ON THE RECONCILIATION OF THEORIES OF PROBABILITY

By M. G. KENDALL

Introduction

1. Few branches of scientific method have been subject to so much difference of opinion as the theory of probability. Even when we put aside numerous points of taste in presentation or axiomatization there remains a stubborn residual variance of viewpoint between different authorities. Everyone agrees that this is undesirable; nobody yet, I think, has dared to maintain that it is avoidable. But that is the theme of the present article. I wish to show that there is nothing necessarily incompatible in the varying views which are currently held; that the authorities are either saying the same thing in different ways or can only disagree because of avoidable latent differences in their premisses or their field of discussion. History, both past and present, teaches us that the role of mediator between contestants is neither safe nor profitable, and I am quite prepared to incur general disfavour for maintaining that most authors have some of the right but that none has a monopoly of it. In plunging into a controversial subject one must run the risk of controversy. I can only say that I have tried not to excite it.*

2. It is convenient to consider the subject under three heads, Foundations, Direct Theory and Inverse Theory. The greatest differences of opinion (or, at least, those which have generated the most heat) have arisen in the third, but there are substantial differences concerning the first. Even the second, as I shall point out, is not free from problems of a character similar to those arising in the other two, though the fact is not normally given much prominence and perhaps has not been generally appreciated.

Frequency and non-frequency theories

3. Although there are many shades of opinion about the appropriate foundations of a theory of probability we may broadly distinguish two main attitudes. One takes probability as 'a degree of rational belief', or some similar idea, and in enunciating the foundations of the subject does not attempt to analyse it into simpler ideas; it is therefore necessary to agree upon certain axioms and postulates concerning probability itself before a definite theory can be founded. The second defines probability in terms of frequencies of occurrence of events, or by relative proportions in 'populations' or 'collectives'; and attempts to base the theory on familiar concepts without invoking a special indefinable 'probability'. In practice both approaches seem to lead to the same kind of direct theory, and it would be rare for exponents of the two to reach different mathematical conclusions from the same premisses. But this pragmatic reconciliation, though perhaps comforting to the onlooker, is not enough. The difference of viewpoint runs down to the heart of the subject and must be resolved if possible, not merely for intellectual comfort, but because it affects the theory of inference in which different authorities sometimes do reach different conclusions.

* The first personal pronoun occurs more frequently in this article than good taste in objective scientific writing would normally permit. The reason is that I wish to leave no doubt about which statements are advanced as personal views.

4. The first point to establish is that the current methods of axiomatizing the *calculus* of probabilities are not sufficient to axiomatize a *theory* of probability. The pure mathematics of a calculus of probabilities proceeds from certain basic rules (such as those governing the addition or multiplication of probabilities of independent events) which are usually laid down without much discussion of their rationale. The simplest approach is exemplified by the procedure adopted in most text-books of algebra, wherein the probability of success of an event which can happen favourably in m out of n possible ways is defined as the ratio m/n. A more sophisticated approach is used by Kolmogoroff (1933) and Cramér (1945), for example, by relating probability to the measure of a set of points. This is probably adequate for the mathematician, who is more concerned with working out the logical consequences of his assumptions than with relating this calculus to the physical world. It fails, however, to solve the problem with which we are here concerned, namely, to found a theory of probability as a branch of scientific method. In other branches of science it may be enough to set up a mathematical model and to accept it if it gives a reasonably good account of observation; but in the most general sense we are here considering how good a 'reasonably good' agreement must be.

5. It has sometimes been suggested that the domain in which the statistician is interested is only part of the whole domain covered by the phrase 'uncertain inference'; that, for example, other people may reasonably be concerned with gauging the degree of doubt in propositions concerning unrepeatable events such as 'Homer was blind', whereas the statistician operates with series of similar propositions or events because his science is the study of aggregates. The suggestion is, I think, that the statistician can avoid some of the difficulties of uncertain inference by confining himself to classes of propositions concerned with experimentally repeatable events. It might even be questioned whether 'probability as the statistician uses the word has the same connotation as when it is used by logicians or scientists in relation to individual hypotheses. Perhaps the psychologist can answer this question for us. My own opinion is that there is no essential difference (other than that of intensity) between the attitudes of doubt towards any propositions of whose truth we are uncertain. Something may be said for distinguishing those hypotheses which are capable of experimental verification from those which are not; it might even be (though I do not believe it) that different kinds of uncertain inference are appropriate to the two cases; but it does not appear to me that the statistician solves any essential problems by confining himself to special classes of case. In practice he is concerned, like any other scientist, with the formation of opinions and the taking of decisions on incomplete evidence, and the mere fact that some of his data can be counted or measured does not, I maintain, relieve him of any major problem arising in the justification of his inferential processes.

6. *I assert that any theory of probability which does not take probability itself as a primitive idea must, in some form or other, introduce an equivalent primitive before it can be applied.*

The rather naïve approach of the algebraic text-book, for example, may take one of two main forms: it can state that if, of a set of n mutually exclusive propositions, m are favourable, the probability that a favourable proposition is true is m/n; or it can state that if m out of n *equally probable* and mutually exclusive propositions are favourable and one must be true, the probability of a favourable proposition is m/n. The first form is untrue (a man can be unidextrous in two ways, as right- or left-handed, but the probability is not $1/2$ that a man is right-handed in any theory that I know); the second begs the question by the use of the expression 'equally probable'. There are, of course, variants on the way in which this idea is put; we may speak of events instead of propositions, or refer to events happening 'at random'. But they all come to much the same thing so far as concerns the point now under discussion. The concept of probability contains more than the mere idea of proportionality of cases. And, of course, the same argument applies when proportionality is replaced by the measure of a set or some more refined mathematical concept.

