STATISTICAL THEORY OF ESTIMATION

Calcutta University Readership Lectures

STATISTICAL THEORY OF ESTIMATION

BY

R. A. FISHER, Sc.D., F.R.S.
GALTON PROFESSOR, UNIVERSITY OF LONDON

PUBLISHED BY THE
UNIVERSITY OF CALCUTTA
1938

PRINTED IN INDIA

PRINTED AND PUBLISHED BY BHUPENDRALAL BANERJEE
AT THE CALCUTTA UNIVERSITY PRESS, SENATE HOUSE, CALCUTTA

Reg. No. 1116B—September, 1928—E.

PREFACE

It is both a pleasure and an honour to address these lectures to the Indian mathematicians gathered in Calcutta for the Indian Science Congress and the first Indian Statistical Conference ; and especially to make contact with the Statistical Laboratory established in this city by the initiative of Professor P. C. Mahalanobis.

The course is based upon numerous research papers written during the period over which the Theory of Estimation was making its most striking progress. To give a proper perspective I have aimed at presenting each important step from more than one point of view. Some repetition has, therefore, been intentional in order to make the course self-contained. For advanced students the lectures should be supplemented by a further study of the papers referred to.

For the orderly presentation of the material in book form, I am indebted to Professor Mahalanobis and his colleagues at the Statistical Laboratory.

January, 1938. R. A. FISHER.

CONTENTS

1. The Logical Situations in which Problems of Estimation Arise

In order to arrive at a distinct formulation of statistical problems, it is necessary to define the task which the statistician sets himself: briefly, and in its most concrete form, the object of statistical methods is the reduction of data. A quantity of data, which usually by its mere bulk is incapable of entering the mind, is to be replaced by relatively few quantities which shall adequately represent the whole, or which, in other words, shall contain as much as possible, ideally the whole, of the relevant information contained in the original data.

This object is accomplished by constructing a hypothetical infinite population, of which the actual data are regarded as constituting a random sample. The selection of such a hypothetical population may be called the problem of specification. Thus specification is the hypothesis in the light of which we interpret our data. The law of distribution of this hypothetical population is specified by relatively few parameters, which are sufficient to describe it exhaustively in respect of all qualities under discussion. Any information given by the sample, which is of use in estimating the values of these parameters, is relevant information. Since the number of independent facts supplied in the data is usually far greater than the number of facts sought, much of the information supplied by any actual sample is irrelevant. It is the object of the statistical processes employed in the reduction of data to exclude this irrelevant information, and to isolate the whole of the relevant information contained in the data.

Consider, for example, a sample in which we have got values of the yield y corresponding to different doses of manure x. Prof. Mitcheslich and his students have taken the formula

$$Y = A[1-e^{-k(x-b)}] \qquad \ldots \quad (1)$$

to denote the relation between the manure x and the yield Y. It should however be noted that this gives us an ideal or undisturbed relation between x and Y. The observed yields y will be distributed about Y, so that Y is only the expectation for a constant value of x. We may reasonably take y to be distributed normally about Y, so that y may be taken to be distributed according to the following law :—

$$df = \frac{1}{\sqrt{(2\pi v)}} e^{-\frac{1}{2v}(y-Y)^2} dy \qquad \ldots \quad (2)$$

where Y is given by (1).

This completes our specification. From the sample the statistician will then want to estimate the values of the unknown parameters A, k, b and v.

The choice of specification is fundamental to the statistician. There are all degrees of empiricism about that choice as the above example would have made clear. Sometimes, however, established theory gives us the specification. For example consider

AB, A*b*, *a*B, *ab*

the four phenotypes in which a progeny of self-fertilised heterozygotes can be distributed with regard to two recognizable contrasts, *e.g.*, the maize factors, Starchy *v.* Sugary and Green *v.* White.

The frequencies with which these four phenotypes are expected to occur are proportional to

$$\frac{2+p^2}{4}, \quad \frac{1-p^2}{4}, \quad \frac{1-p^2}{4}, \quad \frac{p^2}{4}$$

where p is the recombination fraction regarded as equal in both sexes.

Here also errors of random sampling come in, but the specification is complete with p^2 only for the expected frequencies in all distinguishable classes are expressible in terms of p^2.

