# 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 $\theta$, 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, $\phi$ for that of a particular kind of sample, and $\Phi$ 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 $\Sigma$ stands for summation over the possible samples which yield the same estimate; and finally

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

where $\Sigma'$ 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 $\Sigma$ and $\Sigma'$.

If $T$ is distributed normally about $\theta$ 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 $\theta$.

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 $\phi$, 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 $\Sigma''$ 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,

$$X = S(kx),$$

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

If the coefficients $k$ are functions of $\theta$, 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 $\theta$. Differentiating with respect to $\theta$ 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 $\theta$. 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 $2 \times 2$ 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 $2 \times 2$ 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 $2 \times 2$ 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 $\psi$, and for each value of $\psi$ 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 $\psi$ is diminished, and reaches 1 per cent. when $\psi$ 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 $\psi$ 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 $\psi$. 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 $\psi$ 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 $\psi$ 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 $\rho$ is the density of the organisms to be estimated, then the density in the $n$th 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 $\rho$, 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 $n$th dilution, and all of the $n$th 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 $n$th. 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 $\rho$, 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 $\rho$ 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, $\theta$, is chosen at random between 0 and 1. In the first dilution, instead of the dilution ratio $a$ we use the dilution ratio $a^\theta$, using the dilution ratio $a$ for all subsequent dilutions. The probability of any particular configuration occurring is now wholly independent of $\rho$, 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 $\rho$, 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.

## Discussion on Professor Fisher's Paper.

Professor A. L. Bowley: I am not quite sure why I am called upon so frequently to open the discussions at this Society. This is the third time in recent months, and by the law of succession, the chance appears to be three to one that I may open it next time; but I do not accept this law, nor, I imagine, does Professor Fisher.

I am glad to have this opportunity of thanking Professor Fisher, not so much for the paper that he has just read to us, as for his contributions to statistics in general. This is an appropriate occasion to say that I, and all the statisticians with whom I associate, appreciate the enormous amount of zeal he has brought to the study of statistics, the power of his mathematical tools, the extent of his influence here, in America and elsewhere, and the stimulus he has given to what he believes to be the correct application of mathematics. The influence of his work in application to the experimental field is of the very first importance.

It is not the custom, when the Council invites a member to propose a vote of thanks on a paper, to instruct him to bless it. If to some extent I play the inverse rôle of Balaam, it is not without precedent; speakers after me can take the parts of the ass that reproved the prophet, the angel that instructed him, and the king who offered him rewards; and on that understanding I will proceed to deal with some parts of the paper.

The essence of the method of "likelihood," and its relation to earlier ways of approaching the problem of estimating properties of a universe from those of a sample, can be sufficiently appreciated by all those interested by studying Dr. Neyman's paper and the discussion on it in the last *Journal.* Both methods have their importance; the newer one, I think, in choosing the best arrangement of experimental work. Dr. Neyman says that "if all we need consists in the chance that, in the universe which we are sampling, the proportion is within given limits, we certainly cannot go any further than is already known" (p. 624). He also says, "we are interested in the probability of committing an error when applying constantly a certain rule of behaviour" (p. 624). But Professor Fisher claims (p. 40) that "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." And in an earlier place (p. 562) Dr. Neyman said that an approach to problems of this type, where the population is not known *a priori,* "removes the difficulties involved in the lack of knowledge of the *a priori* law." "It is superfluous to make any appeals to Bayes' theorem."

We are therefore left very much where we were, and I must confess that the new method appears to me to tell us only one-half of what we really need, for that *is* to determine "the chance that in the universe, which we are sampling, the proportion is within given limits." That seems to me the fundamental problem; but I had hoped that this subject would not have come up for discussion again to-day.

The chief problem of the earlier part of the paper, apart from the logical discussion on which I hope Professor Wolf will throw light, lies in pp. 42 (foot) to 46. I found the treatment to be very obscure. I took it as a week-end problem, and first tried it as an acrostic, but I found that I could not satisfy all the "lights." I tried it then as a cross-word puzzle, but I have not the facility of Sir Josiah Stamp for solving such conundrums. Next I took it as an anagram, remembering that Hooke stated his law of elasticity in that form, but when I found that there were only two vowels to eleven consonants, some of which were Greek capitals, I came to the conclusion that it might be Polish or Russian, and therefore best left to Dr. Neyman or Dr. Isserlis. Finally, I thought it must be a cypher, and after a great deal of investigation, decided that Professor Fisher had hidden the key in former papers, as is his custom, and I gave it up. But in so doing I remembered that Professor Edgeworth had written a good deal on a kindred subject, and I turned to his studies.

It is of little practical importance, even if it were generally possible, to determine who first gave expression to particular ideas or formulæ. But I wish to call attention to the similarity of part of Fisher's work to that of Edgeworth, who devoted much attention to the measurement of precision in his papers in the *Journal* of this Society in the years 1908 and 1909, especially in his description of "the genuine inverse method."

One of his most important formulæ, which I quote on p. 26 of my study of his work,* is indistinguishable from the value of $V$ that Professor Fisher gives on p. 46.

$$\frac{1}{V} = ni = I = nS\left\{\frac{1}{f}\left(\frac{\delta f}{\delta \theta}\right)^2\right\}$$

Edgeworth writes $n\int \frac{1}{w}\left(\frac{dw}{dx}\right)^2 dx$ for the last expression.

Edgeworth goes on, as I give on pp. 27–8 of my study, departing from what he called the "genuine inverse method," to endeavour to reach the same result of another path. "The result," he says, "can be obtained by a direct method free from the speculative character which attaches to inverse probability." This result is that the fluctuation (called variance by Professor Fisher) of an unknown "average" (statistic) is a minimum when the conditions of genuine inverse probability (method of maximum likelihood) are satisfied, and that the inverse square of the fluctuation is as stated above. It should be observed that Edgeworth gave a different form to the measurement connected with the fluctuation, as contrasted with an average; but I understand that in the paper before us only an illustrative problem is worked out, and it is not improbable that Professor Fisher has also obtained Edgeworth's second formula. In fact, in consideration of the paper read it is important to give close attention to the limitations of the application of the whole method and of the particular formulæ, and to study the hypotheses on which they rest.

* *F. Y. Edgeworth's Contributions to Mathematical Statistics.* Published by the Royal Statistical Society, 1928.

I must confess to dislike of the method or nomenclature that leads to such a phrase as "there 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." Or in other connections, such and such methods use only 90 per cent. of the information. Or the equating of "efficiency" to unity, and saying that a measurement is 90 per cent. efficient. The change from the intelligible statement that the variance of one estimate is $a^2$ and of another is $b^2$, where $b$ is less than $a$, to the phrase $b^2$ is 90 per cent. of $a^2$, does not add information and introduces misleading ideas. The measurement on this basis of the amount of knowledge seems to me to have the same dangers as treating the correlation coefficient or its square as the amount of covariation. In both cases a definite meaning is attached to the maximum called unity, and to the minimum called zero. In neither case does an intermediate value correspond, unless under special conditions, which may or may not be common, to anything otherwise definable.

The identification of minimum variance with maximum information appears to me to be arbitrary, and not of general application, even if it is appropriate to the class where the frequency of the estimate is normal. For example, the whole investigation depends on there being a definite algebraic function describing the frequency group from which the sample is drawn.

Finally, I should wish everyone to consider the "claim that mathematical likelihood supplies a measure of rational belief." (I take one phrase out of a conditioning sentence.) If, in fact, we knew nothing about a universe except that the variance measured in a particular way corresponded to a certain point on the normal curve of error, should we have any grounds for any rational belief, let alone a measurement of it?

Dr. Isserlis: The ground has been cut from under my feet in more ways than one. Professor Bowley, in proposing the vote of thanks, expressed in a very sincere way and in strong terms the appreciation which all of us who work in mathematical statistics have for Professor Fisher's contributions to the whole subject. We have in the *Journal* of the Society one or two papers by Professor Fisher, and I certainly welcome this first occasion on which he has read a paper before the Society and given so many of us an opportunity of hearing at first hand a summary of the special methods and the new concepts that he has introduced.

Before exercising the ordinary privilege of proposer or seconder on these occasions, of treating an author's paper somewhat critically, I should like to say at once that I think these new methods and concepts have very great practical value. There is no doubt in my mind at all about that, but Professor Fisher, like other fond parents, may perhaps see in his offspring qualities which to his mind no other children possess; others, however, may consider that the offspring are not unique.