7. The mathematician may fairly contend that such matters are not his concern, any more than the Euclidean geometer is concerned with the question whether there exist in practice objects which can be regarded as straight lines. One must be a little careful about arguing from the analogy of other branches of applied mathematics because at this stage we are concerned not only with the relationship of theory with the external world but also with the relationship between our calculus and the way we think. Even if we pass over such a point, however, and admit the mathematician's right to deny interest, *qua* mathematician, we must accept responsibility for resolving the difficulty in some other capacity, as psychologists, logicians, or statisticians, unless we are prepared to relegate probability to the domain of pure mathematics and the possibility of its application to the phrontisteries of our academies.

8. Von Mises (1928) has made a valiant attempt to provide a frequency theory of probability, and for a time I was almost convinced that he had succeeded. Myself when young did eagerly frequent. But his theory fails, in my view, in much the same way as the other theories mentioned in § 6. He has, in fact, to introduce the idea of an 'Irregular Kollektiv' (English writers would call it a random series) or a 'Prinzip des ausgeschlossenen Spielssystems'—the impossibility of a winning system in games of chance. The 'Irregular Kollektiv' itself is a new concept outside ordinary mathematics. Attempts by various writers to show that there exist sequences of numbers with the necessary properties break down, I hold, on the point that such sequences can only be shown to obey an enumerable set of conditions, whereas the 'Kollektiv' must obey an innumerable set (unless the definition is to be modified, in which case we arrive at the repugnant conclusion that the laws of probability may be obeyed sometimes, but not always and not even, relatively speaking, frequently).

9. I myself do not believe that anyone will ever succeed in producing a theory of probability which does not, at some point, require a primitive idea equivalent to that of probability itself. It is necessary, I hold, to have some concept of randomness, haphazardness or uncertainty of happening inherent in the subject which either must be introduced explicitly into the axiomatization or, if omitted, must be introduced later before the theory is capable of application. Precisely where the idea is introduced is to some extent a matter of taste or didactic convenience. But appear it must. There can, I assert, be no pure frequency theory of probability any more than (to anticipate a later argument) there can be a useful objective theory of probability without reliance at some point on empirical justification in terms of frequency.

10. One point, however, is worth examination at this stage. Must we take as our primitive a general idea of probability which permits of the comparison of any two propositions, or is it sufficient to use only the idea of equal probability? The axiomatization referred to in § 6 suggests that equi-probability is sufficient if we can analyse our alternative possibilities to the point where they all stand on the same footing. If we can, as it were, break down our situation into atomic propositions with equal probability, a mathematical measure of the probability of a subset follows readily enough.

11. At first sight this does not appear to be possible. The probability that a new-born child is male is (say) 0·51 but it does not look as if we can regard this as one of a subset of

51 favourable propositions out of 100 equi-probable propositions. But I think this is not a fatal objection. We arrive at the probability of 0·51 by counting cases, and it is assumed in doing so that the probability of occurrence of a male in each case is the same; in fact, we *do* base our probability on equi-probable events, and the same is true of any probability based on statistical frequencies. I am not prepared to say that we can in every instance schedule the equi-probable events explicitly. General judgements in probability often have to be made on the basis of a 'feeling of the situation' which is compounded of a multitude of relevant factors half-remembered or not explicitly remembered at all. There is, as Jeffreys has pointed out, an element of uncertainty attributable to the imperfection of the human mind itself. I am, nevertheless, inclined to think that a satisfactory theory can be founded merely on the notion of equi-probability or the equivalent notion of randomness. But it is not necessary to insist on the point. What I do insist upon is the necessity for a primitive idea of probability or randomness of some kind.

12. It might be thought that the differences between the frequentists and the non-frequentists (if I may call them such) are largely due to the difference of the domains which they purport to cover. *I assert that this is not so.* One of the principal modern advocates of the non-frequency approach is Jeffreys, who takes probability as a measure of belief and is concerned with the application of his theory to scientific inference in general. But practically every example in Jeffreys's book can be treated by the methods of the frequentists (or so they would claim). The fact is that in practice both schools deal with the same kind of problem. They differ because they approach the same problems differently, not because they deal with different problems.

13. The essential distinction between the frequentists and the non-frequentists is, I think, that the former, in an effort to avoid anything savouring of matters of opinion, seek to define probability in terms of the objective properties of a population, real or hypothetical, whereas the latter do not, and, indeed, sometimes go further and repudiate the introduction of a population as irrelevant, incompetent and immaterial. To revert to the example of the sex-ratio in births, the frequentist would, at the outset of the inquiry, postulate the existence of a probability p that a birth was male in some population or 'Kollektiv' and then proceed to *estimate* it in the light of experience. The non-frequentist would begin by assuming a prior probability for the ratio and would then modify it in the light of the posterior probabilities given by observation. If the observations are numerous his prior probabilities dwindle into insignificance and he gets much the same formula as the frequentist. But he is not, on the face of it, trying to do the same thing. The frequentist is estimating an unknown constant; the non-frequentist is determining a probability which is not a constant, but varies according to the state of his knowledge.

