124

THE LOGIC OF INDUCTIVE INFERENCE

Author's Note (CMS 26.38a)

In giving this general explanatory account of advances in statistical methods at that time comparatively recent, the opportunity was taken to add a few novelties which might make the evening more interesting to those few among the audience who were already familiar with the general ideas. Of these the modern reader may still find interest in the second half of Example 1. The discussion here may serve to distinguish statements of fiducial probability proper from the "confidence belts" based on tests of significance applied to discontinuous data, which in reality represent inequality statements as to fiducial probability.

Journal of the Royal Statistical Society, 98: 39-54, (1935).

The Logic of Inductive Inference.

By Professor R. A. Fisher, Sc.D., F.R.S.

[Read before the Royal Statistical Society on Tuesday, December 18th, 1934, the President, Professor M. Greenwood, F.R.S., in the Chair.]

When the invitation of your Council was extended to me to address this Society on some of the theoretical researches with which I have been associated, I took it as an indication that the time was now thought ripe for a discussion, in summary, of the net effect of these researches upon our conception of what statistical methods are capable of doing, and upon the outlook and ideas which may usefully be acquired in the course of mathematical training for a statistical career. I welcomed also the invitation, personally, as affording an opportunity of putting forward the opinion to which I find myself more and more strongly drawn, that the essential effect of the general body of researches in mathematical statistics during the last fifteen years is fundamentally a reconstruction of logical rather than mathematical ideas, although the solution of mathematical problems has contributed essentially to this reconstruction.

I have called my paper "The Logic of Inductive Inference." It might just as well have been called "On making sense of figures." For everyone who does habitually attempt the difficult task of making sense of figures is, in fact, essaying a logical process of the kind we call inductive, in that he is attempting to draw inferences from the particular to the general; or, as we more usually say in statistics, from the sample to the population. Such inferences we recognize to be *uncertain* inferences, but it does not follow from this that they are not mathematically rigorous inferences. In the theory of probability we are habituated to statements which may be entirely rigorous, involving the concept of probability, which, if translated into verifiable observations, have the character of uncertain statements. They are rigorous because they contain within themselves an adequate specification of the nature and extent of the uncertainty involved. This distinction between uncertainty and lack of rigour, which should be familiar to all students of the theory of probability, seems not to be widely understood by those mathematicians who have been trained, as most mathematicians are, almost exclusively in the technique of deductive reasoning; indeed, it would not be surprising or exceptional to find mathematicians of this class ready to deny at first sight that rigorous inferences from the particular to the general were even possible. That they are, in fact, possible is, I suppose, recognized by all who are familiar with the modern work. It will be sufficient here to note that the denial implies, qualitatively, that the process of learning by observation, or experiment, must always lack real cogency.

My second preliminary point is this. Although some uncertain inferences can be rigorously expressed in terms of mathematical probability, it does not follow that mathematical probability is an adequate concept for the rigorous expression of uncertain inferences of every kind. This was at first assumed; but once the distinction between the proposition and its converse is clearly stated, it is seen to be an assumption, and a hazardous one. The inferences of the classical theory of probability are all deductive in character. They are statements about the behaviour of individuals, or samples, or sequences of samples, drawn from populations which are fully known. Even when the theory attempted inferences respecting populations, as in the theory of inverse probability, its method of doing so was to introduce an assumption, or postulate, concerning the population of populations from which the unknown population was supposed to have been drawn at random; and so to bring the problem within the domain of the theory of probability, by *making* it a deduction from the general to the particular. The fact that the concept of probability is adequate for the specification of the nature and extent of uncertainty in these deductive arguments is no guarantee of its adequacy for reasoning of a genuinely inductive kind. If it appears in inductive reasoning, as it has appeared in some cases, we shall welcome it as a familiar friend. More generally, however, a mathematical quantity of a different kind, which I have termed *mathematical likelihood*, appears to take its place as a measure of rational belief when we are reasoning from the sample to the population.