I started by saying that the ground has been cut from under my feet by Professor Bowley in two ways. The first I have referred to;

the second is the very lucid way in which he referred to the dependence of the idea of likelihood on a certain narrow field, so that it can only be applied in the case of investigations in which the parameter that one seeks is distributed in samples in some simple way.

There is a third point in which I feel a certain difficulty, which comes in the earlier portion of the paper, dealing with the general subject of inductive inference. My style is cramped because Professor Wolf is sitting next to the seat I have just vacated. The criticism of that part of the subject will have to be undertaken by him as a professional logician.

Man is an inductive animal; we all generalize from the particular to the general; in all branches of science, and not only in statistics, it is the business of those of us who have devoted some attention to our own branch of the subject, to try and act as guides to our followers in preventing rash generalization.

Speaking as a mathematician as well as a statistician, I find it rather difficult to follow the paragraphs on p. 39 of the paper where Professor Fisher tells us that mathematicians trained in deductive methods are apt to forget that rigorous inferences from the particular to the general are even possible. I do not think that is the case with the ordinary mathematician. It may be that in mathematical analysis the fundamental inductions on which the analysis rests are rather remote, but they are there all right, and no mathematician may proceed safely with his work unless he is strongly aware of their existence.

A good deal depends upon how rigorous we are in interpreting the word "rigorous" as used by Professor Fisher. A mathematical statement is surely rigorous when it is a probability statement, just as when it is a statement in ordinary analysis. Let us take the distinction between the problem of learning something about possible samples when the universe is known, and learning something about the universe when all that we know is a sample. If an ordinary pack of 52 cards is dealt, and I get a hand of 13 cards, this is a case where the universe is known, and the question can be put to me, "What kind of a hand do you expect to get? Are you going to get a hand containing three hearts?" I make a statement with regard to the probability that my hand will contain three hearts; that is a prefectly rigorous statement and as precise a statement as we can make. I start by saying at once that I do not know how many hearts I will have in my hand, and that if you ask me to give an estimate of the number, you are putting the wrong question. The right question is, "What kind of rigorous statement can you make?" I am in a similar position if the story is the other way. A pack of 52 cards is taken; I do not know the composition of that pack; a hand of 13 is dealt to me and I find in my hand three hearts; then I ask the question, "What information does this give me with regard to the nature of the pack from which the cards were dealt?" One answer is, "The probability that the pack contains less than 25 hearts is greater than 80." * In making this answer I do not use the

* $P(3 \leq x < 25) > 0{\cdot}801$.

method of inverse probability and make no estimation of a distribution of *a priori* probabilities; it is what Professor Fisher would call a pure deduction from theory. The statement is not as good as it might have been owing to my laziness in doing arithmetical calculations; if I had used more complicated known results I should have been able to do one of two things; I could have had a smaller number than 25 inside the bracket, or I could have a bigger decimal than 0·801 on the right-hand side of the inequality.

If I understand their work correctly, Dr. Neyman and Dr. Pearson have given us many results of this kind, their attention being sometimes directed to the various numbers that can be put inside the bracket when the probability outside is fixed, and sometimes to the various probabilities outside the bracket when the limits are fixed inside it.

I do not mean to say that I agree with Professor Fisher that the method of inverse probability must necessarily be rejected. A good deal of work has been done in showing that in the cases that matter, the extended form of Bayes' theorem will suggest as a reasonable estimate very much the same value as is suggested by the method of maximum likelihood.

If I may detain you a few moments longer, I should like to refer to the portions which Professor Fisher did not read, and to the very interesting table given on p. 48. Looking at that table without examining the first two lines, it might be said, "This table suggests that the monozygotic brother of a twin is seven times more likely to be a convict himself than would be the case with a dizygotic brother. A ratio of about 6 or 7 is suggested. If in that table there were no correlation at all, one would expect 7·5 individuals in each column of the table. If, on the other hand, the marginal frequencies in the table were forced upon one, then in the S.W. corner one would expect not 7·5 but 6·10. As a matter of fact there are only two individuals in that table, and it is this characteristic of the table which would lead the ordinary man to come to his conclusion.

Professor Fisher notes the fact that the S.W. corner of the table contains only two individuals, and asks what are the circumstances which would lead to the probability being of a fair size that that corner should contain so few as 0 or 1 or 2? His answer is, that the circumstances would be suitable if a monozygotic twin brother had a probability 2·4 times greater than a dizygotic brother of being himself a convict and the probability of such suitable circumstances is greater than $\frac{1}{100}$.

With that sort of thing I am in full agreement. I do not think that it introduces any new concepts. The only thing that puzzles me is why it should be necessary to use new terms or to suggest that we cannot make probability statements which are as rigorous as those which are made by any of our confrères who work in the natural sciences.

That is all I have to say excepting, of course, that I do sincerely join with Professor Bowley in his motion that a vote of thanks be given to Professor Fisher for his interesting paper.

Dr. Irwin said he happened recently to be reading that classical old book, Todhunter's *History of the Theory of Probability*, in which he came across the following passage. "Dr. Bowditch himself was accustomed to remark, 'Whenever I meet in Laplace with the words, "Thus it plainly appears," I am sure that hours, and perhaps days of hard study will alone enable me to discover how it plainly appears.'"

To those who had been familiar for some years with Professor Fisher's work, the first five pages of this paper bore a certain air of familiarity; they were no longer thrown completely into confusion by the notions of efficient and sufficient statistics, and were not entirely startled when they found that if there was an efficient statistic, it could be estimated by the method of maximum likelihood. Perhaps it was on that account that he was rather lucky with pp. 43–4 and he found that he got through that without any undue amount of difficulty. It was true that he was rather stuck when he came to the calculus of variations on p. 46, but was not unduly perturbed thereby, because he had discovered another proof that Professor Fisher had given elsewhere which did not involve the calculus of variations. He could well see, however, that anyone coming to the subject for the first time would find in this paper what would appear to be gaps—not so much gaps in the reasoning, but gaps in his own mental processes which he could not at the moment fill up. For example, if Professor Fisher's definition of intrinsic accuracy were accepted and if one were to define the information in a single observation as being equal to the intrinsic accuracy, then a new-comer to the subject would probably feel inclined to say, "Yes, but we have certainly to show that if we adopt that definition, and combine the information from two or more observations, we get the same answer as if we were getting the information from the pair of observations directly." In other words, what the new-comer had to do, and what was not done here, was to write down the distribution of a pair of observations, use Professor Fisher's definition to get the intrinsic accuracy of the distribution of that pair in estimating the parameter, and see that it was equal to the sum of the intrinsic accuracies obtained from the distributions of the two observations separately. As soon as he realized that that was the problem, he would have no difficulty whatever in writing down the solution, and of course Professor Fisher has done it. It followed in a line or two of algebra in his paper on the theory of statistical estimation, but Dr. Irwin thought that to a person coming fresh to the subject, that might not be obvious at first sight.

Of course in this paper Professor Fisher covered a very large amount of ground—the ground of the whole of his ideas of the last ten years—and this naturally led to considerable condensation.

Dr. Irwin thought that the fact that he felt fairly familiar with the first half of the paper was due to the many hours of labour he had tried to put in filling in the gap where Professor Fisher had said "It is easy to see that. . . ."

When he came to the second part of the paper where there were ideas with which he was less familiar, there were still one or two gaps which in his own mind he could not immediately fill. For instance, in the question of ancillary information it was not absolutely clear how one should define an ancillary statistic. On pp. 48 *et seq.* it would be found that the margins were used as ancillary information for the purpose of testing the hypothesis that the proportion of convicted criminals in the two classes, monozygotic and dizygotic, was the same. That problem Professor Fisher pointed out had nothing whatever to do with estimation, but earlier on p. 48 an ancillary statistic was defined as something which told one nothing about the value of the parameter, but told how big an estimate had been made of it. Obviously the margins were not ancillary statistics quite in that sense, and Dr. Irwin felt sure that all would be grateful to Professor Fisher if he would help to clear up that little confusion, that perhaps existed in his own mind only, due to the fact that he had seen the paper for the first time on the previous day.

There was one other point. On p. 52 he had been able to follow down to where Professor Fisher said: "The probability of any particular configuration occurring is now wholly independent of $\rho$." But Professor Fisher continued: "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 $\log \rho - x \log a$." Dr. Irwin had not been able to get there yet, and he would be grateful if Professor Fisher would help him to short-cut that process.