14. Suppose I draw a random example of 1000 births and find that 600 are male. Do I then infer (since this differs significantly from what could happen, to an acceptable degree of probability, in sampling from a population wherein the proportion of males is 51 %) that the proportion of males cannot be 51 %? Most assuredly I do not. There is so much prior evidence in favour of a ratio of about 51 % as against a ratio of 60 % that I require much more posterior evidence than this to shake my prior probability seriously. My previous knowledge must affect my judgement. The frequentist cannot, I think, question this. He must then do one of two things. He must either admit that probability is influenced by prior knowledge if that probability is *by itself* to be the basis of a rational judgement or course of action; or he must concede that the final judgement is based on a mixture of probability and some

different quality of uncertain inference which conditions it. It seems to me that if he takes the second course he lays himself open to the criticism that his theory fails to solve the problem with which he is concerned. One does not obtain an objective theory of inference by separating it into an objective and a non-objective part and ignoring the second. Perhaps the frequentist can maintain that his theory is consistent and logical. He cannot maintain that it provides more than a part of the answer to the fundamental question.

15. I shall revert to this point later in considering the statistical theory of confidence intervals. It is sufficient for the moment to record the point that in arriving at a probability of an observed event even the frequentist must make some assumption about prior probabilities. This may, in the last analysis, mean no more than that he has to postulate of his observations that they are conforming to a random process of specified type; but that itself is an assumption at least of equi-probability, which is an assumption concerning prior probabilities. *I therefore assert that, to assign a probability in any practical case the frequentist as well as the non-frequentist requires a prior probability distribution from which to start.*

16. This raises two problems: (*a*) how, in any given situation, we determine the prior probabilities, and (*b*) how we justify the contention that what we are doing has any application in real life.

The precise determination of prior probabilities may be a matter of practical difficulty, but I do not think this constitutes an objection. We may have a fair idea of the temperature of a room without being able to express our impression very precisely in degrees on a thermometric scale; and similarly, we may have a rough idea of measurable prior probabilities without being able to say exactly what they are. But if we pursue these general impressions back to their source, how do they arise? Not, I think, by any innate knowledge, if there is such a thing, but from even more prior probabilities. The prior values which we assign on any given occasion are themselves posterior values of a previous experience. Where did they start?

17. It is difficult (but not impossible) for an adult to find any situation in which he has no prior knowledge whatever to influence his assessment of a probability. There may, however, have been moments in his childhood when his ignorance was complete. This is not, so far as I can see, necessarily so because in the very act of learning the meaning of a proposition he may have acquired some expectation as to its truth. It is for the psychologist to explore these topics. For my present purpose I need only emphasize that prior probabilities are themselves built up from experience, albeit an unremembered and unconscious experience.

18. Jeffreys and some other writers have endeavoured to meet the 'situation of initial ignorance' by laying down certain rules determining the types of prior probability distribution to be adopted in specified cases. When Jeffreys's book first appeared this seemed to me the least convincing part of his treatment, and it seems so still. In fact, my doubts increase as the rules proposed for different situations multiply. It cannot be necessary to adopt peculiar rules for prior distributions in order to say that we know nothing. As I understand him Jeffreys argues that his rules are necessary in some cases because otherwise there would arise mathematical difficulties. But this seems to me a circular argument, and more a ground for examining the mathematical treatment than a justification of the theory. In reading some of the papers in this vein one sometimes has difficulty in resisting a cynical suspicion that certain kinds of distribution are introduced because they give the right kind of answer, which may be a posterior justification but is unsatisfactory in laying down the bases of a subject.

19. *I assert:*

(*a*) *That the assignment of a probability to observed phenomena requires in all cases the assignment of a prior probability or some equivalent procedure such as the assumption that the generating process is random.*

(*b*) *That situations rarely, if ever arise, in which there is no knowledge of prior probabilities.*

(*c*) *But that, if such a situation arises, the only possible rule to use is that of Bayes in which all the possibilities are given the same prior probability.*

Let me add two scholia:

(i) Some difficulties arise over the mathematical treatment when we are considering parameters which may have an infinite range or even when they are continuous over a finite range. These difficulties, I hold, are mathematical in the sense that ideas of continuity and limiting processes are mathematical. They raise problems, but the existence of such problems is not very relevant to the basis of the theory of probability.

(ii) As data accumulate the posterior probability dominates the total probability so that it makes very little difference what the prior probabilities were. This is an argument for using prior distributions which are plausible but inexact if they make the mathematics easier, at least for large samples. It in no way affects the points now under discussion.

20. So far my summing up has been in favour of the non-frequentists on the points concerning the impossibility of founding the theory without some primitive idea of probability and the necessity of using prior probabilities. I must now restore the balance, and show that the non-frequentist requires notions of frequency in some form or other to make his theory of any practical importance if it is to be a numerical theory.

The frequentist tries to give his probabilities objectivity by basing them on frequency of occurrence. The non-frequentist would also like to obtain objectivity, of course, but he is concerned with the measurement of an attitude of mind which is subjective. He may, as Keynes did, beg the question by speaking of degrees of rational belief as if rational minds could not differ in the assessment of a probability on the same evidence, which in the ultimate analysis implies that a belief founded on the same data is only rational if it agrees with one's own. Jeffreys makes a better attempt, I think, by admitting that there is no logical answer to the solipsist but offering to believe in him on the basis of reciprocal aid. However, one must not make too much of this kind of point. We all assume that there are certain rules of thought which are common to all rational human beings. Let us then suppose that the rules of probability are commonly accepted. Let us further suppose that on given data two different individuals will arrive at the same assessment of prior probabilities. This is a very considerable assumption to make because the 'background' of individuals may vary so much that we cannot be sure that the data can ever be the same; consequently each individual may possess his own schedule of probabilities, and we do not emancipate ourselves from matters of opinion. However, in order to press on to the main difficulty, let us make the assumption. We then arrive at the fundamental question: what use is it to calculate a probability?

