84

INVERSE PROBABILITY

Author's Note (CMS 22.527a) *

This short paper to the Cambridge Philosophical Society was intended to introduce the notion of "fiducial probability," and the type of inference which may be expressed in this measure. It opens with a discussion of the difficulties which had arisen from attempts to extend Bayes' theorem to problems in which the essential information on which Bayes' theorem is based is in reality absent, and passes on to relate the new measure to the likelihood function, previously introduced by the author, and to distinguish it from the Bayesian probability *a posteriori*.

It is emphasised that statements of equality (exact statements) of fiducial probability can only be derived from statistics having continuous distributions. It should also have been emphasised that the information they supply as to the unknown parameter should be exhaustive. Only the case of a single parameter is discussed. The importance of the paper lies, however, in setting forth a new mode of reasoning from observations to their hypothetical causes.

 Proceedings of the Cambridge Philosophical Society, 26: 528-535, (1930).

Inverse Probability. By R. A. FISHER, Sc.D., F.R.S., Gonville and Caius College; Statistical Dept., Rothamsted Experimental Station.

[*Received* 23 July, *read* 28 July 1930.]

I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation. Yet that is quite exactly the position in respect of inverse probability. Bayes, who seems to have first attempted to apply the notion of probability, not only to effects in relation to their causes but also to causes in relation to their effects, invented a theory, and evidently doubted its soundness, for he did not publish it during his life. It was posthumously published by Price, who seems to have felt no doubt of its soundness. It and its applications must have made great headway during the next 20 years, for Laplace takes for granted in a highly generalised form what Bayes tentatively wished to postulate in a special case.

Before going over the formal mathematical relationships in terms of which any discussion of the subject must take place, there are two preliminary points which require emphasis. First, it is not to be lightly supposed that men of the mental calibre of Laplace and Gauss, not to mention later writers who have accepted their views, could fall into error on a question of prime theoretical importance, without an uncommonly good reason. The underlying mental cause is not to be confused with the various secondary errors into which one is naturally led in deriving a formal justification of a false position, such as for example Laplace's introduction into his definition of probability of the unelucidated phrase "equally *possible* cases" which, since we must be taken to know what cases are equally possible before we know that they are equally probable, can only lead to the doctrine, known as the "doctrine of insufficient reason," that cases are equally probable (to us) unless we have reason to think the contrary, and so reduces all probability to a subjective judgment. The underlying mental cause is, I suggest, not to be found in these philosophical entanglements, but in the fact that we learn by experience, that science has its inductive processes, so that it is naturally thought that such inductions, being uncertain, must be expressible in terms of probability. In fact, the argument runs somewhat as follows: a number of useful but uncertain judgments can be expressed with exactitude in terms of probability; our judgments respecting causes or hypotheses are uncertain, therefore our rational attitude towards them is expressible

in terms of probability. The assumption was almost a necessary one seeing that no other mathematical apparatus existed for dealing with uncertainties.

The second point is that the development of the subject has reduced the original question of the inverse argument in respect of probabilities to the position of one of a series of quite analogous questions; the hypothetical value, or parameter of the population under discussion, may be a probability, but it may equally be a correlation, or a regression, or, in genetical problems, a linkage value, or indeed any physical magnitude about which the observations may be expected to supply information. The introduction of quantitative variates, having continuous variation in place of simple frequencies as the observational basis, makes also a remarkable difference to the kind of inference which can be drawn.

It will be necessary to summarise some quite obvious properties of these continuous frequency distributions. The probability that a variate x should have a value in the range $x \pm \frac{1}{2}dx$ is expressed as a function of x in the form

$$df = \phi(x)\, dx.$$

The function depends of course on the particular population from which the value of x is regarded as a random sample, and specifies the distribution in that population. If in the specification of the population one or more parameters, θ_1, θ_2, θ_3, ... are introduced, we have

$$df = \phi(x, \theta_1, \theta_2, \theta_3, \ldots)\, dx,$$

where ϕ now specifies only the form of the population, the values of its parameters being represented by θ_1, θ_2, θ_3, ...