These were the thoughts that occurred to him in reading the paper, but he felt certain that when they had had time to digest it thoroughly they would find new ideas behind it as stimulating and as useful as they had always found Professor Fisher's ideas to be in the past. He would like to join the proposer and seconder in thanking Professor Fisher very heartily for an interesting and stimulating paper.

Professor Wolf thanked the President for inviting him to listen to this paper and the very instructive discussion, and for allowing him to take part in the discussion. He was not a mathematician, nor a statistician, and he could not, therefore, be expected to make any contribution towards the mathematics of the paper, but he had all his life been interested in the study of scientific method. Unfortunately there were very few men of science who had ever seriously thought about the basic methods and principles of science, or, at all events, who had published their reflections upon the principles which underlay their scientific investigations. Therefore when he came across men of science who had the courage to do that kind of thing, he wanted to thank them very gratefully, and he did thank Professor Fisher.

So far as he could make out, Professor Fisher had proposed a very ingenious method of making a little evidence go a very long way, by introducing certain qualifications, consisting of estimations of reliability, of the conclusions. With regard to the particular points stressed in the paper, he would like to ask what was the net result of these estimates to be? Were these estimates finally to be merely of a subjective value, or were they intended to have an objective, scientific character? What he meant by this would be obvious if he took the case of the theory of probability. So far as he was concerned, he had maintained for many years that there were both types of estimates of probability, the deductive and the inductive calculation of probability; but from a scientific point of view he believed that the real value lay in the knowledge of the frequencies. In inductive calculations one started from the sample frequencies, and deduced their probabilities. In the deductive calculations one started from the *a priori* probabilities, and from these it was possible, more or less securely, to deduce the probable frequencies. But, in either case, the real scientific value lay in the frequencies rather than in the probabilities.

Estimates of probability seemed to be more of psychological, rather than of general scientific, importance. When he compared different fractions of probability as the measure of what his rational belief ought to be, he found it impossible to adjust his belief to these different fractions. Even subjectively, therefore, calculations of probability seemed unimportant. He could not find any real, scientific, or strictly objective significance in probabilities as such.

When he said that measures of probability were a matter of psychological or subjective interest, he realized, of course, that they were logical in character, and therefore, in a secondary sense, objective, that is to say, they were not capriciously subjective; but nevertheless it remained true that he did not find it within his competence to adjust his degree of rational belief to the different requirements of the different estimates of probability. In the light of these considerations concerning the nature of probability, he would like to ask Professor Fisher to explain and make intelligible to him what exactly the character of the proposed estimates of reliability was to be? Was it conceived to be objective or subjective, and what was its exact function to be in science?

Professor Wolf said he did not propose to add any comments on the more limited problems with which the lecturer had dealt. He was more interested in the wider problem suggested by the title of Dr. Fisher's paper, namely, the general problem of the logic of induction. It was gratifying to him personally to find that Professor Fisher repudiated the old idea that the whole of induction was based on the calculation of probability. Two or three decades ago that was more or less the prevalent conception of induction.

Nearly twenty-five years ago Professor Wolf wrote a paper on the Philosophy of Probability, in which he had repudiated this idea, and he had tried to show that, so far from induction being based on probability, there could be no rational calculus of probability without the postulates of ordinary induction. Statisticians were sometimes apt to forget that there were advanced sciences long before there was such a thing as a science, or rather a method, of statistics. Galileo and Newton made their exact and far-reaching discoveries while Graunt, King, and Petty were still struggling hard with the elementary arithmetic of statistics, and dealing with simple averages. It was good to find that Professor Fisher discouraged the tendency to exaggerate the place of statistics in induction generally. Dr. Wolf related that his early paper on the Philosophy of Probability had brought him into disgrace in the eyes of an eminent statistician. Not many years later, however, Mr. Keynes also repudiated this view of the basis of induction; and now he was glad to find Professor Fisher did the same.

With regard to some of the misapprehensions which underlay the older conception of the statistical basis of induction, it was not quite clear whether Professor Fisher was entirely free from them, in spite of the fact that in one place he distinctly repudiated them. The storm-centre lay very largely in the conception of mathematics and of its place in science. There was the familiar idea that pure mathematics was entirely deductive; and a great many people held that view. The conception that probability was at the base of all induction was largely the progeny of this conception of pure mathematics. The idea underlying that belief was that pure mathematics was exact and absolutely reliable; it did not make any assumption of an inductive character, and was therefore qualified to serve as a basis of inductive inference. Professor Wolf was very doubtful about this. He did not believe that pure mathematics was purely deductive. There was induction in mathematics, but it was slurred over. Owing perhaps to bad teaching, encouragement had been given to the assumption that mathematics was all deductive, and not at all inductive. How was it that mathematics has thus come to be associated solely with deduction?

The misapprehension was probably due to three contributory factors. (1) The idea was upheld partly by Descartes, who played such an important rôle in the whole development of modern mathematics that his word was accepted without challenge. But if one studied Descartes' use of the term "deduction" it would be seen that he did not use it in the ordinary sense of inference from general propositions, definitely accepted, or assumed provisionally; he used it in a much more complicated sense, which included a good deal of induction.

(2) People were still frequently using the term "deduction" not in its ordinary sense—"inference from the general to the particular or to the less general"—but for inference of any and every kind. A common phrase was, "What deductions do you draw from these facts?" Deductions (properly so called) were not drawn from facts; "inferences" was the word that should be used in such contexts. There was thus a very common use of the term "deduction" for "inference"; and people did not always realize that they were talking about inference in general, and not about deduction in particular.

(3) A third point was perhaps even more important. Mathematicians and scientists generally did not realize sufficiently that in what was called "inductive inference" there was nearly always a moment, or stage, which was deductive, namely, the stage where the hypothesis had to be verified, and this was done by application to suitable cases of the hypothesis, which was a general statement accepted as possibly true. That stage was purely deductive, yet the investigation as a whole was essentially inductive. It was not sufficiently realized that although there might be deductions without

inductions, there could not be—except in very rare cases—induction without a deductive moment or stage. In mathematics, no doubt, the deductive moment loomed very large, and so people jumped to the conclusion that the whole of mathematics was deductive. Professor Wolf did not accept that view; and as soon as it was realized that even mathematics was partly inductive, one could see for oneself that mathematics, or any part of it, could not be made the logical basis of all other forms of induction.

To pass to another point, Professor Wolf sometimes wondered whether the tendency to exaggerate the importance of mathematics, and especially the theory of probability, in inductive science was not due to a very large extent to the disbelief, on the part of the exponents, in the possibility of induction altogether; whether, in fact, it was not due to their conception that not only was so-called "probability" a subjective matter, but that the whole of scientific inference was mainly the subjective play of the human mind attempting to amuse itself, or to satisfy itself, by means of man-made conjectures which might not reflect reality at all. Mr. Bertrand Russell in one of his latest books has made this idea perfectly clear. He has said that, for all that was known, natural phenomena might contain no order at all, and that it was only the cleverness of mathematicians which imposed on Nature an *appearance* of order. Although he was not a mathematician, Professor Wolf did not believe that Mr. Russell could discover a formula showing order among phenomena utterly disordered. Here was a tendency to exaggerate the importance of mathematics, coupled with scepticism as to the real objective value of science—a scepticism as to the real existence of orderliness among natural phenomena. To some extent the same tendency might be found in Professor Karl Pearson. On looking at his *Grammar of Science* it would be seen how he was smitten with Kantian philosophy interpreted in such a way as to make all knowledge the invention or creation of the mind, so that the orderliness that was found in Nature was simply the orderliness which the human mind imposed upon natural phenomena.

Professor Wolf again expressed his thanks to Professor Fisher for coming out into the open in his most interesting paper.

DR. E. S. PEARSON said that there were a number of points he would have liked to have discussed; at this late hour, however, it would be better for him not to go into them, but rather (if he might) add a written contribution to the discussion in the *Journal*. There was, however, one point that he would like to mention now. At the beginning of the paper Professor Fisher had said that he regarded the essential effect of the general body of researches in mathematical statistics during the last fifteen years to be fundamentally a reconstruction of logical rather than mathematical ideas. Dr. Pearson agreed with that statement, but he rather gathered that Professor Bowley and Dr. Isserlis did not. It seemed to him that Professor Fisher had contributed to this development of logical ideas something which was definitely lacking before. When these ideas were fully understood, whether there was final agreement or not with his particular terminology and the details of his theory of inductive inference, it would be realized that statistical science owed a very great deal to the stimulus Professor Fisher had provided in many directions.