21. There is no point in calculating probabilities as an arithmetical exercise unless we are willing to relegate the theory of probability to the position of branch of pure mathematics. They must either be pure measures of a mental attitude or they must correspond to something in the external world. If they are only measures of belief, they may still form the basis of rational actions and rational decisions. But why are they of any use in this respect unless our decisions or actions are improved by them—in fact, unless we are more often right by using them than by ignoring them? The frequentist does not encounter this difficulty, for

his probabilities purport to describe observed frequencies. The non-frequentist, it seems to me, needs a new basic assumption.

*I assert that, for a non-frequentist theory of probability to be applicable it is necessary to assume that propositions with greater probability are true more often in fact than propositions with less probability.**

22. Certain minds, including some types of trained scientific mind, will instantly revolt at any contention that there is a positive correlation between what we believe and what is true. The scientist, very properly, is trained to suspect even his own beliefs and, remembering that whole populations have believed that the world is flat or that heat is a fluid, is not even very impressed by an overwhelming body of collective opinion. And yet the number of instances in which people have believed something wrong is relatively small; and frequently the erroneous propositions which have been most widely believed are those whose meaning is in any case abstract and obscure. Every conscious movement I make is based on the belief that what has happened before will happen again. The use of language itself is a testimony to this belief, and anyone who denies it has to explain why he thinks that the sounds he utters or the marks he inscribes on paper in making the denial will be understood.

23. I am not arguing that there is any less reason than in the past to seek for evidence, and as much of it as one can get, before coming to conclusions. I am only saying, first, that in many ordinary decisions we already have strong prior probabilities, and secondly, that we may have to take decisions on very slight evidence or on prior probabilities which only just favour one course as against the alternative; and I ask, what grounds have we for supposing that our probabilities are a good guide to conduct? It is here, I think, that the frequentist can assert himself; for the only grounds are that we have found in the past that it is so. Our reliance on our probabilities is based on the frequency with which we have found it justified in the past. This, I think, is the pragmatic frequentist sanction for a non-frequentist theory. It works.

24. The line of thought which leads me to suggest that a reconciliation of the frequentist and non-frequentist views is possible should now be clear. The frequentist seeks for objectivity in defining his probabilities by reference to frequencies; but he has to use a primitive idea of randomness or equi-probability in order to calculate the probability in any given practical case. The non-frequentist begins by taking probability as a primitive idea, but he has to assume that the values which his calculations give to a probability reflect, in some way, the behaviour of events. *Frequentiam furca expellas, tamen usque recurret.* Neither party can avoid using the ideas of the other in order to set up and justify a comprehensive theory. I believe that if this is firmly grasped the fundamental differences vanish. There may still be room for argument about the best way of presenting the subject from the logician's or the teacher's point of view, but that is quite a different matter.

FOUNDATIONS

25. In laying down the collection of axioms, postulates and general rules-of-the-game which provide the foundations of a theory, we have a substantial amount of choice which is exemplified by the very different treatments given by different authors under the title of

* In this article I speak sometimes of probabilities of events, sometimes of propositions, sometimes of propositions that *are* true, sometimes of events that *will* happen. Such terms require definition and clarification in a highly rigorous exposition, but to have attempted such a thing here would have obscured the main points I am making. I am satisfied that in omitting such a treatment I am not glossing over any difficulties which affect my main conclusions.

'Theory of Probability'. The mathematician is apt to skate rather lightly over the problems of relationship with experience in order to press on to the calculus, which is his primary interest. The logician goes more deeply into the correct axiomatization but sometimes does not proceed to develop the theory to the point of showing its practical utility. One cannot complain of this, for only an encyclopaedist can do justice to the subject in its entirety, and even a collective work like Borel's *Traité* leaves large areas untouched. One can indeed complain of the lack of sympathetic understanding shown by some authors to the works of others; but it is unfortunately true that among the exponents of the science of objective judgement there seem to be as many emotional judgements as among less enlightened sections of the community.

26. The very extent of the domain to be covered requires an author to be selective, but there is a serious danger that his selection may give a bias to his treatment. For instance, it is a fairly common practice to begin with the throwing of dice or the tossing of coins as illustrations of the 'kind of situation' with which the probabilist is to deal, as if the theory of probability were the same thing as the doctrine of chances. It is not part of my purpose to discuss such matters on the present occasion, but it is fair to point out that a teacher ought to be very careful not to put only one side of the case to his students. Admittedly he must start with simple ideas and not destroy their self-confidence by leading into the difficulties at the beginning of the subject—as in mathematics itself, the fundamentals are the most difficult and the beginnings should come last. But in reading most modern treatments I cannot help feeling that a little more impartiality would be an improvement, and that an author should do more than ignore opposing views or mention them merely to refute them.

27. My discussion of the frequentist and non-frequentist viewpoints in §§ 3–24 has left me little more to say about fundamentals. I have only two points to make. The first concerns the introduction of probabilities associated with continuous variables. From the idea of the probability of events we can build up the idea of the probability of variate values, and hence the idea of a probability function in the discontinuous case. It is then tempting for the mathematician to jump to the continuous case without pausing very long for thought and to write such expressions as

$$dF = f(x)\,dx, \tag{1}$$

which purport to express the fact that the element of probability dF in the range dx is $f(x)\,dx$. A variate transformation is then made by writing $x = x(\xi)$, and we have

$$dF = f\{x(\xi)\}\frac{dx}{d\xi}d\xi, \tag{2}$$

expressing that the probability in the range $d\xi$ is $f\{x(\xi)\}\,dx/d\xi$.