The above is an example of a discontinuous frequency distribution, but statisticians frequently come across continuous distributions, the most familiar being the Normal or Gaussian Law of Error. This is given by

$$df = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-m)^2}{2v}} dx$$

Skew variations may also be considered, for example, we may consider the frequency distribution

$$df = \frac{1}{p\,!}\left(\frac{x-\mu}{a}\right)^{p} . \, e^{-\frac{x-\mu}{a}} . \frac{dx}{a}$$

The specifications we shall consider can thus involve one or more unknown parameters. For example the one given above involves the three unknown parameters p, μ and a. From the data we naturally want to make estimates of these parameters. Whether these estimates are good or bad will be our problem. But for this we want to know also how these estimates are distributed in random samples.

Following specification, we thus come to the problem of distribution, especially the distribution of the statistics which we put forward as estimating our parameters. Such problems therefore have a great importance for us. For it is in their light that we can judge of the merits of our estimates.

2. Methods of Solution of Problems of Distribution

We shall consider in broad outline the methods available for solving problems of distribution. There are three general methods :

(*i*) Method of Euclidean Hyperspace.

(*ii*) Method of Mathematical Induction.

(*iii*) Method using Characteristic Functions.

(*i*) For example, suppose we have n quantities x_1, x_2, ... x_n all distributed independently and normally about zero with unit variance. Consider

$$\chi^2 = S(x^2)$$

and let us try to find the distribution of χ^2.

We can consider $x_1, x_2, \ldots, x_n$ as the co-ordinates of a point P in a space of n dimensions. If O is the origin of co-ordinates then

$$OP^2 = \chi^2$$

Therefore χ^2 is constant over concentric hyperspheres. The hypervolume included between two such hyperspheres is proportional to

$$\chi^{n-1}\, d\chi$$

The joint distribution of $x_1, x_2, \ldots, x_n$ is given by

$$df = \left(\frac{1}{\sqrt{2\pi}}\right)^n . \, e^{-\frac{1}{2}S(x^2)}\, dx_1 . \, dx_2 . \ldots . \, dx_n$$

so that the distribution of χ is

$$\text{Const. } e^{-\frac{1}{2}\chi^2} \chi^{n-1} d\chi$$

The value of the constant can easily be shown to be

$$\frac{1}{2^{(n-2)/2}\left(\frac{n-2}{2}\right)!}$$

using what is known as the Eulerian integral of the second kind, and remembering that the total frequency is unity.

This problem was actually solved by the method of characteristic functions by Helmert in 1875. It emerged in a different form when Pearson put forward his test of goodness of fit, and was rediscovered by " Student " in connection with other problems.

(*ii*) The same problem can be tackled by the method of induction. If $x_1, x_2, \dots, x_n$ are independent values of a variate distributed normally about zero, with unit variance, then the quantity

$$\chi^2 = x_1{}^2 + x_2{}^2 + \dots + x_n{}^2$$

has a distribution given by :—

$$df = \frac{1}{\left(\frac{n-2}{2}\right)!} (\tfrac{1}{2}\chi^2)^{\frac{1}{2}(n-2)} \, . \, e^{-\frac{1}{2}\chi^2} \, d(\tfrac{1}{2}\chi^2)$$

To prove this by induction, for $n=1$ the expression reduces to

$$\sqrt{\frac{2}{\pi}} \, . \, e^{-\frac{1}{2}\chi^2} \, d\chi$$

which is clearly the distribution of χ^2 for a single observation. If, now, $2u$ is the sum of squares of n independent values of the variate, and has the distribution

$$df = \frac{1}{\left(\frac{n-2}{2}\right)!} u^{\frac{1}{2}(n-2)} e^{-u} \, du$$

and x is an additional observation independent of the others, then

$$\chi^2 = 2u + x^2,$$

and its distribution is to be inferred from the simultaneous distribution

$$df = \frac{1}{\left(\frac{n-2}{2}\right)!} \sqrt{\frac{2}{\pi}}\, u^{\frac{1}{2}(n-2)}\, e^{-u-\frac{1}{2}x^2}\, du\, dx$$

If we now substitute

$$u = \tfrac{1}{2}(\chi^2 - x^2), du = d(\tfrac{1}{2}\chi^2)$$

we have

$$df = \frac{1}{\left(\frac{n-2}{2}\right)!} \sqrt{\frac{2}{\pi}}\, e^{-\frac{1}{2}\chi^2} d(\tfrac{1}{2}\chi^2) \left(\frac{\chi^2 - x^2}{2}\right)^{\frac{1}{2}(n-2)} dx$$

in which x takes all values from 0 to χ. Integration with respect to x will therefore yield a factor χ^{n-1} giving the distribution

$$df = \frac{1}{\left(\frac{n-1}{2}\right)!} (\tfrac{1}{2}\chi^2)^{\frac{1}{2}(n-1)}\, e^{-\frac{1}{2}\chi^2}\, d(\tfrac{1}{2}\chi^2)$$

in accordance with the general formula.