He spoke of himself as one also very keen on the development of the logical processes of reasoning, and if he should appear critical it was not because he regarded Professor Fisher's theory of estimation as unnecessary or unoriginal, but because it seemed to him that there were certain important statistical problems to which there were other rather simpler and more direct methods of logical approach. In his further contribution he would try to make clear what he had in mind.

[Dr. E. S. Pearson's further contribution follows.]

Professor Fisher has emphasized and illustrated again and again the fundamental point, whose significance was not before, I am certain, fully realized, that while there may be many ways of using a given set of statistical data, it is both possible and desirable to set about in a systematic manner determining how best to make use of these data for the purpose we have in view. His researches have provided an extraordinary number of interesting ideas and conceptions, and in a broad sense I think it would be difficult to dispute the value of this idea, which he has discussed specially in the present paper, of extracting from the data the maximum amount of information. With regard, however, to the particular mathematical definition of the "amount of information," the position is a little different, and I personally, while recognizing the fascination of a well-rounded theory, am not yet convinced that it is of quite so far-reaching importance as Professor Fisher believes. This is largely because there are certain very important statistical problems the approach to which can be made, so it seems to me, by a simpler and more direct route.

In the course of teaching statistical theory one becomes after a time rather sensitive to shortcomings in logical reasoning; one realizes, both in one's own development of a subject and in that of other writers, that there are certain steps in argument which it is difficult to get across to students without blurring the issue. I think perhaps that my own approach has been forced upon me because I have found apparent gaps in argument in the writing and teaching of others. To me there are such gaps in Professor Fisher's philosophy which I cannot bridge; the bridges that I have built are not his bridges, but there has been no alternative. I cannot, for example, be clear of the exact form of his logical connection between the theory of estimation and his tests of significance, yet it is clear that these latter must fall within the scope of any theory of inductive inference.

The problems with which the mathematical theory of statistics is largely concerned may perhaps be divided into three categories:

*Category* 1. Problems in which there is to be obtained from the data a *single estimate* of each of one or more parameters.

Professor Fisher's theory of estimation was primarily developed to deal with problems in this field.

*Category* 2. Problems in which it is desired to *estimate an interval* within which we may have confidence that the value of the unknown parameter lies. These were the problems that the old probable error theory attempted to tackle by giving as the interval that which lay between

single estimate $\pm$ 2 (or 3) $\times$ probable error

For large samples this method of definition of the interval was good enough for most practical purposes, but it failed when applied to small samples, and I fancy that it was in any case lacking in logical completeness. Recently there have been a number of contributions, of which Professor Fisher's was the first,* developing a new and, it is suggested, sounder method of solving many such problems. Professor Fisher terms these problems in *fiducial probability*, but I think that the method may be described as a new form of the method of estimation. Also I believe that this form of solution in terms of an interval is fundamentally more important than the solution which gives a single valued estimate. The two solutions are clearly closely interrelated, but may not the simplest conception from which to approach the interval problem be based upon

(*a*) the breadth of interval,

(*b*) the risk of error in the assertion that the value of the unknown parameter falls within the interval,

rather than on some definition of the amount of information in the data?

But whatever may be recognized ultimately as the simplest method of approach, the pathway is still in the making and I have been much intrigued by the suggestion, contained in Professor Fisher's second example, of a device for coping with the case in which the variable considered is discrete rather than continuous.

*Category* 3. Problems of testing hypotheses; these fall under Professor Fisher's heading of *tests of significance*. It is here that I am in real doubt in trying to follow Professor Fisher's logical approach. Let me illustrate my difficulty.

In his example, Professor Fisher is in the first place testing the hypothesis that $p = p^1$, where $p$ and $p^1$ are the chance of a monozygotic and dizygotic twin being criminal respectively. He measures the significance of the departure from proportionality by summing the chances that a quantity $x$ has a value equal to or less than that observed, *i.e.* that $x = 2, 1$, or $0$.

A student of the logic of statistical inference might well ask, "Why sum these three terms?" It is a type of question which I have, in fact, often been asked, and I think the answer I should give would be somewhat as follows:—

As $x$ decreases from 6 towards 0 the likelihood of alternative

* *Proceedings of Cambridge Philosophical Society*, 26 (1930), p. 528. See also Dr. Neyman's paper and the discussion following, contained in the last part of this *Journal*.

hypotheses with $p > p^1$ becomes greater and greater, compared with that of the hypothesis $p = p^1$. At a certain point we shall feel inclined to reject the latter in favour of some one of the former hypotheses. In determining that point we shall be influenced by the following consideration: we do not wish to incur a large risk of rejecting the hypothesis $p = p^1$ when it is really true. To measure this risk we sum the tail terms of the series of probabilities of $x$ assuming values 0, 1, 2 . . ., when $p = p^1$.

In this argument we take into account (*a*) the alternative hypotheses, and (*b*) the risk of making a wrong judgment. These are simple but, of course, not necessarily the only conceptions on which an answer could be based.

Take next a situation in which a more careful answer would be needed. Suppose that a sample of $n$ is drawn from some normal population and it is wished to test the hypothesis that the population variance, $\sigma^2$, is unity; further, suppose that as an alternative to this, $\sigma^2$ might be either greater or less than unity. How should we proceed? Following the usual practice we should consider the sampling distribution of $s^2$ (the squared standard deviation in the sample) for the case $\sigma^2 = 1$, and from this determine two values $s_1{}^2$ and $s_2{}^2$, cutting off equal tail areas. If the two 1 per cent. levels of significance are taken, then (if $\sigma^2 = 1$) the chance is ·01 that $s^2 < s_1{}^2$ and ·01 that $s\sigma^2 > s^2$. We might then decide to reject the hypothesis $\sigma^2 = 1$ unless $s_1{}^2 \leq s^2 \leq s_2{}^2$.

If, however, we adopt this course a curious result follows. Take by way of illustration, the case of a sample of $3 (n = 3)$; then if $\sigma^2 = 1$ the chance of our rejecting this hypothesis that we are testing is, of course, ·020. But suppose that in fact, though we did not know it, $\sigma^2 = 0{\cdot}75$, then it can be shown that the chance of $s^2$ falling outside the range $s_1{}^2$ to $s_2{}^2$ is now ·015; that is to say, we are more likely to reject the hypothesis we are testing when it is true, than to reject it when certain alternatives are true! This result suggests that from the logical point of view there may be something wrong in this customary approach; it appears, in fact, that when dealing with tests of hypotheses it is not enough to know the statistic appropriate for the solution of problems in estimation, and then to calculate levels of significance.

Let me make it clear that I am not here criticizing the theory of estimation, but asking whether there may not be among statisticians a too ready assumption that the conception of "levels of significance" needs no explanation. Is not a logical analysis of each step in the framework of argument as necessary when dealing with problems falling into categories (2) and (3) as in the case of category (1)? Professor Fisher's philosophy of inductive inference must, I am sure, embrace all these aspects of making sense of figures, but I for one am left uncertain as to whether he is dealing with the whole or only a part of that philosophy in the present paper. He has set out clearly much that I was uncertain of before, and it is perhaps ungrateful to ask him for more, but as he has suggested in an early paragraph of the paper that he is dealing with the whole field of logical situations at present explored, I feel that this question should be raised.

Professor Greenwood said that at this late hour it would not be fair to call upon any other speaker, but he would like to add to what Dr. Pearson had said and repeat that the speakers in the discussion had full liberty to amplify in their written report points that could not be taken up now; it would be extremely unfair to ask Dr. Fisher to reply in other than a formal way at this late hour. Written communications had been received from Dr. Jeffreys and from Dr. Neyman, and these would be given to Dr. Fisher.

Before putting the vote, which he felt sure would be carried unanimously, Professor Greenwood said he wished to add his tribute to those already paid to Dr. Fisher.

Lord George Hamilton, a predecessor in the chair, said that the Society ought for dignity's sake to make a rule that at least one paper in each session should be unintelligible to all those present save the reader and the mover and seconder of the vote of thanks. On the present occasion Lord George Hamilton's rule had been broken; others than the mover and seconder had displayed intelligence, yet Dr. Greenwood felt he might not be the only person present to whom much which had been said had been difficult to grasp and, since he had neither the technical knowledge nor the intellect to express his gratitude to Professor Fisher in the way he would most value, viz. by an expert criticism, he would pay a small dividend of it by trying to explain how, as he thought, Professor Fisher might help, still more efficiently than he already did, the very large number of statisticians who wished to use intelligently the instruments he had fashioned for descrying the truths hidden in numbers.