Mathematical likelihood makes its appearance in the particular kind of logical situation which I have termed a *problem of estimation.* In logical situations of other kinds, which have not yet been explored, possibly yet other means of making rigorous our uncertain inferences may be required. In a problem of estimation we start with a knowledge of the mathematical form of the population sampled, but without knowledge of the values of one or more parameters which enter into this form, which values would be required for the complete specification of the population; or, in other words, for the complete specification of the probabilities of the observable occurrences which constitute our data. The probability of occurrence of our entire sample is therefore expressible as a function of these unknown parameters, and the likelihood is defined merely as a function of these parameters proportional to this probability. The likelihood is thus an observable property of any hypothesis which specifies the values of the parameters of the population sampled. Neyman and Pearson have attempted to extend the definition of likelihood to apply, not to particular hypotheses only, but to classes of such hypotheses. With this extension we are not here concerned. The best use I can make of the short time at my disposal is to show how it is that a consideration of the problem of estimation, without postulating any special significance for the likelihood function, and of course without introducing any such postulate as that needed for inverse probability, does really demonstrate the adequacy of the concept of likelihood for inductive reasoning, in the particular logical situation for which it has been introduced.

In the theory of estimation we proceed by building up a series of criteria for judging the merits of the estimates arrived at by different methods. Each criterion is thus a method of forming a judgment that some one estimate or group of estimates is better than others. An initial difficulty here arises, best expressed in the question, "Better for what?" and it is remarkable that this preliminary difficulty does not frustrate our enquiry. Whatever other purpose our estimate may be wanted for, we may require at least that it shall be fit to use, in conjunction with the results drawn from other samples of a like kind, as a basis for making an improved estimate. On this basis, in fact, our enquiry becomes self-contained, and capable of developing its own appropriate criteria, without reference to extraneous or ulterior considerations.

This logical characteristic of our approach naturally requires that our edifice shall be built in two stories. In the first we are concerned with the theory of *large samples,* using this term, as is usual, to mean that nothing that we say shall be true, except in the limit when the size of the sample is indefinitely increased; a limit, obviously, never attained in practice. This part of the theory, to set off against the complete unreality of its subject-matter, exploits the advantage that in this unreal world all the possible merits of an estimate may be judged exclusively from its variability, or sampling variance. In the second story, where the *real* problem of finite samples is considered, the requirement that our estimates from these samples may be wanted as materials for a subsequent process of estimation is found to supply the unequivocal criteria required. Let me sketch the two stages, with special emphasis on the staircase, relegating all mathematical demonstrations to the written word.

First, we may distinguish consistent from inconsistent estimates. An inconsistent estimate is an estimate of something other than that which we want an estimate of. If we choose any process of estimation, and imagine the sample from which we make our calculations to increase without limit, our estimate will usually *tend,* in the special sense in which that word is used in statistics, to a limiting value, which is some function of the unknown parameters. Our method is then a consistent one for estimating this particular parametric function, but would be inconsistent for estimating any different function. The limiting value is easily recognized by inserting for the frequencies in our sample their mathematical expectations.

Having satisfied ourselves that our method is consistent, we may now confine our attention to the class of estimates which, as the sample is increased without limit, tend to be distributed about their limiting value in the normal distribution; that is, to the class to which the theory of large samples is applicable. The normal distribution has only two characteristics, its mean and its variance. The mean determines the bias of our estimate, and the variance determines its precision.

The consideration of bias need not detain us. With consistent estimates it must tend to zero; if we wish to use our estimates for tests of significance it is as well that it should tend to zero more rapidly than $n^{-\frac{1}{2}}$. We can always adjust our estimate to make the bias absolutely zero, but this is not usually necessary, for in estimating any parameter we must remember that we are at the same time estimating its reciprocal, or its square, or any other such function, and zero bias in one of these usually implies bias of the order of n^{-1} in the others. This is therefore the normal rate for the bias to approach zero.

Variance is a more serious affair; for a knowledge of the variance of our estimate does not provide us with any means for producing one which shall be less variable. In the cases which we are considering the variance falls off with increasing size of sample always ultimately in inverse proportion to n. The criterion of efficiency is that the limiting value of nV, where V stands for the variance of our estimate, shall be as small as possible. The first point which needs mathematical proof is that the limiting value of $\frac{1}{nV}$ is necessarily less than or equal to a certain quantity, i, which is independent of the method of estimation used.