(*iii*) Consider any frequency distribution

$$df = f(x)dx$$

Then

$$\int e^{itx} f(x)dx$$

always exists and always lies within the unit circle.

If now we put

$$M_x(t)=\int e^{itx}\ f(x)dx$$

then $M_x(t)$ is the characteristic function of the distribution

$$df = f(x)dx.$$

Similar definitions can be given for the multivariate case. Consider now a bivariate (or multivariate) population with unknown parameters θ, ϕ, ψ,

$$df = f(x,\ y,\ \theta,\ \phi,\ \psi)\ dxdy$$

Consider any statistic T

$$T = F\begin{pmatrix} x_1,\ x_2,\ \dots, & \\ y_1,\ y_2,\ \dots, & y_n \end{pmatrix}$$

based on a sample of size n.

Then the characteristic function for the distribution of T would be

$$\int\cdots\int\int e^{itT}\ f(x_1,\ y_1,\ \theta,\ \phi,\ \psi).\ f(x_2,\ y_2,\ \theta,\ \phi,\ \psi)\dots f(x_n,\ y_n,\ \theta,\ \phi,\ \psi$$

$$\times\ dx_1dy_1dx_2dy_2\dots dx_ndy_n.$$

From this we could at least theoretically write down the distribution of T. Hence the method of characteristic functions is potentially a very powerful method.

If x and y are two independently distributed variates then the characteristic function of the distribution of their sum is given by

$$\mathrm{M}_{x+y}(t) = \mathrm{M}_x(t).\ \mathrm{M}_y(t)$$

If we put

$$\log \mathrm{M}_x(t) = \mathrm{K}_x(t)$$

then $\mathrm{K}_x\ (t)$ may be called the cumulative function of the distribution of x. Then

$$\mathrm{K}_{x+y}(t) = \mathrm{K}_x(t) + \mathrm{K}_y(t)$$

If

$$\mathrm{S}(x) = x_1 + x_2 + \ldots + x_n$$

then

$$\mathrm{K}_{S(x)}(t) = n\mathrm{K}_x(t)$$

or again

$$\mathrm{K}_{\frac{1}{n}S(x)}(t) = n\mathrm{K}_x(t/n)$$

giving us the characteristic function for the distribution of the mean of n quantities x_1, x_2, ..., x_n.

3. The Limiting Values of Statistics

If we have decided on a method of estimation and if we apply it to larger and larger samples, then, for our method to be of any value, our estimates must show increasing agreement with one another. More rigorously, our estimating statistic T must tend to a limit in the following sense :—

If T_n be any statistic calculated from a sample of n observations, there must be a limiting value T_∞ such that if ϵ be any positive number, however small, the frequency (or probability) with which $| T_n - T_\infty |$ exceeds ϵ, tends to zero as n tends to infinity.

In symbols

$$P\{ | T_n - T_\infty | > \epsilon \} \longrightarrow 0$$

In the large majority of cases this quality of tending to a limit is possessed. The criterion of consistency then simply states that this limit T_∞ is the same as the parameter θ estimated by T. The criterion of consistency is defined only in the case when a limit exists. The question is left open when the limit does not exist. A good example of a statistic not tending to a limit is given by Cauchy's distribution,

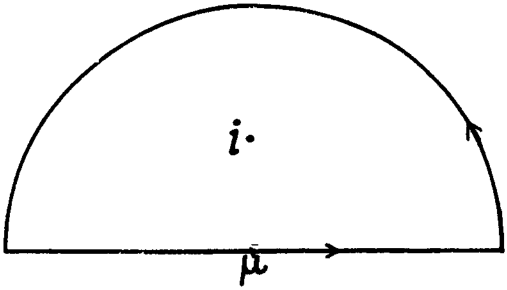
$$df = \frac{1}{\pi} \cdot \frac{dx}{1+(x-\mu)^2}\,.$$

E(x) does not exist as

$$\frac{xdx}{1+(x-\mu)^2}$$

cannot be integrated, since the integral does not converge.