Knowing the distribution of the variate x, we also know the distribution of any function of x, for if

$$x = \chi(\xi)$$

we may substitute for x and obtain the distribution of ξ in the form

$$df = \phi\{\chi(\xi)\} \frac{d\chi}{d\xi} d\xi.$$

Obviously the form of the distribution has changed; thus, if we know the frequency distribution of the time in which a number of men run 100 yards, we may derive the distribution of their velocities, which will be a different distribution, obtained simply by transforming df as a differential element. In particular we must notice that the mean of the distribution is not invariant for such transformations, thus, if $\bar{x}$ and $\bar{\xi}$ are the means of their respective distributions, we shall not in general find that

$$\bar{x} = \chi(\bar{\xi}).$$

Similarly, the *mode*, that is, the point, if there is one, at which ϕ has a maximum for variation of x, will not be invariant, for the equations

$$\frac{d^2f}{dx^2}=0, \quad \frac{d^2f}{d\xi^2}=0$$

will not normally be satisfied by corresponding values. The central measure which is invariant, at least if $d\chi/d\xi$ is positive for all values, is the *median*, the value which divides the total frequency into two equal halves. For this point $f=\frac{1}{2}$, and the values of x and ξ will be necessarily in agreement. The same will be true of all other points defined by the value of f, so that we may have deciles, centiles, etc., dividing the frequency into 10 or 100 equal parts, and these will be invariant for any transformation for which $d\chi/d\xi$ is always positive.

All the above applies with no essential change to the more general case in which we have several observable variates $x, y, z, \ldots$ in place of one.

The general statement of the inverse type of argument is as follows; we shall first cloak its fallacy under an hypothesis, and then examine it as an undisguised assumption.

Suppose that we know that the population from which our observations were drawn had itself been drawn at random from a super-population of known specification; that is, suppose that we have *a priori* knowledge that the probability that $\theta_1, \theta_2, \theta_3, \ldots$ shall lie in any defined infinitesimal range $d\theta_1\, d\theta_2\, d\theta_3 \ldots$ is given by

$$dF = \Psi(\theta_1, \theta_2, \theta_3, \ldots)\, d\theta_1 d\theta_2 d\theta_3 \ldots,$$

then the probability of the successive events (a) drawing from the super-population a population with parameters having the particular values $\theta_1, \theta_2, \theta_3, \ldots$ and (b) drawing from such a population the sample values $x_1, \ldots, x_n$, will have a joint probability

$$\Psi(\theta_1, \theta_2, \theta_3, \ldots)\, d\theta_1 d\theta_2 d\theta_3 \ldots \times \prod_{p=1}^{n} \{\phi(x_p, \theta_1, \theta_2, \theta_3, \ldots)\, dx_p\}.$$

If we integrate this over all possible values of $\theta_1, \theta_2, \theta_3, \ldots$ and divide the original expression by the integral we shall then have a perfectly definite value for the probability (in view of the observed sample and of our *a priori* knowledge) that $\theta_1, \theta_2, \theta_3, \ldots$ shall lie in any assigned limits.

This is not inverse probability strictly speaking, but a perfectly direct argument, which gives us the frequency distribution of the population parameters θ, from which we may, if we like, calculate their means, modes, medians or whatever else might be of use.

The peculiar feature of the inverse argument proper is to say something equivalent to "We do not know the function Ψ specifying the super-population, but in view of our ignorance of the actual values of θ we may take Ψ to be constant." Perhaps we might add that all values of θ being equally possible their probabilities are by definition equal; but however we might disguise it, the choice of this particular *a priori* distribution for the θ's is just as arbitrary as any other could be. If we were, for example, to replace our θ's by an equal number of functions of them, θ_1', θ_2', θ_3', ... all objective statements could be translated from the one notation to the other, but the simple assumption $\Psi(\theta_1, \theta_2, \theta_3, \ldots) = \text{constant}$ may translate into a most complicated frequency function for

$$\theta_1', \theta_2', \theta_3', \ldots.$$

If, then, we follow writers like Boole, Venn, and Chrystal in rejecting the inverse argument as devoid of foundation and incapable even of consistent application, how are we to avoid the staggering falsity of saying that however extensive our knowledge of the values of x may be, yet we know nothing and can know nothing about the values of θ? Inverse probability has, I believe, survived so long in spite of its unsatisfactory basis, because its critics have until recent times put forward nothing to replace it as a rational theory of learning by experience.

The first point to be made belongs to the theory of statistical estimation; it has nothing to do with inverse probability, save for the historical accident that it was developed by Gauss in terms of that theory.