In the first place, he suspected that Professor Fisher's nomenclature had not been very helpful to the layman. He imagined that Professor Fisher recoiled from the Victorian practice of coining Greek vocables—a practice which gave occasion for a cruel practical jest in a sister learned society. But perhaps the introduction of what rude people called "gibberish" was less confusing than attaching particular meanings to words well established in the current speech of educated people. It did not, perhaps, give people much difficulty to distinguish between variance in the sense of the second moment coefficient and in the more usual sense of the attitude of any one mathematical statistician to any other mathematical statistician. But a confusion between statistics as the object of their pious founders and as a Fisherian plural was more troublesome. This, however, was only a trifle. The Galton Professor might surely claim the right exercised by Humpty Dumpty.

More serious was Professor Fisher's extreme reluctance to bore his readers—surely a defect rare in statisticians. He seemed to be a little over-anxious not to incur the sneer of—whom?—perhaps of some of the speakers that evening—that something he had said was "obvious" or "self-evident." He was in a little too much danger of dichotomizing his public into a tiny class of persons who were his intellectual peers, and a much larger class of persons who were to behave like the gallant six hundred. A trichotomy was practicable, and the class for which he did not yet cater was important. Professor Greenwood did not himself feel learned enough to illustrate this properly, but he would try to do so. He would refer to the problem discussed on pages 48 and those following. He *thought* he understood the steps by which Professor Fisher reached his conclusion. That the marginal frequencies should be preserved, the binomial exponents must be $p^{12}$ and $(1 - p)^{18}$, and on that condition it was not necessary to make any hypothesis at all as to the value of $p$. But what he could not do was to demonstrate that the approach to the problem which, he supposed, would be made by an ordinary unmathematical student whose knowledge of the general subject went no further than, say, the chapter on Probability in an ordinary text-book of algebra, was illogical. He thought such a person would reason in the following way. The probability that he would draw a sample of 13 with 10 black balls and a sample of 17 with 2 black balls, was $13!/10!\,.\,3!$ multiplied by $17!/2!\,.\,15!$ multiplied by $p^{12}(1 - p)^{18}$, and this product would be a maximum if $p = 12/30$. He would take that value of $p$ then, and would multiply the tails of the two binomials with exponents 13 and 17, viz. the sum of terms giving 10, 11, 12 and 13 convictions for one by the sum of terms giving 2, 1 or 0 convictions for the other. This product would measure the probability of as bad a result as he observed or a worse result on the hypothesis most favourable to concurrence. Naturally he would not get Professor Fisher's answer, but a smaller value. He would have used the vertical marginal totals to obtain a value of $p$ but have disregarded the totals in his working out of probability. But he might retort that he was testing the hypothesis that these two samples did come out of the same bag and that the constitution of the bag assumed *was* the constitution most favourable to the assumption that they did come from the same bag. Why was this an improper test? Professor Greenwood was not sure he could give a satisfactory answer. He was sure Professor Fisher could, and he thought readers of the class he had in mind would be grateful to him if he would proceed by steps; if he would begin by taking the "naïve" method of approach and by pointing out why it was unsatisfactory and why another method was more logical. This might be, and no doubt was, utterly unnecessary for trained mathematical logicians, but they were a small class and could certainly look after themselves. It might be that the rationale of the most precise methods of statistical analysis could only be made intelligible to a small minority of naturally gifted and highly-trained minds; but he hoped this was not true, because if it were, then, apart from a tiny minority, only those who were willing to be intellectual helots would use the best statistical methods. It might be said that biologists who knew nothing of applied optics could and did use efficiently compound microscopes; why, therefore, should they not use statistical formulæ the construction of which they did not understand? The answer was that by very little use a biologist could distinguish between a well- and an ill-made microscope. But without a remarkable pooling of laboratory experience it would be hard to determine that Mr. A.'s method of evaluating a probability was more efficient than Professor B.'s even if Mr. A. systematically over-estimated the probability by a considerable amount. In truth, he had a higher opinion of Professor Fisher's powers than Professor Fisher had himself; Professor Fisher was a great investigator; he might be a great teacher too.

The following contributions to the discussion were received in writing:—

DR. HAROLD JEFFREYS: I should be interested to know the source of Professor Fisher's remark that in the theory of inverse probability the method was to introduce a postulate concerning the population from which the unknown population was supposed to be drawn. I have not come across it in a fairly wide reading of the subject. It was certainly not the procedure of Bayes, Laplace, W. E. Johnson, or F. P. Ramsey, nor is it that of Keynes or myself. The estimation of the distribution of probability corresponding to previous ignorance is perfectly direct and rests on no postulate beyond the fundamental notion of probability itself. The difficulties on this point seem to me to have arisen partly from the hesitation to say just what it is that one does not know, and partly from the insistence of opponents that when a man has said " I do not know " he must necessarily mean something else. Most writers take probability as a primitive idea, but Bayes and Ramsey define it in terms of expectation; the latter procedure has some advantages, but the results are the same either way.

Professor Fisher seems to set up his use of likelihood in opposition to the theory of probability. I cannot see why he does this, since the theory of probability provides the use of likelihood with its best justification. The theorem of inverse probability states that as between different hypotheses, the posterior probability is proportional to the product of the prior probability and the likelihood. When the number of observations is large the likelihood has a sharp peak, outside of which its values are much smaller than that at the maximum. Outside a limited range of hypotheses the posterior probability is negligible on account of the small likelihood, while in ordinary cases the prior probability varies so little within the range that it may be treated as constant. Hence we are entitled to say that if the number of observations is large the posterior probability is proportional to the likelihood, within the range of hypotheses that have enough probability to be of any interest.

As a matter of fact I think that Professor Fisher's argument would be made much easier by an explicit use of probability. Thus if $T$ is normally distributed about $\theta$ with variance $V$, the probability of a sample value $T$ in a range $dT$ is

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

If we have $n$ samples yielding values $T_1 \ldots T_n$, their likelihood is

$$(2\pi V)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2V}\Sigma(T-\theta)^2\right\} \Pi(dT) = (2\pi V)^{-\frac{1}{2}n}$$
$$\exp\left[-\frac{n}{2V}\left\{(T_m-\theta)^2+\sigma'^2\right\}\right] \Pi(dT)$$

where $T_m$ is the mean of the observed values of $T$, and $\sigma'$ their mean square departure from this mean $\Pi(dT)$ does not involve the hypotheses, and the likelihood is therefore proportional to

$$(2\pi V)^{-\frac{1}{2}n} \exp\left[-\frac{n}{2V}\left\{(T_m - \theta)^2 + \sigma'^2\right\}\right].$$

The prior probability by definition does not involve the observations, so that the contribution to the posterior probability made by the observations is wholly summed up in the quantities $T_m$ and $\sigma'$. Consequently we need calculate only these two quantities; no other can give any additional information, and if we attempt to replace them by any other functions of the data we necessarily sacrifice accuracy. Fisher has discussed the point at much greater length, but I cannot see why he has thought it necessary. In his present paper he averages with regard to $T$, for no very obvious reason except that it enables him to get an answer; but the probability method gives a quite definite procedure. In most practical cases we are interested in $\theta$ much more than in $V$; what we want to know is the probability that the true value of the quantity we are trying to measure lies within certain limits. Consequently we formally fix these limits, and add up the posterior probabilities that $\theta$ may lie between these limits given by all the values of $V$; in other words, we integrate the posterior probability with regard to $V$ to get that of $\theta$ by itself. The likelihood takes us a long way, but the theory of probability finishes the job.

The contingency problem treated by Professor Fisher can be dealt with on the same lines. It resolves itself into two alternatives, each with various sub-alternatives. If $p$ and $p'$ are the proportions in the two populations sampled that possess the property sought for, the first question is, is $p'$ equal to $p$? If the answer is yes, then we have to ask further, what is the distribution of the probability of $p$ among its possible values? If the answer is no, we have to ask for the distributions of the probabilities of both $p$ and $p'$ among their possible values. Now if anybody asks a question it is a fair presumption that he does not know the answer; the prior probabilities that $p'$ is equal to or different from $p$ are the same, both being $\frac{1}{2}$. If $p' = p$, the prior probability of $p$ is uniformly distributed; if not, those of $p$ and $p'$ are both uniformly distributed. Then it is a straightforward application of the theory to find the distribution of the posterior probabilities among the various possibilities. Those of all values of $p$ on the hypothesis that $p' = p$ must be added up to give the posterior probability that $p' = p$; those of all values of $p$ and $p'$ on the hypothesis that $p$ and $p'$ are different must be added up to give the posterior probability that they are different. It is found that if the two sampling ratios are nearly equal, the posterior probability that $p$ and $p'$ are equal is very high; if they are very different, it is very low; the results are in accordance with what we should expect. In the case in question the ratio turns out to be

$$\frac{12!\ 18!\ 14!\ 18!}{2!\ 3!\ 10!\ 15!\ 31!} = \frac{1}{273{\cdot}7}.$$

The use of the fiducial probability in place of the posterior probability seems to me to be open to a number of objections. It shows insufficient respect to the observed data and does not answer the right question. When the actual difference in the two sampling ratios is given exactly, the possibility that a greater difference might have been obtained seems irrelevant; but actually it is these greater differences that contribute most of the fiducial probability. I think that when a question is proposed for statistical solution the questioner is always saying that he does not know the answer to certain questions and that he wants the answer in terms of the posterior probability with respect to the observed data. If he accepts the fiducial probability as an answer it is because he mistakenly interprets it as a posterior probability.