The retention of the differential element is thus very important in the expression of a probability. It is well known that there is something arbitrary in the determination of probabilities in a continuum. But on occasion we ignore the differentials and concentrate on the frequency function f, as, for example, in methods involving the likelihood function. There are contexts where it is important to remember that likelihoods in this sense are not invariant under variate transformations, so that tests based on likelihood ratios contain an implicit assumption about the random process generating the observations.

28. My second point is that in setting up a theory the non-frequentist has to take one hurdle which the frequentist by-passes. It is necessary for him to establish that his probabilities are measurable on a numerical scale unless (like Keynes and some other writers

on logic) he is content to see his theory so indefinite that practical application is confined within very narrow limits. Most authors are prepared to make whatever assumptions are necessary to permit their probabilities to be numerically measurable. There is evidently some latitude of choice in postulates for the purpose. The simplest course is merely to postulate that probability *is* measurable. Jeffreys starts from a slightly anterior point and requires the axiom that of two probabilities p and q, p is either greater than, less than, or equal to q. This puts his probabilities in order; he requires a further axiom that if probabilities are put in order they can be associated by a $(1, 1)$ correspondence with a set of real numbers in increasing order. This ensures that there are enough numbers for the purpose, and he then calibrates his scale by reference to the theorem that of n equally probable and exclusive events the probability of a subset m is m/n, and hence arrives at the *theorem* that any probability can be expressed by a real number. There may be other ways of arriving at the same result, but it appears unlikely to me that any of them could avoid making assumptions equivalent to those of Jeffreys.

DIRECT THEORY

29. The direct theory of probability, as I am using the expression, consists mainly of the formulation of rules for building up from the probabilities of simple propositions the probabilities of more complex collections of propositions. So far as I know there are no serious difficulties or differences of opinion about the simpler rules such as those expressing the probabilities of the sum of propositions. Nor are there any problems, other than those of pure mathematics, once certain fundamental rules have been established. Such valuable and ingenious little books as Whitworth's *Choice and Chance* contain many problems, but they are all essentially mathematical. There is, however, one rule which offers a stumbling block and merits particular attention, namely, the product rule which is usually expressed in some form such as

$$\begin{aligned} P(pq \mid h) &= P(p \mid h)\, P(q \mid ph) \\ &= P(q \mid h)\, P(p \mid qh), \end{aligned} \tag{3}$$

that is to say, the probability of both p and q on data h is the product of the probabilities of p on h and of q on p and h (or, equivalently, of q on h and p on q and h). I find it a useful practice, whenever meeting with a new book on probability, to go straight to the derivation of the product rule. It provides a kind of touchstone for the author's whole treatment.

30. For the frequentist the derivation is simple. If, of a set of n equi-probable and exclusive propositions k are favourable to p and q, l to p and m to q we have

$$\frac{k}{n} = \frac{l}{n}\frac{k}{l} = \frac{m}{n}\frac{k}{m}, \tag{4}$$

which establishes the proposition. There can be little doubt, I think, that the simple arithmetical identities of equation (4) are the basis of our readiness to accept the product-rule as generally valid in any theory. The idea of equi-probability or randomness contains within it the product-rule for a finite number of alternatives. There may be difficulties in dealing with limiting cases where infinities are involved, but I think they are far from insuperable and need not be brought up here to confuse the issue.

31. The non-frequentist has a much harder task to establish the rule. Johnson (*Logic*, vol. 3) treats it as an axiom, and his view is important because he influenced both Keynes

and Jeffreys. Keynes himself, I think, fell into error on this point. He *defines* the product of probable relations by

$$P(p \mid qh)\, P(q \mid h) = P(pq \mid h), \tag{5}$$

where, it must be remembered, equation (5) is merely a rule for manipulating logical symbols, not the expression of numerical probabilities. Although there is some latitude of choice in which of our elementary propositions are introduced as definitions and which as postulates, it is repugnant to general expectation that an important rule such as the product-rule should be introduced by mere definition. I find Keynes's subsequent attempt to establish a numerical theory of probability hard to follow, and I cannot see that in his theory the product-rule states anything more than that we may multiply numerical probabilities when we may, which is true but carries us no further forward.

32. Jeffreys takes what I think is the only possible course and reverts to Johnson's treatment by taking (3) as axiomatic.

Now the product-rule is not obvious. We are therefore entitled to consider why it should be introduced at all. The somewhat cynical reply that without it we could not get the answers we want is insufficient; we still have to explain why we want that kind of answer. Here, again, I think the frequentist comes into his own. The only justification for the product-rule that I know is based on the fact that it can be established for sets of equi-probable exclusive alternatives as in equation (4). *I therefore assert that both frequentist and non-frequentist theories are compelled to rely, for the justification of the product-rule, on the properties of equi-probable exclusive alternatives.* We are, I hope, one stage nearer an understanding how the two approaches are interlocked.

Inverse theory

33. A precisian might reasonably contend that there is no such thing as 'inverse' probability, but the term is so widely used that one cannot disturb it. 'Probabilité des causes' comes rather nearer to expressing what we mean by it. The essence of the process lies in its attempt to reason from observation to the general law governing observation. In its broadest aspect it seeks to provide a science of induction.