To show that if T be an estimate of an unknown parameter θ, normally distributed with variance V, then the limit as $n \longrightarrow \infty$, of $\frac{1}{nV}$ cannot exceed a value, i, defined independently of methods of estimation.

Let f stand for the frequency of a particular kind of observation, ϕ for that of a particular kind of sample, and Φ for that of all the kinds of sample which yield a particular value T of the statistic chosen as an estimate. Then in general

$$\log \phi = S(\log f),$$

where S stands for summation over the sample; next

$$\Phi = \Sigma(\phi),$$

where Σ stands for summation over the possible samples which yield the same estimate; and finally

$$1 = \Sigma'(\Phi),$$

where Σ' stands for summation over all possible values of the statistic. When continuous variation is in question, symbols of integration will replace the symbols of summation Σ and Σ'.

If T is distributed normally about θ with variance V,

$$\Phi = \frac{1}{\sqrt{2\pi V}} e^{\frac{-(T-\theta)^2}{2V}} dT.$$

Hence

$$-\frac{\partial^2}{\partial\theta^2}\log\Phi = \frac{1}{V}.$$

Since this is independent of T, we may take the average for all values of T, and obtain

$$\begin{aligned}\frac{1}{V} &= -\Sigma'\Phi\frac{\partial^2}{\partial\theta^2}\log\Phi \\ &= -\Sigma'\frac{\partial^2}{\partial\theta^2}\Phi + \Sigma'\frac{1}{\Phi}\left(\frac{\partial\Phi}{\partial\theta}\right)^2.\end{aligned}$$

Hence

$$\frac{1}{V} = \Sigma'\frac{1}{\Phi}\left(\frac{\partial\Phi}{\partial\theta}\right)^2,$$

since $\Sigma'(\Phi)$ is independent of θ.

Now consider

$$x = \frac{1}{\phi}\frac{\partial\phi}{\partial\theta}$$

as a variate, among the samples which lead to the estimate T. Each value of x occurs with frequency ϕ, so the variance of x is

$$\begin{aligned}&\frac{1}{\Phi}\Sigma(\phi x^2) - \frac{1}{\Phi^2}\Sigma^2(\phi x) \\ &= \frac{1}{\Phi}\left\{\Sigma\frac{1}{\phi}\left(\frac{\partial\phi}{\partial\theta}\right)^2 - \frac{1}{\Phi}\left(\frac{\partial\Phi}{\partial\theta}\right)^2\right\};\end{aligned}$$

but the variance of x is positive, or, the limiting case zero; in taking the mean for all values of T it follows that

$$\Sigma'\Sigma\frac{1}{\phi}\left(\frac{\partial\phi}{\partial\theta}\right)^2 - \Sigma'\frac{1}{\Phi}\left(\frac{\partial\Phi}{\partial\theta}\right)^2$$

is positive or zero. In other words,

$$\frac{1}{V} \leq \Sigma'\Sigma\frac{1}{\phi}\left(\frac{\partial\phi}{\partial\theta}\right)^2,$$

where it is to be noted that the quantity on the right is the average value for all possible samples of

$$\left(\frac{1}{\phi}\frac{\partial\phi}{\partial\theta}\right)^2,$$

and is therefore independent of the method of estimation. To evaluate it we may note that

$$\Sigma'\Sigma\frac{1}{\phi}\left(\frac{\partial\phi}{\partial\theta}\right)^2 = -\Sigma'\Sigma\phi\frac{\partial^2}{\partial\theta^2}\log\phi,$$

which is the average value in all possible samples of

$$-\frac{\partial^2}{\partial\theta^2}\log\phi,$$

or the average value for all possible individual observations of

$$-n\frac{\partial^2}{\partial\theta^2}\log f,$$

or of

$$n\left(\frac{1}{f}\frac{\partial f}{\partial\theta}\right)^2.$$

It appears then that, in large samples in which the statistic is normally distributed,

$$\frac{1}{nV}\leq i,$$

where i is the average value of

$$\left(\frac{1}{f}\frac{\partial f}{\partial\theta}\right)^2;$$

or, if Σ'' stand for summation over all possible observations,

$$i = \Sigma''\left\{\frac{1}{f}\left(\frac{\partial f}{\partial\theta}\right)^2\right\}.$$