MR. M. S. BARTLETT.: I should like to take this opportunity of asking Professor Fisher about a difficulty that I have experienced in connection with the initial postulates of the theory of statistical probability, on which the methods he has summarized for us in his paper are based. I trust that he will understand that it is my very real appreciation of the value of these methods that has made me anxious to be sure in my own mind what is to be the starting-point for their derivation.

If I understand him correctly, he would divorce probability theory entirely from the subjective theory of degrees of rational belief postulated by advocates of inverse probability; but I confess I am not yet certain as to the implications of this step.

Statistical probability, and probability as used in any science incorporating statistical ideas, is associated with the idea of taking a random sample from a specified population. It will, I think, be clear that the use of the word "random" here already implies the notion of equal objective probabilities or chances. It seems to be, therefore, that one must either accept the laws of probability to some extent in an empirical way for these equal probabilities, or else refer them back, as I have tended to do myself, as a particular though highly important type of probability in the subjective theory, the use of which there has been so much reason to criticize.

I should like to add that I see no objection to accepting them as they stand, with any corresponding definition of probability already assuming the usual probability laws. Professor Fisher's own definition, based on the limit of frequencies, I would personally place in this category. In the experimental sciences it is not, as far as I am aware, regarded as a drawback that the definition of any physical measures, such as the weight of a body or its temperature, or the electrical resistance of a conductor, should assume the laws of physics that have given rise to the concepts corresponding to those measures. I do not know whether I have made myself sufficiently clear for Professor Fisher to realize the point I am trying to make, or whether he himself finds any difficulty here. I hope I have not misrepresented his own views with regard to these preliminary stages in the logical development of our subject.

DR. JERZY NEYMAN: There is an interesting detail which may be observed in the history of almost any mathematical science. Starting with several problems of purely practical character, the branch of mathematics is being developed primarily as a set of methods useful for the solution of similar problems. These methods are then compiled in theories already forming what could be called a science. However, the early theories contain usually many gaps and inaccuracies. When these are noticed a new period comes, in which a marked effort to penetrate into the depth rather than into the width of the science may be observed. This is a period of criticism and reconstruction of the existing systems or of systems about to exist. Such was approximately the history of modern mathematical analysis. After the works of Newton and Leibnitz there was a considerable period of solving problems. The end of the nineteenth century started the period of criticism and foundations.

Mathematical statistics is following the same steps. It may be considered as born as an independent science about forty years ago with the first papers of Karl Pearson. Now the period of "solving problems" is over. The next period of criticism and laying foundations has been started by R. A. Fisher in his *Phil. Trans.* paper of 1921. A series of other revolutionary papers followed and to-day we are discussing one of them.

These papers of R. A. Fisher have been criticized. But this is comprehensible: the first efforts in one direction contain but rarely the last word to be said, and any critical review of basic ideas is always unpleasant and difficult to be accepted by those who are perhaps too much attached and accustomed to these ideas. Considering the position we have to remember the silent but eloquent critics buying and buying the fourth (since 1925!) edition of R. A. Fisher's book. It is not the only one on the market. It is born of the ideas expressed in Fisher's theoretical papers. *Vox populi* . . .

Professor Fisher's papers are interesting not only because of the many important problems stated and solved, but, perhaps still more, because they contain so many hints and questions which the author did not have time or perhaps did not care to solve himself. Therefore, going through this paper one has not only to follow the writer's ideas, but is, as it were, compelled to think of many other problems, not directly discussed there.

The ways of thought followed in these matters may be of different kinds according to the psychology of the reader. The psychology of some readers leads them to develop a theory already started along the way indicated. They probably think: "What an interesting problem is raised! how could I develop it further?" This is certainly the origin of the considerable literature surrounding the writings of Professor Fisher. I personally seem to have another kind of psychology and can't help thinking: "What an interesting way of asking and answering questions, but can't I do differently?"

The present paper of R. A. Fisher is not an exception from the general rule. Professor Fisher states that in many cases the conception of mathematical probability is not adequate to express the nature and the extent of the uncertainty of statements and that "more generally the mathematical likelihood appears to take its place as a measure of rational belief when we are reasoning from the sample population."

Now I am tempted to raise two questions: (1) Is it really true to say that the conception of mathematical likelihood in itself contains some elements which form a new basis for our reasoning from the sample to population and which are independent of the theory of probability? To explain this question I shall illustrate it by an example. The conception of the standard deviation is certainly an element of our reasoning in many statistical problems, but it is not any sort of new basis, as its importance consists in its properties discussed in the theory of probability. This is merely a very important conception *within* the theory of probability itself.

Now what is the conception of likelihood? Is it a conception of the theory of probability like geodesic line is one of differential geometry or is it something independent?

(2) My other question is: granted that the conception of likelihood is independent of the classical theory of probability, isn't it possible to construct a theory of mathematical statistics which would be based solely upon the theory of probability (thus independent of the conception of likelihood) and be adequate from the point of view of practical statistical work?

As far as I can see, the present rôle of the likelihood in mathematical statistics is similar to that of the geodesic lines in geometry: it is an interesting and important conception, the importance of which is based upon the properties demonstrable by means of the theory of probability. The review of these properties is to be found in the present paper of Professor Fisher.

(*a*) It is shown here that "if $T$ be an estimate of an unknown parameter $\beta$, normally distributed with variance $V$, then $\frac{1}{nV}$ cannot exceed a value, $i$, defined independently of methods of estimation."

(*b*) "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 $i$."

I may add a third point, of which I do not know whether it is already known.

(*c*) If a sufficient estimate exists, then it is a maximum likelihood estimate.

(In a recent paper in the *Proceedings of the Royal Society*, Professor Fisher gives what could be considered as a sufficient condition for the existence of a sufficient statistic. This condition is, in fact, also the necessary one.)

The above three points prove by themselves that the conception of the likelihood function is extremely important. However, I do not think that we have left the ground of the theory of probability. It may be that if this point is realized, the criticism raised by some others against anything connected with the idea of likelihood will stop.

In fact we are calculating the maximum likelihood estimates not because we believe blindly in some magic properties of this function, but because there are mathematical proofs of the important properties, easy to explain in terms of other conceptions of the theory of probability, such as the variance, etc.

Now I turn to my second question—whether it is possible to construct a theory of mathematical statistics independent of the conception of likelihood. In order to discuss this we must consider carefully the rôle of likelihood in the theory of statistics built up by R. A. Fisher. This may be shortly presented as follows :—When dealing with the majority of the problems we need statistics with minimum variances or, what comes to the same thing, with maximum possible amount of information. These statistics may be obtained by the method of maximum likelihood. Besides, we may use ancillary information, also obtainable from the function of likelihood. The ancillary statistics are used to recover the information lost, and thus to diminish (if possible) the variance.

This is—roughly and shortly presented—the skeleton of Fisher's theory of mathematical statistics, at least a part of the skeleton. It contains the principle of choice among the estimates, that is, the interesting and important conception of the amount of information. The likelihood function seems to play a secondary rôle as a source of estimates with the maximum amount of information.

Now, a system of the theory of statistics, if it is to be built "differently," must differ from that of Professor Fisher in something fundamental, that is to say, in the principle of choice. I personally feel that the amount of information, if the actual meaning of the term is explained in detail, is a conception too complicated and too remote to serve as a principle. Of course if someone says, "Don't do this because you will lose some information," one will be inclined to behave in accordance with the advice. But such action may be due to the suggestive power of the words "the amount of information lost." If these words are, as it were, translated into ordinary though more complicated language and if it is realized that what is true in the limit may not be so when we are dealing with finite samples, where the distributions are often skew, then the readiness to follow the principle which may be termed that of "maximum amount of information" may be diminished.