34. The non-frequentist can set up a theory of inverse probability, or perhaps it would be better to say a theory of inference, without a great deal of difficulty. Some frequentists would deny this, but I think the majority of the grounds of their denial would be found, on analysis, to relate to the more basic problems referred to in the earlier parts of this paper. It is, for example, legitimate for the non-frequentist to consider the prior probability distribution of an unknown parameter (which may be a natural constant such as the mass of a proton) and to modify it in the light of experiment according to the relation

$$\text{posterior probability} = \text{prior probability} \times \text{likelihood}. \tag{6}$$

In this way the non-frequentist can proceed, by continued experiment if necessary, to concentrate his probabilities in narrowing ranges, or in short, to get continually nearer to the truth; *always provided that he can obtain a prior probability to start with.*

35. It is in the problem of determining the initial probabilities that most of the modern controversy has centred. In fact, at least one statistical school of frequentists has felt the difficulty so acutely that much of their work on inference has been devoted to finding methods which avoid the use of prior probabilities altogether. In doing so they encounter troubles of their own which I discuss below. Even the non-frequentists, however, are not free from

troubles in formulating rules for the determination of prior probabilities. I have already referred to the point in § 18. The postulate of Bayes, that if nothing is known about prior probabilities they are to be assumed equal, appears to me, after many years' reflexion, to be inevitably right. When we are in a genuine state of indecision we 'toss up for it'. But it is very rarely indeed that we are in a complete state of ignorance about the alternatives we are considering. The mere fact that we understand what we are talking about implies some kind of knowledge. We may, indeed, be unable to set exact values on our vague appreciation of the prior position, but that is not really the point.

36. In an endeavour to avoid Bayes's postulate several modern writers, notably Fisher, Neyman and E. S. Pearson, have propounded methods which are widely accepted in one form or another by statisticians. The methods do not always agree between themselves, and in one famous case—the Behrens test—give different results; but they have the common object of attempting to escape from the necessity of using prior probability distributions. I shall discuss, in that order, maximum likelihood, fiducial inference and confidence intervals with the object of showing that there is, in fact, no escape except by the introduction of new assumptions.

MAXIMUM LIKELIHOOD

37. If a probability function $f(x, \theta)$ depends on a single unknown parameter θ the principle of maximum likelihood enunciates that to estimate θ we should take that value which, for variations in θ, maximizes the likelihood of the observations, for instance, in a sample of n independent observations, the function

$$L = \prod_{j=1}^{n} f(x_j, \theta). \tag{7}$$

I have discussed this principle on a previous occasion (1940). All I need say here is

(a) that in large samples the method gives much the same results as the method of Bayes, as is emphasized by Jeffreys; and

(b) that the strong posterior recommendations of the method (many of which relate to large samples) do not obviate the necessity for recognizing that the principle of maximum likelihood constitutes a new postulate.

38. The principle, in fact, is not obvious in the sense of being immediately and intuitively acceptable. It states that we are to proceed on the assumption that the most likely event has happened. Now, why? Even if we ignore any prior knowledge, for which the principle makes no allowance, and assume for the sake of argument that the most likely event is the most probable event (an assumption which appears to me to involve Bayes's postulate), we know, or at least we believe, that the most probable event does not always happen. The justification for our principle can then only be that it leads us closer to the truth *on the whole* than other methods. This appears to me to be another form of the postulate which is referred to above as necessary for the non-frequentist theory, namely, that what we believe is, on the whole, true. *I therefore assert that the principle of maximum likelihood requires a new postulate which is, in the last analysis, equivalent to one of the assumptions required to validate the non-frequentist theory.*

39. Here I interpolate one comment of general application. In seeking for a satisfactory logical basis of modern methods, and in pointing out that some of them are not so firmly based as they seem, I am in no way trying to undermine the use of those methods. What I am trying to do is to bring to light the latent assumptions on which they are founded. The

same applies to the other methods which I consider below. I think it is desirable to make this statement, partly to disarm those writers who may jump to the conclusion that I am laying the axe to the roots of their work (which is not my intention), and partly because there is a danger that critics of the new methods may swing too far the other way, and in finding that there are legitimate grounds for doubt about the logical foundations, may be tempted to reject the methods altogether. English writers may not go to these extremes, but the following extract from Pietra (1948) will illustrate my meaning. Referring to some recent publications by Gini on inverse probability, Pietra says (p. 70): 'but there is no doubt now that [Gini's] revision [of the basis of probability in statistics] has brought to light the fallacious character of the numerous illusions on which the great success of the Anglo-Saxon development of our subject is based.* An Anglo-Saxon may be pardoned for feeling a little nettled by this kind of statement; but it represents a point of view which he would be wise not to ignore.

Fiducial inference

40. The so-called method of 'fiducial' inference also requires a new postulate. The results which it gives in some cases agree with those given by the theory of confidence intervals and by Jeffreys's form of non-frequentist probability, but since it can give results which are not true in the former and is mostly advocated by those who repudiate the latter it can hardly be regarded as equivalent to either. The essence of the fiducial process appears to be this: if a frequency function of a sufficient statistic t and a parameter θ is

$$dF = \frac{\partial F(t,\theta)}{\partial t}\,dt, \tag{8}$$

the fiducial distribution of θ is given by

$$dF = -\frac{\partial F(t,\theta)}{\partial \theta}\,d\theta. \tag{9}$$

This is used to give a range within which the value of θ may be regarded as lying to an acceptable degree of 'probability'.

