Inverse probability and the use of Likelihood. By R. A. FISHER, Sc.D., F.R.S., Gonville and Caius College.

[*Received* 8 March, *read* 16 May 1932.]

Logicians have long distinguished two modes of human reasoning, under the respective names of deductive and inductive reasoning. In deductive reasoning we attempt to argue from a hypothesis to its necessary consequences, which may be verifiable by observation; that is, to argue from the general to the particular. In inductive reasoning we attempt to argue from the particular, which is typically a body of observational material, to the general, which is typically a theory applicable to future experience. In statistical language we are attempting to argue from the sample to the population, from which it was drawn. Since recent statistical work has shown that this type of argument can be carried out with exactitude in a usefully large class of cases (2, 3), by means of conceptions somewhat different from those of the classical theory of probability, it may be useful briefly to restate the logical and mathematical distinctions which have to be drawn.

The mathematical work on inverse probability of the eighteenth and nineteenth centuries, beginning with Bayes' *Essay on the doctrine of chances* (4) in 1763, has made it perfectly clear that, if we can assume that our unknown population has been chosen at random from a super-population, or population of populations, the characteristics of which can be completely specified from *a priori* knowledge, then the statement of our inferences from the sample to the population can be put into a purely deductive form, and expressed in terms of mathematical probability. This is hardly surprising, since our data now supply us with precise information as to the generality of the populations of the kind under discussion, and we are thus in a position from the first to approach the problem deductively. Mathematicians have, however, often been tempted to apply the procedure, appropriate to this rather special case, to types of problem in which our *a priori* knowledge is certainly not of the definite kind postulated; partly perhaps because they had been trained in a great tradition of exact deductive inference, but were without example or precedent in the exact use of inductive processes; and partly because of a very remarkable feature of the mathematics, which early attracted attention, namely that as the observational material is made more and more ample, uncertainty, with respect to our *a priori* premises, makes in our result less and less difference.

In the notation used by Mr J. B. S. Haldane(1) in a recent note in these *Proceedings*, if x is an unknown probability, that is the unknown fraction of the population from which our observations are drawn, which is characterised by some observational peculiarity, and if it is known *a priori* that our population has been chosen at random from a super-population in which the frequency with which x lies in the range dx is given by the known frequency element

$$f(x)\,dx \equiv e^{\phi(x)}\,dx,$$

then the joint probability that a population shall have been chosen in the range dx, and that of n observations drawn from that population a shall be of specified kind, will be

$$\frac{n!}{a!\,(n-a)!}\,x^a\,(1-x)^{n-a}\,f(x)\,dx.$$

The numerical coefficient, which is independent of the unknown x, may be ignored in further discussion. The remainder may be interpreted as proportional to the frequency, when the data have in fact given a successes out of n, with which the unknown probability will have fallen in the range dx. Knowing the frequency distribution of x we could, of course, calculate its mean value, its median—that value which would be exceeded in 50 trials out of 100—or any other characteristic that might be required, and the fact with which we are here concerned is that, of the two factors of which our frequency element is composed, that which is contributed by, and may be calculated from, our observations, becomes, as the sample is increased, more and more influential, while the factor $f(x)\,dx$, contributed by our *a priori* knowledge, becomes less and less influential in determining these quantities; so that, subject to the very broad reservation that $f(x)$ shall be non-vanishing and differentiable at that value of x towards which a/n tends, we may say that our conclusions tend to be the same, as the abundance of our data is increased without limit, whatever the particular form of our *a priori* information.

We have of course no such assurance of the harmlessness of erroneous *a priori* assumptions, when our observations are finite in number, as is invariably the case in practice. Nevertheless, the fact under discussion has been used to justify the procedure of assigning an arbitrary function, such as $f(x)=1$, to the *a priori* distribution, in cases where it is, in reality, unknown; on the ground that such errors as we introduce in doing so, since they tend to vanish with increasingly abundant data, will not infect our conclusions with a greater uncertainty than that to which, as based on finite material, they are inevitably prone. The obvious objection to this line of argument is that, if the function $f(x)$ is in reality irrelevant to our conclusions, it should have no place in our reasoning; and that if the form of our reasoning requires its introduction, the fault lies with our adoption of this form of reasoning. The conclusion to be drawn from the decreasing importance of our *a priori* information is not the trivial one that, by introducing false *a priori* data, we may quite possibly not be led far astray, but, rather, that it indicates the fundamentally different position that conclusions can be drawn from the data alone, and that, if the questions we ask seem to require knowledge prior to these, it is because, through thinking only in terms of mathematical probability, and of the deductive processes appropriate to it, we have been asking somewhat the wrong questions. The assumption which has misled us is that, because many statements of uncertain inference can be made with precision in terms of mathematical probability, therefore this same concept is competent for the exact specification of all forms of uncertain inference of which the human mind is capable.