Now what could be considered as a sufficiently simple and unquestionable principle in statistical work? I think the basic conception here is the conception of frequency of errors in judgment. Statistical problems may be roughly divided into two categories: the problems of estimation and the problems of testing hypotheses. The question, what is the density of bacteria in a given liquid, forms a problem of estimation. The similar but not identical question, whether this density exceeds a specified level, is an instance of the problem of testing hypothesis. The difference between these two problems becomes obvious when we realize that there are two possible different kinds of errors in testing hypothesis and only one kind of error in estimation. Whatever the answers to these questions, they may be true and they may be false. Any attempts to answer will be associated with a wish to avoid being wrong. Therefore if some two methods are given to select from, and if it is possible to prove that the chance or the mathematical probability of an erroneous judgment of method A is smaller than that by method B, then I do not think anybody will use method B instead of A unless the latter is too difficult to apply. Of course the variation in human psychology is enormous, but the above principle is so simple and so persuasive that I cannot imagine the psychology which will not be ready to adopt it. Still, I grant it is only a principle. If the.principle is accepted, then we have to deal with mathematical problems of finding criterion for testing hypothesis and rules for estimating population parameters such as to minimize not the variance but the actual probability of an error in judgment. The complex of results in this direction may be considered as a system of mathematical statistics alternative to that of Professor Fisher, and entirely based on the classical theory of probability.

As a matter of fact, the results which have been already reached suggest that the two systems of mathematical statistics, very different at their basis, are very close in their final results. The resulting methods of testing hypothesis and of estimation differ only in the ultimate details. Quite lately I found in dealing with one problem that the conception of the amount of information was closely related to the width of the confidence or fiducial intervals. The problem was that of estimating the density of bacteria in a liquid by means of experiments similar to those discussed by R. A. Fisher in his paper in *Philosophical Transactions*. Unfortunately the connection could be made only numerically.

Before concluding I should like to compliment Professor Fisher on the remarkable device of making continuous a discontinuous variate. At a recent meeting of this Society Professor Fisher raised the question whether the probability statements in the form of inequalities in the fiducial argument concerning a discontinuous variate were necessary or whether they were due to the use of an unsatisfactory method in my solution of the problem. Later, I was able to solve the question and show that inequalities can be avoided only in quite exceptional cases. Before this result was published, however, Professor Fisher succeeded in altering the problem in such an ingenious way that the necessity of dealing with the troublesome discontinuous variables seems to be abolished.

Professor Fisher replied in writing as follows :—

The acerbity, to use no stronger term, with which the customary vote of thanks has been moved and seconded, strange as it must seem to visitors not familiar with our Society, does not, I confess, surprise me. From the fact that thirteen years have elapsed between the publication, by the Royal Society, of my first rough outline of the developments, which are the subjects of to-day's discussion, and the occurrence of that discussion itself, it is a fair inference that some at least of the Society's authorities on matters theoretical viewed these developments with disfavour, and admitted them with reluctance. The choice of order in speaking, which puzzles Professor Bowley, seems to me admirably suited to give a cumulative impression of diminishing animosity, an impression which I should be glad to see extrapolated.

In his fourth paragraph Professor Bowley provides a medley of remarkably disconnected quotations, and of this I need only say that he is mistaken in thinking that Dr. Neyman's paper was based on the use of likelihood, or discussed the same topics as that which he had just heard. However true it may be that Professor Bowley is left very much where he was, the quotations show at least that Dr. Neyman and myself have not been left in his company.

Professor Bowley's allusion to Edgeworth recalls an intricate chapter in the history of probable errors, to which, in a short space, it is difficult to do justice. It must suffice that in 1898 Filon and Pearson put forward a general method of obtaining the sampling variance of a statistic, which leads to the formula to which Professor Bowley refers. We now know that the formula is correct (in the theory of large samples) for efficient estimates, such as the first and second moments in the case of the normal distribution. Pearson and Filon, however, used it for other estimates, derived by the method of moments, for which the formula is invalid. In 1903 the correct formulæ for the sampling variances of the inefficient statistics found by the method of moments were given in *Biometrika*, using a method due to Sheppard. These facts were presumably known to Edgeworth, writing in 1908 and 1909. He refers to Pearson and Filon's paper, though without calling attention to any error. It was also, I presume, his deep entanglement with the theory of inverse probability which prevented him from perceiving that, in this large-sample result, when properly understood, lay the key to the problem of finite samples.

For the rest, I find that Professor Bowley is offended with me for "introducing misleading ideas." He does not, however, find it necessary to demonstrate that any such idea is, in fact, misleading. It must be inferred that my real crime, in the eyes of his academic eminence, must be that of "introducing ideas."

With respect to Dr. Isserlis's remarks I have only to clear up some few confusions which have survived in the final proof circulated to me for reply. He mentions "the dependence of the idea of likelihood on a certain narrow field, so that it can only be applied in the case of investigations in which the parameter that one seeks is distributed in samples in some simple way." If Dr. Isserlis will look into the matter he will find that it can be applied whenever the parameter that one seeks is well defined, quite independently of the simplicity of the sampling distribution. I should add, of course, that the parameters should be disinguished from the statistical estimates of them, which may properly be said to be distributed in samples. It is possible that here Dr. Isserlis is reflecting a misapprehension which shows itself in Professor Bowley's remarks, where he takes the statistics referred to in my paper as being only averages, and contrasts them with "the fluctuation." The theorem given in my paper refers to statistical estimates of all kinds, and only requires that we should know what they are estimates of.

Dr. Isserlis says, " A mathematical statement is surely rigorous when it is a probability statement, just as when it is a statement in ordinary analysis." This is an echo of my first preliminary point, and in other places I find Dr. Isserlis using phrases from my writings as though he were expostulating with me. He seems to follow Neyman and Pearson in deducing an inequality statement of fiducial probability, but does not indicate any source prior to their work in which their form of argument can be found. He says that he is in full agreement with my treatment of the fourfold table, but that it does not introduce any new concept. I shall await with interest the results of a search, if he is willing to make one, for a prior use of this method.

In reply to Dr. Irwin I should like to say that when I read the valuable summaries of recent work on Mathematical Statistics which he compiles for the Society from year to year, I am quite sure that nothing in my paper would have offered any difficulty to him, even if he had not been one of those who for years had been familiar with the fundamental processes and ideas discussed.

With respect to the fourfold table, the margins provide ancillary information relevant to the estimation of the unknown parameter $\psi$, and one point which the example illustrates is that they may be used in the same way, even when no estimation is in question, but merely a test of the significance of deviation from any hypothetical value of $\psi$, such as unity. I rather wanted to show that ancillary information was useful not only in questions of estimation proper.

With respect to the functional relationship in the dilution problem, perhaps the simplest approach is to consider that, knowing the configuration, we know how many samples are fertile and how many sterile at the dilution $r$ steps after $x$, for all values of $r$. At this dilution if $m_r$ stand for the average number of organisms per sample we have,

$$\log m_r = \log \rho - x \log a - r \log a.$$

From the values of $m$ are inferred in succession :—

(i) the probabilities of a sample being sterile or fertile;

(ii) the probabilities of observing given numbers of fertile and sterile samples,

(*a*) at a given dilution, and

(*b*) at all dilutions.

Through these processes $\log \rho$ is followed by $x \log a$ as faithfully as Mary by her little lamb. The experimental variation of $\theta$ serves only to make the continuum of values of $x$ compatible with any given $\rho$, and consequent upon it with frequencies, in ranges $dx$, proportional to the probabilities calculated as above.

In reply to Professor Wolf I should probably have explained that, following Bayes, and, I believe, most of the early writers, but unlike Laplace, and others influenced by him in the nineteenth century, I mean by mathematical probability only that objective quality of the individual which corresponds to frequency in the population, of which the individual is spoken of as a typical member.

It is of great interest that Professor Wolf had concluded long ago that the concept of probability was inadequate as a basis for inductive reasoning. I believe we may add that, in so far as an induction can be cogent, it must be capable of rigorous mathematical justification, and that the concept of mathematical likelihood makes this possible in the important logical situation presented by problems of estimation.

I did not suggest that mathematics could be entirely deductive, but that the current training of pure mathematicians gave them no experience of the rigorous handling of inductive processes. Professor Wolf expresses my thought well when he says "there is induction in mathematics, but it is slurred over," but I should myself prefer to say "in mathematical applications," for some mathematical reasoning is purely deductive.

