*), which is that used by Jeffreys in his derivation of the form of the distribution *a priori* of the constant of precision, may be very easily shown to be untrue, save in an exceptional case of zero probability.

If we represent (fig. 1) the probabilities of an observation chosen at random being less, respectively, than the first and the second observation, by the co-ordinates of a point P, then it is easy to see that P will lie with equal probability within any two regions of equal area inside the unit square, for which both co-ordinates lie between 0 and 1. The probability of a third observation lying between the first two will then be the absolute value of the difference between the co-ordinates. This difference will exceed any chosen value, p, in two regions at opposite corners of the square, the aggregate of which is $(1-p)^2$. Consequently, in one case out of four p will exceed $\frac{1}{2}$, in four cases out of nine p will exceed $\frac{1}{3}$, and in seven cases out of sixteen p will be less than $\frac{1}{4}$. Obviously, also, the chance of p lying in the range $\frac{1}{3} \pm \frac{1}{2}dp$ will be $\frac{4}{3}\,dp$, and will tend to zero as dp is decreased indefinitely. The probability, p, is therefore only exceptionally in the near neighbourhood of $\frac{1}{3}$, and in general it takes all values from 0 to 1 with a calculable frequency. It is only its average value that is equal to $\frac{1}{3}$.

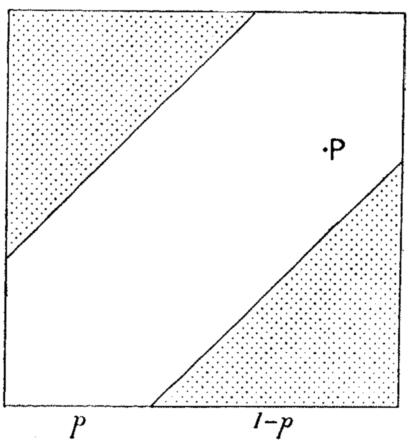

FIG. 1.

In another paper* the author has shown that when, in Jeffreys' analysis, allowance is made for variation in the values of the first two observations by integrating their frequencies over the possible range of these values the equation arrived at reduces to a mere identity. This was, indeed, to be expected, when a fact, true equally of all the distributions under discussion, is adduced to discriminate the probabilities of each separately.

In a rejoinder,† Jeffreys objects to this process of integration :—

> "Fisher proceeds to reduce my theory to absurdity by integrating with respect to all values of the observed measures. This procedure involves a fundamental confusion, which pervades the whole of his statistical work, and deprives it of all meaning."

* 'Proc. Roy. Soc.' A, vol. 139, p. 343 (1933).

† 'Proc. Roy. Soc.' A, vol. 140, p. 532 (1933).

Any defence which Jeffreys might have to offer of his omission to perform these integrations is thus lost in a polemical haze which his subsequent paragraphs do nothing to elucidate. I am not inclined to deny that the integrations reduce Jeffreys' theory to absurdity. Their purpose, however, is merely to justify the principle on which Jeffreys' reasoning is avowedly based, *i.e.*, to draw the conclusions derivable from the true proposition (a) above, in place of those which Jeffreys has derived from the untrue proposition (b).

This point can, I hope, be made plain independently of any general criticism of the system of notions respecting probability, which Jeffreys has elsewhere developed, and which he has reiterated in his rejoinder to my note. Since, however, he seems to complain of my neglect of these notions, I may be permitted to put forward as briefly as possible the reasons which have weighed with me in this neglect.

Criticism of Jeffreys' Theory

Jeffreys' definition of probability is subjective and psychological*: "We introduce the idea of a relation between one proposition p and another proposition q, expressing the *degree* of knowledge concerning p provided by q." In this it resembles the more expressive phrase used by Keynes, "the degree of rational belief." Obviously no mathematical theory can really be based on such verbal statements. Any such theory which purports to be based upon them must in reality be derived from the supplementary assumptions and definitions subsequently introduced, a "series of conventions, involving no further hypotheses," in Jeffreys' explanatory phrase. Thus Keynes establishes the laws of addition and multiplication of probabilities, by stating these laws in the form of definitions of the processes of addition and multiplication. The important step of showing that, when these probabilities have numerical values, "addition" and "multiplication," as so defined, are equivalent to the arithmetical processes ordinarily known by these names, is omitted. The omission is an interesting one, since it shows the difficulty of establishing the laws of mathematical probability, without basing the notion of probability on the concept of frequency, for which these laws are really true, and from which they were originally derived.