We shall come later to regard i as the amount of information supplied by each of our observations, and the inequality

$$\frac{1}{V}\leq ni = I,$$

as a statement that the reciprocal of the variance, or the *invariance,* of the estimate, cannot exceed the amount of information in the *sample.* To reach this conclusion, however, it is necessary to prove a second mathematical point, namely, that for certain estimates, notably that arrived at by choosing those values of the parameters which maximize the likelihood function, the limiting value of

$$\frac{1}{nV}=i.$$

Of the methods of estimation based on linear functions of the frequencies, that with smallest limiting variance is the method of maximum likelihood, and for this the limit in large samples of $\frac{1}{nV}$ is equal to i.

Let x stand for the frequency observed of observations having probability of occurrence f and let $m = nf$, the expected frequency in a sample of n. Consider any linear function of the frequencies,

$$X \equiv S(kx),$$

the summation being for all possible classes of observations, occupied or unoccupied.

If the coefficients k are functions of θ, the equation,

$$X = 0,$$

may be used as an equation of estimation. This equation will be consistent if

$$S(kf) = 0$$

for all values of θ. Differentiating with respect to θ it appears that

$$S\left(f\frac{\partial k}{\partial\theta}\right) + S\left(k\frac{\partial f}{\partial\theta}\right) = 0.$$

Since the mean value of X is zero, the sampling variance of X is

$$S(k^2m) = nS(k^2f),$$

but as the sample is increased indefinitely, the error of estimation bears to the sampling error of X the ratio

$$\frac{-1}{\frac{\partial X}{\partial\theta}} = \frac{-1}{S\left(x\frac{\partial k}{\partial\theta}\right)}.$$

If, therefore,

$$\frac{-n}{S\left(x\frac{\partial k}{\partial\theta}\right)}$$

tends to a finite limit,

$$\frac{-1}{S\left(f\frac{\partial k}{\partial\theta}\right)}.$$

the sampling variance of our estimate is

$$\frac{S(k^2f)}{nS^2\left(f\frac{\partial k}{\partial\theta}\right)},$$

or, using the condition for consistency,

$$\frac{S(k^2f)}{nS^2\left(k\frac{\partial f}{\partial\theta}\right)}.$$

We may now apply the calculus of variations or simple differentiation to find the functions of k, which will minimize the sampling variance. Since the variance must be stationary for variations of each several value of k, the differential coefficients of the numerator and the denominator, with respect to k, must be proportional for all classes. Hence,

$$kf \propto \frac{\partial f}{\partial \theta},$$

which is satisfied by putting

$$k = \frac{1}{f}\frac{\partial f}{\partial \theta}.$$

This also satisfies the requirement that

$$S(kf) = 0$$

for all values of θ. The equation of estimation

$$S\left(\frac{x}{f}\frac{\partial f}{\partial \theta}\right) = 0$$

is the equation of maximum likelihood. The limiting value of the sampling variance given by the analysis above is

$$nV = \frac{1}{S\left\{\frac{1}{f}\left(\frac{\partial f}{\partial \theta}\right)^2\right\}}$$

or

$$\frac{1}{nV} = S\left\{\frac{1}{f}\left(\frac{\partial f}{\partial \theta}\right)^2\right\} = i.$$

The condition for the validity of the approach to the limit is seen to be merely that i shall be finite. Cases where i is zero or infinite can sometimes be treated by a functional transformation of the parameter; other cases deserve investigation. The proof shows, in fact that where i is finite there really are I and no less units of information to be extracted from the data, if we equate the information extracted to the invariance of our estimate.

This quantity i, which is independent of our methods of estimation, evidently deserves careful consideration as an intrinsic property of the population sampled. In the particular case of error curves, or distributions of estimates of the same parameter, the amount of information of a single observation evidently provides a measure of the intrinsic accuracy with which it is possible to evaluate that parameter, and so provides a basis for comparing the accuracy of error curves which are not normal, but may be of quite different forms.