In cases therefore in which we allow our total ignorance of the super-population from which our population might be supposed to have been drawn, or in which, having some vague knowledge, we are unwilling to admit it as the basis of precise mathematical inference, the information supplied by the sample, as the basis for a purely inductive process of reasoning, by which the properties of the population are to be inferred, is summed up in the factor

$$x^{a}(1-x)^{n-a}.$$

This factor is a function of x, and not a differential increment of such a function. It is not a probability and does not obey the laws of probability. It can, however, be shown to provide, not only in the estimation of a probability, but in the whole field of statistical estimation, as satisfactory a measure of "degree of rational belief" as a probability could do. For this reason I have termed it, or some arbitrary multiple of it, the *likelihood*, based on the information supplied by the sample, of any particular value of x. Obviously the claim that the likelihood possesses these properties, and provides a rational basis for exact inference, can only be made in the light of a theory of estimation applicable to finite samples. In (2) I have developed such a theory, and have demonstrated that the most likely value of x, that is the particular estimate found by the method of maximum likelihood, possesses uniquely those sampling properties which are required of a satisfactory estimate.

For the details of this work the reader must be referred to the paper in question. At the present it will suffice to mention that when a *sufficient* statistic exists, that is one which in itself supplies the whole of the information contained in the sample, that statistic is the solution of the equation of maximum likelihood: that when no sufficient statistic exists, the solution of the equation of maximum likelihood is *efficient* in the sense that the quantity of information lost tends in large samples to a zero fraction of the whole, and that this solution contains more information than other efficient statistics. Further, setting aside, for simplicity of statement, cases involving discontinuities, the limiting value of the amount of information lost may be reduced to zero by calculating, again from the likelihood, a second statistic ancillary to the primary estimate, and indeed may be reduced to a zero of any order by a series of such ancillary statistics. These latter properties are of interest in showing that, though the primary problem of estimation is solved by one feature only, namely the *maximum* of the likelihood, yet when we press for the fullest information obtainable from a finite body of data, it is the whole course of the function which has to be used.

The outline above will show sufficiently clearly that a correction is needed on one or two points on which Mr Haldane alludes to this work. On page 60 in reference to the method of maximum likelihood he says:

"In this case Fisher showed (subject to the tacit assumption that all values of x in the neighbourhood of the optimal value are equally probable *a priori*) that the probability density of x is given by

$$dp = \frac{1}{(2\pi)^{\frac{1}{2}}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}} dx,$$

where $\bar{x}$, the optimal value, is the root of

$$L'(x) = 0 \text{ and } \sigma^{-2} = -L''(\bar{x})."$$

I had hoped that it should be clear that my work was based not on the tacit assumption of equal *a priori* probability, but upon the explicit rejection of this assumption. A closer reading of the passage in its context shows, however, that the theorem I am credited with does not belong to me at all; for Haldane's distribution is that of the unknown parameter x, and the theorem to which he evidently alludes deals with the sampling distribution about this true value of the optimal estimates, which we should obtain from different samples.

The text continues as follows:

"Fisher defines the likelihood of x as a quantity proportional to $e^{L(x)}$. This is a convenience of statement, but the introduction of the *a priori* probability density $f(x)$ allows the deduction of Fisher's results without introducing concepts other than those found in the theory of direct probability. The method of maximum likelihood in its complete form is only applicable where σ is somewhat smaller than the difference between $\bar{x}$ and the upper or lower limits which it can possibly attain, and where the graph of

$L(x)$ can be adequately fitted by a parabola in the neighbourhood of $\bar{x}$.”

I suppose I need not protest against the comprehensive phrase “Fisher’s results”, for, as far as I can judge, the rather vague conclusion that the maximum likelihood estimate will be pretty near the middle of any reasonable inverse probability distribution. The phrase “the method of maximum likelihood in its complete form” seems, if I read it aright, to refer to Haldane’s use of the method of inverse probability, and it would be less misleading to students if he had used some such term. I imagine that it is this same method which is ascribed to me lower in the page:

“Hence it is a sufficient condition for the validity of Fisher’s theory that $[\phi'(x)]^2$ should be small compared with $-L''(\bar{x})$ in the neighbourhood of $x=\bar{x}$”, and I must insist that, in so far as I have been guilty of a theory, it is entirely independent of the properties of ϕ, and its derivatives, and that the principal point of it lies in this independence.

REFERENCES.

(1) J. B. S. Haldane, “A note on inverse probability”, *Proc. Cambridge Phil. Soc.*, 28 (1932), 55–61.

(2) R. A. Fisher, “The mathematical foundations of theoretical statistics”, *Phil. Trans.*, A, 222 (1922), 309–368.

(3) R. A. Fisher, “Inverse probability”, *Proc. Cambridge Phil. Soc.*, 26 (1930), 528–535.

(4) T. Bayes, “An essay towards solving a problem in the doctrine of chances”, *Phil. Trans.*, 53 (1763), 370–418.