With Professor Wolf's third point I am inclined to disagree. He says: "As soon as it is realized that even mathematics was partly inductive, one could see for oneself that mathematics, or any part of it, could not be made the logical basis for all other forms of induction." This suggests that mathematics can be made the logical basis of deductive reasoning, but I doubt if this is what Professor Wolf means. I should rather say that all reasoning may properly be called mathematical, in so far as it is concise, cogent, and of general application. In this view mathematics is always no more than a means of efficient reasoning, and never attempts to provide its logical basis.

I am rather sorry that Dr. Pearson still wishes to regard problems in fiducial probability "as a new form of the method of estimation," for they constitute only a small branch of the subject. It is not clear what he means by "the method of estimation"; in what I have called the "Theory of Estimation" the categories which he distinguishes do not arise. What he calls a "single estimate," and, elsewhere, a "single valued estimate" must mean an estimate unaccompanied by any appropriate measure of precision. For maximum likelihood estimation the large-sample variance is always known. In the two cases for which the theory of estimation has been completed, by the utilization of the whole of the information supplied, even by small samples, I have shown that the exact distribution of the estimate may be derived from the likelihood function.

Dr. Pearson's paradox on the choice of fiducial intervals is an entertaining one, but is far from confirming his statement that "there are certain very important statistical problems the approach to which can be made by a simpler and more direct route." On the contrary, the difficulty, such as it is, arises solely from Dr. Pearson's own route of approach. If he wants a fiducial interval which shall reject the hypothesis to be tested more frequently when any chosen alternative hypothesis is true, he must choose an interval in the distribution of logs bounded by equal ordinates. If he wants one that rejects the true value equally frequently as too large or too small, he must use equal areas. I must add that I cannot understand the statement that I "suggested in an early paragraph of the paper that he is dealing with the whole field of logical situations at present explored." On the contrary, I limited my discussion to the description and illustration of the ideas which have arisen in the theory of estimation and, even within this field, set aside its application to the calculation of fiducial probabilities, since Dr. Neyman's paper had been largely devoted to this topic.

In reply to the President I should like to say that I am glad that he chose to allude to the discussion of the fourfold table. That example, however, does not seem well chosen to illustrate the view that my paper is only within the reach of "a tiny class of persons." The problem is one in which a large part of the audience must be in the habit of instructing their students or assistants, and with which, indeed, many junior clerks in the service of Government must be expected to deal competently. I could not have claimed the time needed to describe the very many methods which have been proposed for discussing the fourfold table; or to enumerate the objections which may be raised to each. With respect to the method of the "ordinary unmathematical student," with which the President confronts me, I suppose the simplest objection is that the probabilities assigned to the cases in which the hypothesis of independence is rejected, and those assigned to the cases in which it is accepted, do not add to unity, and therefore cannot be the probabilities of occurrence of an exhaustive enumeration of mutually exclusive events. An objection of a different kind is that the hypothesis proposed for testing is encumbered with the clause "and that the constitution of the bag assumed *was* the constitution most favourable to the assumption that they did come from the same bag." Since this clause itself implies an occurrence which is known to be improbable, apart from any question of the independence of the variates, its inclusion in a test of this independence is open to objection. To state these objections is, of course, different from detecting the logical error in the argument on which the method is supposed to be justified; but to do this it would be necessary for that argument to be set out explicitly.

I ought to be surprised that Dr. Harold Jeffreys should quarrel with the remark that, in the theory of inverse probability, the method was to introduce a postulate concerning the population from which the unknown population was supposed to be drawn. For the procedure of Bayes is quite explicit. He demonstrates a theorem in which the datum is that the probability specifying the unknown population is distributed in a given manner; and in a *scholium* following his proposition proposes the postulate in question.

The data of Bayes' propositions 8 and 9 are set out in the following paragraphs :—

> "I suppose the square table or plane *ABCD* to be so made and levelled, that if either of the balls *O* or *W* be thrown upon it, there shall be the same probability that it rests upon any one equal part of the plane as another, and that it must necessarily rest somewhere upon it.
>
> "I suppose that the ball *W* shall be first thrown, and

> through the point where it rests a line *os* shall be drawn parallel to *AD*, and meeting *CD* and *AB* in *s* and *o*; and that afterwards the ball *O* shall be thrown $p + q$ or $n$ times, and that its resting between *AD* and *os* after a single throw be called the happening of the event *M* in a single trial."

The casts with the ball *O* constitute a sample of events drawn from a population characterized by a certain frequency of the "happening of the event," which is later taken to be unknown. This population is itself explicitly obtained by the previous cast of the ball *W*.

Dr. Jeffreys further adduces the well-known fact that errors in the knowledge postulated *a priori* will, in a large class of cases, produce less and less effect on the conclusions drawn, as the observations are made more and more abundant. This seems to be a reason, not for thinking that the postulate is true, but rather that it must be possible to draw valid conclusions without its aid. This is the basis of the methods I have put forward.

Dr. Jeffreys cannot see why I have thought it necessary to demonstrate that certain statistics are sufficient, and that no others can, in these cases, supply additional information. The answer is that the theory of estimation is not confined to the normal curve of error, but is applicable to all cases in which a hypothetical population can be specified. In many cases, as I mentioned in my paper, the estimate derived from maximum likelihood does not contain the whole of the information available.

Finally, I must doubt whether any living statistician agrees with Dr. Jeffreys that the prior probability that two unknown quantities are equal is the same as that they are unequal. If this were, indeed, a property of prior probabilities this fact would, in my opinion, alone suffice to justify their exclusion from any argument having practical aims.

In reply to Mr. Bartlett, he is perhaps wrestling with a difficulty the force of which I have not myself felt. I agree that the use of the word "random" implies the notion of equal objective probabilities, or of equal hypothetical frequencies. The notion of random sampling also implies a hypothetical population characterized by these frequencies. For frequencies, the laws of probability are directly demonstrable, after the manner of Bayes, in the earlier sections of his essay. My own definition is not based on the limit of frequencies, if by this Mr. Bartlett means experimental frequencies, for I believe we have no knowledge of the existence of such limits.

With reference to Dr. Neyman's interesting contribution, I must confine myself to some very brief notes.

(i) In the present paper I have been particularly concerned to show that all the properties of mathematical likelihood, which make it valuable, can be demonstrated independently of any postulated value. From this it seems to me to follow that the concept of likelihood could be eliminated completely from discussions of estimation, and these discussions be adequately, though perhaps more cumbrously, carried out in other terms. A like argument could, of course, be used for the elimination of the notion of temperature from physical and thermodynamic discussions. The fact that likelihood has been an aid to thought in such progress as has so far been made in the subject will suggest the advisability of using it for what it is worth, even though, ultimately, we may find ourselves able to do better. That there are logical situations in which the uncertainty of our inferences is expressible in terms of likelihood, but not in terms of probability, is one solid step gained, even though more comprehensive notions may later be developed.

(ii) I ought to mention that the theorem that if a sufficient statistic exists, then it is given by the method of maximum likelihood was proved in my paper of 1921, to which Dr. Neyman refers. It was this that led me to attach especial importance to this method. I did not at that time, however, appreciate the cases in which there is no sufficient statistic, or realize that other properties of the likelihood function, in addition to the position of the maximum, could supply what was lacking.

(iii) In saying " We need statistics with minimum variances, or, what comes to the same thing, with maximum possible amount of information," Dr. Neyman must be taken as speaking only of the preliminary part of the theory, dealing with the properties of statistics in " large " samples. The concept of *amount of information* as a measurable quantity, not identical with the invariance, was developed for the theory of finite samples, where the distributions are often skew, and it is only in studying these that the advantage of assessing and utilizing the whole of the information available will be fully appreciated.

It has been, naturally, of great interest to me to follow the attempts which Drs. Neyman and Pearson have made to develop a theory of estimation independently of some of the concepts I have used. That, whenever unequivocal results have been obtained by both methods they have been identical is, of course, a gratifying confirmation of the hope that we are working along sound lines.

As a result of the ballot taken during the meeting, the candidates named below were elected Fellows of the Society :—

Eric Lester Bunce.
Thomas Douglas Carnwath, B.A.
William Harry Coombs.
David Dolovitz.
Moritz John Elsas.
Ishwar Das Mahendru.
Jakob Marschak.
Ernest Stanley Tucker.
E. Ashworth Underwood, M.A., M.B., D.P.H.
Matthew Young, M.D., D.Sc., D.P.H.