We are now in a position to consider the real problem of finite samples. For any method of estimation has its own characteristic distribution of errors, not now necessarily normal, and therefore its own intrinsic accuracy. Consequently, the amount of information which it extracts from the data is calculable, and it is possible to compare the merits of different estimates, even though they all satisfy the criterion of efficiency in the limit for large samples. It is obvious, too, that in introducing the concept of quantity of information we do not want merely to be giving an arbitrary name to a calculable quantity, but must be prepared to justify the term employed, in relation to what common sense requires, if the term is to be appropriate, and serviceable as a tool for thinking. The mathematical consequences of identifying, as I propose, the intrinsic accuracy of the error curve, with the amount of information extracted, may therefore be summarized specifically in order that we may judge by our pre-mathematical common sense whether they are the properties it ought to have.

First, then, when the probabilities of the different kinds of observation which can be made are all independent of a particular parameter, the observations will supply no information about the parameter. Once we have fixed zero we can in the second place fix totality. In certain cases estimates are shown to exist such that, when they are given, the distributions of all other estimates are independent of the parameter required. Such estimates, which are called *sufficient,* contain, even from finite samples, the whole of the information supplied by the data. Thirdly, the information extracted by an estimate can never exceed the total quantity present in the data. And, fourthly, statistically independent observations supply amounts of information which are additive. One could, therefore, develop a mathematical theory of quantity of information from these properties as postulates, and this would be the normal mathematical procedure. It is, perhaps, only a personal preference that I am more inclined to examine the quantity as it emerges from mathematical investigations, and to judge of its utility by the free use of common sense, rather than to impose it by a formal definition. As a mathematical quantity information is strikingly similar to *entropy* in the mathematical theory of thermo-dynamics. You will notice especially that reversible processes, changes of notation, mathematical transformations if single-valued, translation of the data into foreign languages, or rewriting them in code, cannot be accompanied by loss of information; but that the irreversible processes involved in statistical estimation, where we cannot reconstruct the original data from the estimate we calculate from it, may be accompanied by a loss, but never by a gain.

Having obtained a criterion for judging the merits of an estimate in the real case of finite samples, the important fact emerges that, though sometimes the best estimate we can make exhausts the information in the sample, and is equivalent for all future purposes to the original data, yet sometimes it fails to do so, but leaves a measurable amount of the information unutilized. How can we supplement our estimate so as to utilize this too? It is shown that some, or sometimes all of the lost information may be recovered by calculating what I call ancillary statistics, which themselves tell us nothing about the value of the parameter, but, instead, tell us how good an estimate we have made of it. Their function is, in fact, analogous to the part which the *size* of our sample is always expected to play, in telling us *what reliance* to place on the result. Ancillary statistics are only useful when different samples of the same size can supply different amounts of information, and serve to distinguish those which supply more from those which supply less.

Example 1.

The use of ancillary statistics may be illustrated in the well-worn topic of the 2 × 2 table. Let us consider such a classification as Lange supplies in his study on criminal twins. Out of 13 cases judged to be monozygotic, the twin brother of a known criminal is in 10 cases also a criminal; and in the remaining 3 cases he has not been convicted. Among the dizygotic twins there are only 2 convicts out of 17. Supposing the data to be accurate, homogeneous, and unselected, we need to know with what frequency so large a disproportion would have arisen if the causes leading to conviction had been the same in the two classes of twins. We have to judge this from the 2 × 2 table of frequencies.

Convictions of Like-sex Twins of Criminals.

	Convicted.	Not Convicted.	Total.
Monozygotic	10	3	13
Dizygotic	2	15	17
Total	12	18	30

To the many methods of treatment hitherto suggested for the 2 × 2 table the concept of ancillary information suggests this new one. Let us blot out the contents of the table, leaving only the marginal frequencies. If it be admitted that these marginal frequencies by themselves supply no information on the point at issue, namely, as to the proportionality of the frequencies in the body of the table, we may recognize the information they supply as wholly ancillary; and therefore recognize that we are concerned only with the relative probabilities of occurrence of the different ways in which the table can be filled in, subject to these marginal frequencies. These ways form a linear sequence completely specified by giving to the number of dizygotic convicts the 13 possible values from 0 to 12. The important point about this approach is that the relative frequencies of these 13 possibilities are the same whatever may be the probabilities of the twin brother of a convict falling into the four compartments prepared for him, provided that these probabilities are *in proportion*.