If we make the assumption that $\Psi(\theta_1, \theta_2, \theta_3, \ldots) = \text{constant}$, and if then we ignore everything about the inverse probability distribution so obtained except its mode or point at which the ordinate is greatest, we have to maximise

$$\prod_{p=1}^{n} \{\phi(x_p, \theta_1, \theta_2, \theta_3, \ldots)\}$$

for variations of $\theta_1, \theta_2, \theta_3, \ldots$; and the result of *this* process will be the same whether we use the parameters $\theta_1, \theta_2, \theta_3, \ldots$ or any functions of them, $\theta_1', \theta_2', \theta_3', \ldots$. Two wholly arbitrary elements in this process have in fact cancelled each other out, the non-invariant process of taking the mode, and the arbitrary assumption that Ψ is constant. The choice of the mode is thinly disguised as that of "the most probable value," whereas had the inverse probability distribution any objective reality at all we should certainly, at least for a single parameter, have preferred to take the mean or the median value. In fact neither of these two processes has a logical justification, but each is necessary to eliminate the errors introduced by the other.

The process of maximising $\Pi(\phi)$ or $S(\log\phi)$ is a method of estimation known as the "method of maximum likelihood"; it has in fact no logical connection with inverse probability at all. The facts that it has been accidentally associated with inverse probability, and that when it is examined objectively in respect of the properties in random sampling of the estimates to which it gives rise, it has shown itself to be of supreme value, are perhaps the sole remaining reasons why that theory is still treated with respect. The function of the θ's maximised is not however a probability and does not obey the laws of probability; it involves no differential element $d\theta_1 d\theta_2 d\theta_3 \ldots$; it does none the less afford a rational basis for preferring some values of θ, or combination of values of the θ's, to others. It is, just as much as a probability, a numerical measure of rational belief, and for that reason is called the *likelihood* of $\theta_1, \theta_2, \theta_3, \ldots$ having given values, to distinguish it from the probability that $\theta_1, \theta_2, \theta_3, \ldots$ lie within assigned limits, since in common speech both terms are loosely used to cover both types of logical situation.

If A and B are mutually exclusive possibilities the probability of "A or B" is the sum of the probabilities of A and of B, but the likelihood of A or B means no more than "the stature of Jackson or Johnson"; you do not know what it is until you know which is meant. I stress this because in spite of the emphasis that I have always laid upon the difference between probability and likelihood there is still a tendency to treat likelihood as though it were a sort of probability.

The first result is thus that there are two different measures of rational belief appropriate to different cases. Knowing the population we can express our incomplete knowledge of, or expectation of, the sample in terms of probability; knowing the sample we can express our incomplete knowledge of the population in terms of likelihood. We can state the relative likelihood that an unknown correlation is $+0{\cdot}6$, but not the probability that it lies in the range $\cdot 595$—$\cdot 605$.

There are, however, certain cases in which statements in terms of probability can be made with respect to the parameters of the population. One illustration may be given before considering in what ways its logical content differs from the corresponding statement of a probability inferred from known *a priori* probabilities. In many cases the random sampling distribution of a statistic, T, calculable directly from the observations, is expressible solely in terms of a single parameter, of which T is the estimate found by the method of maximum likelihood. If T is a statistic of continuous variation, and P the probability that T should be less than any specified value, we have then a relation of the form

$$P = F(T, \theta).$$

If now we give to P any particular value such as ·95, we have a relationship between the statistic T and the parameter θ, such that T is the 95 per cent. value corresponding to a given θ, and this relationship implies the perfectly objective fact that in 5 per cent. of samples T will exceed the 95 per cent. value corresponding to the actual value of θ in the population from which it is drawn. To any value of T there will moreover be usually a particular value of θ to which it bears this relationship; we may call this the "fiducial 5 per cent. value of θ" corresponding to a given T. If, as usually if not always happens, T increases with θ for all possible values, we may express the relationship by saying that the true value of θ will be less than the fiducial 5 per cent. value corresponding to the observed value of T in exactly 5 trials in 100. By constructing a table of corresponding values, we may know as soon as T is calculated what is the fiducial 5 per cent. value of θ, and that the true value of θ will be less than this value in just 5 per cent. of trials. This then is a definite probability statement about the unknown parameter θ, which is true irrespective of any assumption as to its *a priori* distribution.