The alternative method of bridging this gulf is adopted by Jeffreys: "The fundamental rule is the Principle of Non-sufficient Reason according to which propositions mutually exclusive on the same data must receive equal probabili-

* 'Proc. Roy. Soc.' A, vol. 140, pp. 527-8 (1933).

ties if there is nothing to enable us to choose between them." It will be noticed that the idea that a probability can have an objective value, independent of the state of our information, in the sense that the weight of an object, and the resistance of a conductor have objective values, is here completely abandoned. The ideas, familiar to all writers on mathematical probability, that a probability may in certain circumstances be unknown and in other circumstances may be known with greater or less accuracy, are quite foreign to Jeffreys' system. His rejection also of frequency as an observational measure of probability (" By ' probability ' I mean probability, and not frequency, as Fisher seems to think," p. 523) makes it impossible for any of his deductions to be verified experimentally.

The one merit of a system of thought, founded on Non-sufficient Reason, and denied access to experimental verification, might be its internal consistency. As a succession of writers has shown, however, this supposed principle leads to inconsistencies which seem to be ineradicable, as in the example which Jeffreys quotes from Keynes :—

> " Keynes* writes as follows : ' Let us suppose as before that there is no positive evidence relating to the subjects of the propositions under examination which would lead us to discriminate in any way between certain alternative predicates. If, to take an example, we have no information whatever as to the area or population of the countries of the world, a man is as likely to be an inhabitant of Great Britain as of France, there being no reason to prefer one alternative to the other. He is also as likely to be an inhabitant of Ireland as of France. And on the same principle he is as likely to be an inhabitant of the British Isles as of France. And yet these conclusions are plainly inconsistent. For our first two propositions together yield the conclusion that he is twice as likely to be an inhabitant of the British Isles as of France.
>
> " Unless we argue, as I do not think we can, that the knowledge that the British Isles are composed of Great Britain and Ireland is a ground for supposing that a man is more likely to inhabit them than France, there is no way out of the contradiction. It is not plausible to maintain, when we are considering the relative populations of different areas, that the number of *names* of sub-divisions which are within our knowledge, is, in the absence of any evidence as to their size, a piece of relevant evidence.' "

* *Treatise on Probability*, p. 44, 1921.

Jeffreys' attempt to rebut this argument is as follows :—

> " Keynes here commits the fallacy, against which he argues effectively elsewhere, of supposing that the probability of a proposition is a function of that proposition and nothing else, instead of an expression of our state of knowledge of the proposition relative to particular data. Suppose, to make the issue a little more precise, that a man in Buenos Aires receives a message to the effect that a European of unspecified nationality is coming to visit him. He must then assess the probability of the various possible nationalities with respect to his available knowledge. If his data are that Great Britain, Ireland, and France are three different countries, and he has no further information as to the number and mobility of their inhabitants, he must assess their probabilities equally, and the probability that the visitor comes from Great Britain or Ireland *is* twice the probability that he comes from France. If, on the other hand, he considers that the British Isles are one country, of which Great Britain and Ireland are divisions, he must assign to the British Isles and France the same probability, dividing that assigned to the British Isles equally between Great Britain and Ireland. Keynes's dilemma does not exist and is merely an indication of incomplete analysis of the nature of the data."

It will be observed that Jeffreys' defence of his principle depends wholly on the verbal use of the word " country," and has been anticipated and answered by Keynes in the passage quoted, in a way that Jeffreys seems to overlook. Even as a formal solution based on a verbal convention, however, Jeffreys' solution breaks down in the case where the information in the possession of the Argentino is that the British Isles are one country, in one sense of that word, and two countries in another sense in which the word might be employed; and where he is intelligent enough to recognize that his preferences, if any, among the different definitions of the word " country " are irrelevant to the probability which is under his consideration.

The failure of all assumptions of the same nature as the principle of insufficient reason in logical situations, in which the subject is not only ignorant of the relative probabilities of the different hypotheses he might make, but at the same time knows of his own ignorance, leads naturally to the consideration that the logical situations in which uncertain inference may be attempted are various and diverse in character, and that an initial mistake is introduced in all such definitions as that of Jeffreys in assuming that the degree of knowledge or degree of rational belief is, in all cases, measurable by a quantity of the same kind. In Jeffreys' definition, indeed, it is evident that at least two different kinds of quantities are admissible :—

(i) We may consider the amount of information which the proposition, or set of propositions, q, has to offer respecting the truth or falsehood of p; this is evidently a different quantitative element in the logical relationship from

(ii) The extent or degree to which the information provided by q favours the truth rather than the falsehood of p.

In the logical situation presented by problems of statistical estimation, I have shown that a mathematical quantity can be identified which measures the quantity of information provided by the observational data, relevant to the value of any particular unknown parameter. That it is appropriate to speak of this quantity as the quantity of information is shown by the three following properties :

(i) The quantity of information in the aggregate of two independent sets of observations is the sum of the quantities of information in the two sets severally; each observation thus *adds* a certain amount to the total information accumulated.

(ii) When, on increasing our observations, the sampling error of an efficient estimate tends to normality, the quantity of information is proportional to the precision constant of the limiting distribution.