For, suppose that, knowing him to be of monozygotic origin, the probability that he shall have been convicted is p, it follows that the probability that of 13 monozygotic $(12 - x)$ shall have been convicted, while $(1 + x)$ have escaped conviction, is

$$\frac{13\,!}{(12 - x)\,!\,(1 + x)\,!}\,p^{12-x}\,(1 - p)^{1+x}.$$

But, if we know that the probabilities are in proportion, the probability of a criminal's brother known to be dizygotic being convicted will also be p, and the probability that of 17 of these x shall have been convicted and $(17 - x)$ shall have escaped conviction will be

$$\frac{17\,!}{x\,!\,(17 - x)\,!}\,p^{x}(1 - p)^{17-x}.$$

The probability of the simultaneous occurrence of these two events, being the product of their respective probabilities, will therefore be

$$\frac{13\,!\;17\,!}{(12 - x)\,!\,(1 + x)\,!\,x\,!\,(17 - x)\,!}\,p^{12}(1 - p)^{18},$$

in which it will be noticed that the powers of p and $1 - p$ are independent of x, and therefore represent a factor which is the same for all 13 of the possibilities considered. In fact the probability of any value of x occurring is proportional to

$$\frac{1}{(12 - x)\,!\,(1 + x)\,!\,x\,!\,(17 - x)\,!},$$

and on summing the series obtained by varying x, the absolute probabilities are found to be

$$\frac{13\,!\,17\,!\,12\,!\,18\,!}{30\,!}\cdot\frac{1}{(12 - x)\,!\,(1 + x)\,!\,x\,!\,(17 - x)\,!}.$$

Putting $x = 0, 1, 2, \ldots$ the probabilities are therefore

$$\frac{13\,!\,18\,!}{30\,!}\left\{1, \frac{12\,.\,17}{2}, \frac{12\,.\,11\,.\,17\,.\,16}{2\,!\,3\,!}, \ldots\right\}$$

$$= \frac{1}{6{,}653{,}325}\,\{1, 102, 2992, \ldots\}$$

The significance of the observed departure from proportionality is therefore exactly tested by observing that a discrepancy from proportionality as great or greater than that observed, will arise, subject to the conditions specified by the ancillary information, in exactly 3,095 trials out of 6,653,325, or approximately once in 2,150 trials. The test of significance is therefore direct, and exact for small samples. No process of estimation is involved

The use of the margins as ancillary information suggests a more general treatment. Had the hypothesis we wish to examine made the chances of criminality different for monozygotic and dizygotic twins, *e.g.* p in one case and p' in the other, the probability of observing any particular value of x would have included an additional factor

$$\left(\frac{p'q}{pq'}\right)^x.$$

If

$$\frac{p'q}{pq'} = \psi,$$

the frequency distribution is determined by the parameter ψ, and for each value of ψ we can make a test of significance by calculating the probability,

$$(1 + 102\psi + 2992\psi^2)/(1 + 102\psi + \ . \ . \ . + 476\psi^{12}),$$

the ratio of the partial sum of the hypergeometric series to the hypergeometric function formed by the entire series. This probability rises uniformly as ψ is diminished, and reaches 1 per cent. when ψ is just less than 0·48. We may thus infer that the observations differ significantly, at the 1 per cent. level of significance, from any hypothesis which makes ψ greater than 0·4798. That is to say, that any hypothesis, which is not contradicted by the data at this level of significance, must make the ratio of criminals to non-criminals at least 2·084 times as high among the monozygotic as among the dizygotic cases.

Similarly, the probability rises to 5 per cent. when $\psi = \cdot 28496$, so that any hypothesis which is not contradicted by the data at the 5 per cent. level of significance must make the ratio of criminals to non-criminals at least three and a half times as high among the monozygotic as among the dizygotic.