41. The first thing to note about an expression such as (9) is that it has a differential element $d\theta$. To a frequentist this means nothing in terms of his probabilities because θ is an unknown constant and has no probability distribution other than the trivial one of unity when θ has the true value. The introduction of fiducial inference then implies not merely a new postulate about the behaviour of probability but a new kind of uncertainty. Exactly what this kind of uncertainty may be has never been explained, but it is not very unlike probability in the sense of degree of belief, at least in some contexts.

42. Secondly, the transition from (8) to (9) is not by any means an obvious process, and I cannot myself see by what argument it is to be supported. At the least it amounts to a new postulate. Whether it is acceptable is a matter of taste, which perhaps is not worth discussing. The important point is that if one follows R. A. Fisher in rejecting Bayes's postulate and the notion of prior probability in favour of the principles of maximum likelihood and fiducial inference, one is not making any economy in new concepts—in fact, the contrary. My personal feeling is that Bayes's postulate is much more acceptable than the fiducial postulate. There is another complication in the use of fiducial inference which equally affects confidence intervals and I may as well proceed to it at once.

* 'Ma è anche fuori di dubbio ormai che dalla revisione stessa è emersa la fallacia delle molte illusioni sulle quali fondava il suo maggiore successo l'indirezzo Anglo-Sassone della nostra disciplina.'

CONFIDENCE INTERVALS

43. The theory of confidence intervals is an ingenious attempt to set up a method of making statements in probability without the use of prior probabilities or Bayes's postulate, and equally without the use of an alternative postulate. It acknowledges the impossibility of making assertions that a parameter (an unknown constant) will lie, to specified degrees of probability, within specified limits, but meets the situation by substituting for such limits random variables whose values can be observed. It is then possible to choose two random variables t_1 and t_2, and to assert, with given probability of being right, that

$$t_1 \leqslant \theta \leqslant t_2,$$

whatever t_1 and t_2 actually turn out to be when the observations are made. Since it is possible, at least in some cases, to choose t_1 and t_2 to be independent of the unknown parameters, the probability of being right remains the same whatever values of θ are actually encountered in practice. The method thus appears to be independent of any prior distribution of θ, and to be quite independent of any approach such as is embodied in Bayes's postulate.

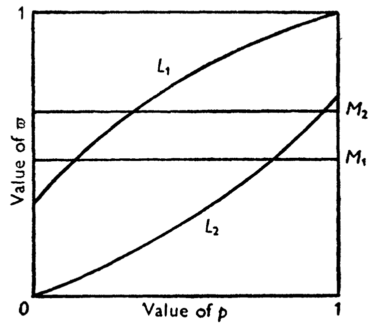

Fig. 1

44. Consider again the example referred to in § 14. Suppose I assume that a sampling process is such as to reproduce a binomial distribution—there is a good deal of evidence for this in the case of births. I observe a value of 0·60 as the ratio of male to total births in a sample of 10,000. The theory of confidence intervals says that I may assert that the proportion p lies between 0·59 and 0·61 with the probability that, if I make this type of assertion systematically in all similar cases, I shall be about 95 % right in the long run. But I do not then make such an assertion because I know too much about birth-rates to believe any such thing. The theory of confidence intervals gives no place to prior knowledge of the situation. How, then, can that theory provide a guide to conduct in making decisions?

45. This difficulty arises in most of the accepted techniques of modern statistics which, while using judgements in probability, make no allowance for prior probabilities. I shall discuss confidence intervals to fix the ideas; but much the same considerations apply, for example, to the theory of testing hypotheses, to fiducial inference, to the analysis of variance and to sequential analysis. The fundamental problem is to find some method of incorporating prior knowledge or prior probabilities into the final probabilistic judgement of the situation. In order to consider it I need to recall briefly the nature of the theory of confidence intervals.

46. The diagram of Fig. 1 is a familiar presentation of confidence intervals for the binomial distribution.

For any given value ϖ and fixed sample number n we may calculate the binomial distribution $(\chi+\varpi)^n$, and, having decided on a confidence coefficient α, find a range of values from p_0 to p_1, such that the proportion of the total distribution lying inside the range is α. (A minor question as to discontinuities in values of p_0 and p_1 I ignore as not affecting the argument.) Having further decided (as I shall for simplicity, again without affecting the argument) to take central confidence intervals, we may map the confidence lines L_1 and L_2 by plotting, for each value of ϖ, the appropriate values of p_0 and p_1. The confidence lines are, as it were, constructed horizontally by plotting the abscissae for selected ordinates.

To use the diagram we read it vertically. For a given abscissa (observed value of p) we read off on the confidence lines the corresponding ordinates (values of ϖ), say ϖ_1 (on L_2) and ϖ_2 (on L_1). We then assert that

$$\varpi_1 \leqslant \varpi \leqslant \varpi_2, \tag{10}$$

with an assurance of being in the long run right in proportion α of the cases in which we make this type of assertion.

47. From the way in which the diagram is constructed it follows that, whatever the frequency distribution of ϖ may be in the cases to which we apply the method, a proportion α of the happenings will lie in the confidence belt between the confidence lines; for the proportion is α in each horizontal elementary strip, and hence is so for the belt as a whole even if different weights are assigned to different strips. Now let us suppose that we know something about the prior probabilities of ϖ; to take an extreme case I shall suppose that we know that ϖ lies in a certain range, say M_1 to M_2. If we adhere to the confidence rules we shall still assert (10) and shall still be right in proportion α of the cases. But nobody in his senses would assert (10) if the range ϖ_1 to ϖ_2 was outside the range M_1 to M_2 at any point. We shall continue to be right in proportion α of the cases if we assert an inequality

$$\gamma_1 \leqslant \varpi \leqslant \gamma_2, \tag{11}$$

where γ_1 is the greater of ϖ_1 and M_1 and γ_2 is the smaller of ϖ_2 and M_2. In short, we can modify *and shorten* our confidence intervals in the light of prior information without loss of accuracy.