(iii) The quantity of information supplied by any statistic or group of statistics can never exceed the total contained in the original data.

Even if, in a logical situation providing a basis for uncertain inference, we confine attention to quantitative characters, measuring the extent to which some inferences are to be preferred to others, different situations, among the kinds which have already been explored, provide measures of entirely distinct kinds. Thus, a knowledge of the construction and working of apparatus, such as dice or roulettes, made for gaming, gives a knowledge of the probabilities of the different events or sequences of events on which the result of the game may depend. This is the form of uncertain inference for which the theory of probabilities was developed, and to which alone the laws of probability are known to apply. Since the term "mathematical probability" and its equivalents in several foreign languages have been used in this sense, and almost exclusively in this sense, for over 200 years, it is impossible to accept Mr. Bartlett's* suggestion, in his thoughtful discussion of the topic, that only the word "chance" should be used for the objective probabilities with this meaning, and that the word "probability" should be confined to the recent and perhaps ephemeral meaning which Dr. Jeffreys has assigned to it.

It is difficult to understand the difficulty expressed by Jeffreys as to the definition of probability, when incommensurable, as the limit of the ratio of two numbers, when these both become infinite or increase without limit. All the sampling properties of hypothetical infinite populations can be expressed rigorously as limits of the sampling properties of finite populations if, as these are increased indefinitely, the frequency ratios of their elements tend to the values assigned in the hypothetical infinite population. This is quite another matter from the difficulty experienced by those who attempted to define probability as the limit of the frequency ratio of experimental events, for we can have no direct knowledge of the existence, or the nature, of the limits approached when any experimental procedure is repeated indefinitely. In contrast, the limits approached by repeating mathematical operations may be investigated with precision.

The logical situation which arises in the Theory of Estimation is of quite a different character. Here we are provided with a definite hypothesis, involving one or more unknown parameters, the values of which we wish to estimate from the data. We are either devoid of knowledge of the probabilities *a priori* of different values of these parameters, or we are unwilling to introduce such vague knowledge as we possess into the basis of a rigorous mathematical argument. Knowledge *a priori* may be, and often is, used in arriving at the specification of the forms of population we shall consider. The chief logical characteristic of this line of approach is that it separates the question of specification from the subsequent question of estimation, which can arise only when a specification is agreed on.

When a definite specification has been adopted, we can obtain a function of the parameters proportionate to the probability that, had these been the true values, the observations would have been those actually observed. This function is known as the mathematical likelihood of any value of a single parameter, or of a set of values, if the parameters are more than one. With respect to the parametric values the likelihood is not a probability, and does not obey the laws of probability. Maximizing the likelihood provides a

* 'Proc. Roy. Soc.' A, vol. 141, p. 518 (1933).

method of estimation which has been shown to possess the following relevant properties :—

(i) In certain cases an estimate is possible which, even from finite samples, contains the whole of the information contained in the sample. Such estimates are known as "sufficient." The method of maximum likelihood provides such sufficient estimates when they exist.

(ii) When no sufficient estimate is possible, and there exist only estimates which conserve a fraction of the total information, tending to unity as the sample is increased indefinitely, the value with the highest likelihood is one such estimate, and contains not less information than any other estimate of the same kind.

(iii) When the likelihood function is differentiable at its maximum, ancillary statistics may be formed from its successive differential coefficients, which reduce the amount of information lost, as the sample is indefinitely increased, to zero of any required order.

(iv) In certain instances, as I have more recently shown, the whole of the information contained in the sample may be recovered by using the whole course of the likelihood function.

Thus we have, in addition to the probability of the classical theory, already two other quantitative characteristics, appropriate to different logical situations admitting of different sorts of uncertain inference. It is to be anticipated that a detailed study of logical situations of other kinds might reveal other quantitative characteristics equally appropriate to their own particular cases. However reasonable such a supposition may have appeared in the past, it is now too late, in view of what has already been done in the mathematics of inductive reasoning, to accept the assumption that a single quantity, whether "probability" or some other word be used to name it, can provide a measure of "degree of knowledge" in all cases in which uncertain inference is possible.

Jeffreys attempts the more difficult task of justifying our procedure in arriving at particular specifications by means of the Theory of Probability. It is not, however, obvious that probability provides our only, or chief, guide in this matter. Simpler specifications are preferred to more complicated ones, not, I think, necessarily because they are more probable or more likely, but because they are simpler. As more abundant data are accumulated certain simplifications are found to be very unlikely, or to be significantly contradicted by the facts, and are, in consequence, rejected ; but among the theoretical possibilities which are not in conflict with any existing body of fact, the calculation of probabilities, even if it were possible, would not, in the writer's opinion, afford any satisfactory ground for choice.