This is not a probability statement about ψ. It is a formally precise statement of the results of applying tests of significance. If, however, the data had been continuous in distribution, on the hypothesis considered, it would have been equivalent to the statement that the fiducial probability that ψ exceeds 0·4798 is just one chance in a hundred. With discontinuous data, however, the fiducial argument only leads to the result that this probability does not exceed 0·01. We have a statement of inequality, and not one of equality. It is not obvious, in such cases, that, of the two forms of statement possible, the one explicitly framed in terms of probability has any practical advantage. The reason why the fiducial statement loses its precision with discontinuous data is that the frequencies in our table make no distinction between a case in which the 2 dizygotic convicts were only just convicted, perhaps on venial charges, or as first offenders, while the remaining 15 had characters above suspicion, and an equally possible case in which the 2 convicts were hardened offenders, and some at least of the remaining 15 had barely escaped conviction. If we knew where we stood in the range of possibilities represented by these two examples, and had similar information with respect to the monozygotic twins, the fiducial statements derivable from the data would regain their exactitude. One possible device for circumventing this difficulty is set out in Example 2. It is to be noticed that in this example of the fourfold table the notion of ancillary information has been illustrated solely in relation to tests of significance and fiducial probability. No problem of estimation arises. If we want an estimate of ψ we have no choice but to take the actual ratio of the products of the frequencies observed in opposite corners of the table.

Example 2.

On turning a discontinuous distribution, leading to statements of fiducial inequality, into a continuous distribution, capable of yielding exact fiducial statements, by means of a modification of experimental procedure.

Consider the process of estimating the density of micro-organisms in a fluid, by detecting their presence or absence in samples taken at different dilutions. A series of dilutions is made up containing densities of organisms decreasing in geometric progression, the ratios most commonly used being tenfold and twofold. We will suppose, to simplify the reasoning, that the series is effectively infinite, in the sense that it shall be scarcely possible for the organism to fail to appear in the highest concentration examined, or for it to appear in the highest dilution. A number, s, of independent samples are examined at each dilution. The dilution ratio we shall call a, and we shall suppose the dilutions to be numbered consecutively, with the number n increasing as dilution is increased.

If ρ is the density of the organisms to be estimated, then the density in the nth dilution, reckoned on the size of the sample taken, is

$$m = \rho a^{-n}.$$

The chance of a sterile sample is, therefore,

$$p = e^{-m}.$$

The probability of securing t sterile and u fertile cultures at this dilution will therefore be

$$\frac{s\,!}{t\,!\,u\,!}\,p^t(1-p)^u;$$

and the probability of a complete series of observations specified by t_n and u_n at each dilution will be

$$\prod_{n=-\infty}^{n=\infty} \frac{s\,!}{t_n\,!\,u_n\,!}\,p_n^{t_n}(1-p_n)^{u_n},$$

which, regarded as a function of ρ, gives the likelihood of any particular value of the unknown density.

The form of the likelihood function, and therefore the amount of information supplied by a series of observations, depends very greatly on the distribution of the numbers of sterile and fertile samples in that part of the range of dilutions in which both occur. Thus, if there were three samples at each dilution, an experiment in which all were fertile before the nth dilution, and all of the nth and higher dilutions were sterile, would give a higher precision to the estimate than if there were one sterile at the $(n-1)$th dilution, and one fertile at the nth. Consequently, it would be advantageous, if possible, to take account of the configuration of the observed series, that is, of the succession of numbers of sterile samples from the first observed, irrespective of the particular dilution in which this appears, as information ancillary to the interpretation of our estimate, which itself must depend greatly on where the series starts.

The objection to doing this is that, for a given series of dilutions, the frequency with which any particular configuration appears will not be entirely independent of ρ, but will be a periodic function of $\log \rho$, since it evidently does not change when $\log \rho$ is increased or diminished by a multiple of $\log a$. In order to make these frequencies entirely independent of ρ it is, however, sufficient that the particular series of dilutions used should themselves be chosen at random by a process equivalent to the following :—A number, θ, is chosen at random between 0 and 1. In the first dilution, instead of the dilution ratio a we use the dilution ratio a^θ, using the dilution ratio a for all subsequent dilutions. The probability of any particular configuration occurring is now wholly independent of ρ, and, for any configuration the probability of the first sterile sample being drawn from the dilution :—