Fiducial 5 % ρ	95 % r	Fiducial 5 % ρ	95 % r	Fiducial 5 % ρ	95 % r
−·995055	−·968551	−·761594	+·145340	+·761594	+·989816
−·993963	−·961623	−·716298	+·270475	+·800499	+·991770
−·992632	−·953179	−·664037	+·388574	+·833655	+·993335
−·991007	−·942894	−·604368	+·496089	+·861723	+·994593
−·989027	−·930375	−·537050	+·590725	+·885352	+·995608
−·986614	−·915151	−·462117	+·671557	+·905148	+·996427
−·983675	−·896661	−·379949	+·738849	+·921669	+·997091
−·980096	−·874240	−·291313	+·793711	+·935409	+·997628
−·975743	−·847110	−·197375	+·837715	+·946806	+·998066
−·970452	−·814372	−·099668	+·872590	+·956237	+·998421
−·964028	−·775019	0	+·900000	+·964028	+·998711
−·956237	−·727916	+·099668	+·921432	+·970452	+·998646
−·946806	−·671918	+·197375	+·938146	+·975743	+·999139
−·935409	−·605881	+·291313	+·951174	+·980096	+·999296
−·921669	−·528824	+·379949	+·961338	+·983675	+·999424
−·905148	−·440127	+·462117	+·969286	+·986614	+·999529
−·885352	−·339761	+·537050	+·975519	+·989027	+·999615
−·861723	−·228562	+·604368	+·980424	+·991007	+·999685
−·833655	−·108446	+·664037	+·984298	+·992632	+·999742
−·800499	+·017528	+·716298	+·987371	+·993963	+·999789
−·761594	+·145340	+·761594	+·989816	+·995055	+·999827

For example, if r is a correlation derived from only four pairs of observations, and ρ is the correlation in the population from which the sample was drawn, the relation between ρ and the 95 per cent. value of r is given in the following table, which has been calculated, from the distribution formula I gave in 1915, by Miss F. E. Allan. From the table we can read off the 95 per cent. r for any given ρ, or equally the fiducial 5 per cent. ρ for any given r. Thus if a value $r = \cdot 99$ were obtained from the sample, we should have a fiducial 5 per cent. ρ equal to about $\cdot 765$. The value of ρ can then only be less than $\cdot 765$ in the event that r has exceeded its 95 per cent. point, an event which is known to occur just once in 20 trials. In this sense ρ has a probability of just 1 in 20 of being less than $\cdot 765$. In the same way, of course, any other percentile in the fiducial distribution of ρ could be found or, generally, the fiducial distribution of a parameter θ for a given statistic T may be expressed as

$$df = -\frac{\partial}{\partial \theta} F(T, \theta)\, d\theta,$$

while the distribution of the statistic for a given value of the parameter is

$$df = \frac{\partial}{\partial T} F(T, \theta)\, dT.$$

I imagine that this type of argument, which supplies definite information as to the probability of causes, has been overlooked by the earlier writers on probability, because it is only applicable to statistics of continuous distribution, and not to the cases in regard to which the abstract arguments of probability theory were generally developed, in which the objects of observation were classified and counted rather than measured, and in which therefore all statistics have discontinuous distributions. Now that a number of problems of distribution have been solved, for statistics having continuous distribution, arguments of this type force themselves on our attention; and I have recently received from the American statistician, Dr M. Ezekiel, graphs giving to a good approximation the fiducial 5 per cent. points of simple and multiple correlations for a wide range of cases. It is therefore important to realise exactly what such a probability statement, bearing a strong superficial resemblance to an inverse probability statement, really means. The fiducial frequency distribution will in general be different numerically from the inverse probability distribution obtained from any particular hypothesis as to *a priori* probability. Since such an * hypothesis may be true, it is obvious that the two distributions must differ not only numerically, but in their logical meaning. It would be perfectly possible, for example, to find an *a priori*

* It must not be known to be true - R.A.F.

frequency distribution for ρ such that the inverse probability that ρ is less than ·765 when $r = ·99$ is not 5 but 10 in 100. In concrete terms of frequency this would mean that if we repeatedly selected a population at random, and from each population selected a sample of four pairs of observations, and rejected all cases in which the correlation as estimated from the sample (r) was not exactly ·99, then of the remaining cases 10 per cent. would have values of ρ less than ·765. Whereas apart from any sampling for ρ, we know that if we take a number of samples of 4, from the same or from different populations, and for each calculate the fiducial 5 per cent. value for ρ, then in 5 per cent. of cases the true value of ρ will be less than the value we have found. There is thus no contradiction between the two statements. The fiducial probability is more general and, I think, more useful in practice, for in practice our samples will all give different values, and therefore both different fiducial distributions and different inverse probability distributions. Whereas, however, the fiducial values are expected to be different in every case, and our probability statements are relative to such variability, the inverse probability statement is absolute in form and really means something different for each different sample, unless the observed statistic actually happens to be exactly the same.