48. To take a slightly more general case let us now suppose that we have a known prior probability distribution of ϖ, $f(\varpi)\,d\varpi$ which varies effectively in the range 0 to 1. We can still improve on the inequality (10) without loss of accuracy. We merely have to determine a domain in the square of Fig. 1 such that the total probability (allowing for variation of ϖ as well as of p) shall be α and obeying certain elementary requirements as to connexity and convexity of the confidence belt. The situation is essentially the same as if, in determining the confidence line, we started with probability $f(\varpi)\,(\chi+\varpi)^n$ instead of the second factor only.

49. Provided, then, that we know the prior distribution $f(\varpi)$ we can incorporate it into the determination of our confidence belt without much difficulty. The essence of the method of confidence intervals is not that it takes $f(\varpi)$ to be unity, to comply with Bayes's postulate (though that is its effect) but that it ignores $f(\varpi)$ and can still obtain accurate statements in probability. It obtains (in this case and all cases where *ad hoc* prior distributions are not introduced for special reasons) the same results as if Bayes's postulate were applied. But to avoid the use of the postulate and at the same time to maintain rigour it has to sacrifice a great deal. It pretends, so to speak, either that there is no prior probability in the frequency sense which is being used to set up the confidence intervals, or that any prior knowledge must

be blended with the results of the theory in some manner unspecified before the final judgement is made. I suspect that the advocates of confidence intervals have often forgotten this fact; I think that they often do not appreciate its importance; I am certain that they fail to bring it out adequately in their expositions. *I therefore assert that the theory of confidence intervals offers only a partial solution to the problem of estimation, and that to provide a basis of rational action it is necessary either to introduce prior probabilities into the theory or to find some new way of linking the theory with prior knowledge or prior expectation.*

50. Two further points require stressing in connexion with confidence intervals. First, it is part of the hypothesis that the observations are being generated by a random process. We believe this, as a rule, on the basis of collateral evidence, i.e. on the basis of prior knowledge. I know of no convincing explanation why we are supposed to use this prior knowledge but to ignore any we may have about the parameter under estimate. Secondly, the method requires some theorem such as Bernoulli's to the effect that if the probability of an event is p it will happen in proportion p of the cases in the long run. There has been a good deal of discussion and some misunderstanding about the role of Bernoulli's theorem in the theory of probability. Essentially it is a proposition in pure mathematics which may be expressed by saying that the proportion of total frequency in the binomial series $(\chi + \varpi)^n$ in the neighbourhood of the mode $n\varpi$ tends to unity as n tends to infinity, 'neighbourhood' meaning within any *fixed* multiple of the dispersion $\sqrt{(n\varpi\chi)}$ however small. It is not a justification of the frequency theory of probability (as Bernoulli himself seems to have thought) in the sense that it asserts anything about the frequency of happenings of equi-probable alternatives. No amount of mathematics, in fact, can assert anything about events which was not latent in the premisses concerning those events. Consequently, in relying on the result of the theorem (or anything equivalent to the effect that events of probability p will happen in proportion p of the cases), the theory of confidence intervals, like the theory of direct probability from the frequentist viewpoint, requires the basic assumption that the processes with which it is concerned do behave in the manner required by the theory; in short, that random processes do exist.

51. It would be tedious to trace the same points through other statistical techniques such as the theory of testing hypotheses. I will merely note that they occur in much the same form and, so far as I know, do not raise any essentially different problems. Not that there are no other parallel problems in statistical theory—conditional inference and order-statistics are two examples of cases where further examination of fundamentals is required—but that the main points have been covered in the foregoing.

52. A friend of mine once remarked to me that if some people asserted that the earth rotated from east to west and others that it rotated from west to east, there would always be a few well-meaning citizens to suggest that perhaps there was something to be said for both sides, and that maybe it did a little of one and a little of the other; or that the truth probably lay between the extremes and perhaps it did not rotate at all. I wish, in conclusion, to emphasize that I am not attempting a compromise of this kind in endeavouring to reconcile the different theories which have been put forward by different authorities. It is not so much a question of choosing between viewpoints as of synthesizing them in order to get a complete picture of the whole.

REFERENCES

An account of the various statistical techniques mentioned in this paper is given in my *Advanced Theory of Statistics* together with extensive references which I need not repeat. The specific works alluded to are:

CRAMÉR, H. (1945). *Mathematical Methods of Statistics.* Uppsala: Almqvist and Wicksell.
JEFFREYS, H. (1939). *Theory of Probability.* Oxford University Press. (2nd ed. 1948.)
JOHNSON, E. W. (1926). *Logic.* Cambridge University Press.
KENDALL, M. G. (1940). On the method of maximum likelihood. *J. R. Statist. Soc.* **103**, 387.
KEYNES, J. M. (1921). *A Treatise on Probability.* London: Macmillan.
KOLMOGOROFF, A. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung.* Berlin: Springer. (Reprint, 1946, by the Chelsea Publishing Company, New York.)
PIETRA, G. (1948). *Studi di Statistica Methodologica.* Milan: Giuffrè.
VON MISES, R. (1928). *Wahrscheinlichkeit, Statistik und Wahrheit,* 3rd ed. revised 1936. Berlin: Springer. (English translation, 1939, W. Hodge, London.)