$$n + \theta = x$$

will be a continuous function of the variate

$$\log \rho - x \log a,$$

which can be completely calculated from the configuration. Consequently, fiducial limits of any chosen probability could be calculated for ρ, merely by observing at what dilution the first sterile sample occurs. For any chosen values of a and s to be used in such tests, the fiducial limits of the commoner configurations could be listed in advance, so reducing the calculation to little more than looking up an anti-logarithm. The artifice of varying the initial dilution in accordance with a number chosen at random for each series thus obviates the need for expressing our conclusions as to the fiducial probability of any proposed density in the form of an inequality.

If we are satisfied of the logical soundness of the criteria developed, we are in a position to apply them to test the claim that mathematical likelihood supplies, in the logical situation prevailing in problems of estimation, a measure of rational belief analogous to, though mathematically different from, that supplied by mathematical probability in those problems of uncertain deductive inference for which the theory of probability was developed This claim may be substantiated by two facts. First, that the particular method of estimation, arrived at by choosing those values of the parameters the likelihood of which is greatest, is found to elicit not less information than any other method which can be adopted. Secondly, the residual information supplied by the sample, which is not included in a mere statement of the parametric values which maximize the likelihood, can be obtained from other characteristics of the likelihood function; such as, if it is differentiable, its second and higher derivatives at the maximum. Thus, basing our theory entirely on considerations independent of the possible relevance of mathematical likelihood to inductive inferences in problems of estimation, we seem inevitably led to recognize in this quantity the medium by which all such information as we possess may be appropriately conveyed.

To those who wish to explore for themselves how far the ideas so far developed on this subject will carry us, two types of problem may be suggested. First, how to utilize the whole of the information available in the likelihood function. Only two classes of cases have yet been solved. (*a*) Sufficient statistics, where the whole course of the function is determined by the value which maximizes it, and where consequently all the available information is contained in the maximum likelihood estimate, without the need of ancillary statistics. (*b*) In a second case, also of common occurrence, where there is no sufficient estimate, the whole of the ancillary information may be recognized in a set of simple relationships among the sample values, which I have called the configuration of the sample. With these two special cases as guides the treatment of the general problem might be judged, as far as one can judge of these things, to be ripe for solution.

Problems of the second class concern simultaneous estimation, and seem to me to turn on how we should classify and recognize the various special relationships which may exist among parameters estimated simultaneously. For example, it is easy to show that two parameters may be capable of sufficient estimation jointly, but not severally, because each estimate contributes the ancillary information necessary to complete the other.

In considering the future progress of the subject it may be necessary to underline certain distinctions between inductive and deductive reasoning which, if unrecognized, might prove serious obstacles to pure mathematicians trained only in deductive methods, who may be attracted by the novelty and diversity of our subject.

In deductive reasoning all knowledge obtainable is already latent in the postulates. Rigour is needed to prevent the successive inferences growing less and less accurate as we proceed. The conclusions are never more accurate than the data. In inductive reasoning we are performing part of the process by which new knowledge is created. The conclusions normally grow more and more accurate as more data are included. It should never be true, though it is still often *said*, that the conclusions are no more accurate than the data on which they are based. Statistical data are always erroneous, in greater or less degree. The study of inductive reasoning is the study of the embryology of knowledge, of the processes by means of which truth is extracted from its native ore in which it is fused with much error.

Secondly, rigour, as understood in deductive mathematics, is not enough. In deductive reasoning, conclusions based on any chosen few of the postulates accepted need only mathematical rigour to guarantee their truth. All statisticians know that data are falsified if only a selected part is used. Inductive reasoning cannot aim at a truth that is less than the whole truth. Our conclusions must be warranted by the whole of the data, since less than the whole may be to any degree misleading. This, of course, is no reason against the use of absolutely precise forms of statement when these are available. It is only a warning to those who may be tempted to think that the particular precise code of mathematical statements in which they have been drilled at College is a substitute for the use of reasoning powers, which mankind has probably possessed since prehistoric times, and in which, as the history of the theory of probability shows, the process of codification is still incomplete.