Suppose a radio-active source is placed at a unit distance in front of a screen so that the foot of the perpendicular from the source on the screen is at a distance μ from the origin.



If then x is the distance from the origin of any point where an α-particle emerging from the radio-active source strikes the screen, then x is distributed in the Cauchy distribution.

For if θ is the angle that the direction of motion of the particle makes with the perpendicular then

$$df = \frac{1}{\pi}\, d\theta$$

$$x-\mu = \tan\theta$$

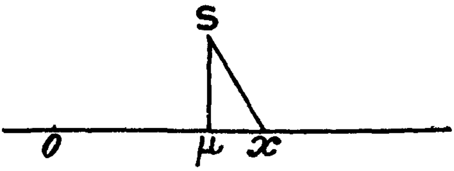
$$dx = \sec^2\theta d\theta$$

Therefore,
$$df = \frac{dx}{\pi(1+\tan^2\theta)}$$
$$= \frac{1}{\pi} \cdot \frac{dx}{1+(x-\mu)^2}$$

If we take n observations of this kind and if we take an average of a sample of n, the average is distributed in the same way as a single observation. To prove this we evaluate

$$E\left(e^{it(x-\mu)}\right)$$

as a contour integral.



Writing $x-\mu=z$ and taking t positive

$$M_x(t) = E(e^{it(x-\mu)}) = \frac{1}{\pi}\int \frac{1}{1+z^2} e^{itz} dz$$

Equating this to the contour integral around the pole, $z=i+\epsilon$, gives

$$\frac{1}{\pi}\int \frac{d\epsilon}{2i\epsilon+\epsilon^2} e^{-t} = e^{-t}$$

Similarly if t is negative, $M_x(t)=e^t$, so that in general we can write $M_x(t) = e^{-|t|}$

or, $$K_x(t) = -|t|$$

Therefore

$$K_{\frac{1}{n}S(x)}(t) = nK_x\left(\frac{t}{n}\right) = -n\left|\frac{t}{n}\right| = -|t|$$

Hence the characteristic function for the distribution of the average of n readings is exactly the same as the characteristic function for the original distribution. This proves that the average is distributed in the same way as a single observation.

If now one should want on the basis of our observations to locate the radio-active source, or in other words to find the value of μ, then taking the mean of a 100 readings, say, would amount to throwing away 99 readings and retaining only a single one of them chosen at random. In this way onē loses therefore 99 per cent. of the information.

The proof of the fact that the distribution of the mean is the same as that of a single reading can also be given by direct analysis.

Let x and y be two quantities independently distributed in Cauchy's distribution so that their joint distribution is

$$df = \frac{dy}{\pi\{1+(y-\mu)^2\}} \cdot \frac{dx}{\pi\{1+(x-\mu)^2\}}$$

If we put

$$u = \frac{x+y}{2}$$

$$v = \frac{x-y}{2}$$

then, transforming the variables x, y to u, v we get

$$df = \frac{2dudv}{\pi^2\{1+(u-v-\mu)^2\}\ \{1+(u+v-\mu)^2\}}$$

Integrating out for v from $-\infty$ to $+\infty$ we get the distribution of u again in the Cauchy form

$$df = \frac{du}{\pi\{1+(u-\mu)^2\}}$$

This shows that a sample of 2, 4, 8, ... , 2^k readings is distributed in the same way as x. To complete the proof we may now show by the same method that if

$$u = px + qy$$

$$|p| + |q| = 1$$

then u is again distributed in the same way.

Here the first moment of the distribution is a very unsatisfactory statistic. But we can demonstrate that the data are not worthless for the point for which we want to use them. It is only our method of estimation that is wrong.

4. The Distribution of the Median

Let us consider how the median is distributed. Let us suppose there are $n = 2s+1$ observations from a population distributed according to the law

$$df = ydx$$

Let $$\phi(x) = \int_{med.}^{x} ydx$$

Then, the distribution of the median is

$$\frac{(2s+1)!}{s!s!}\{\tfrac{1}{2}-\phi(x)\}^s\{\tfrac{1}{2}+\phi(x)\}^s ydx$$

or

$$\text{const. }\{1-4\phi^2(x)\}^s ydx$$

If now $y=y_0$ at the population median then for large values of s the distribution of the median reduces to the normal limiting form

$$\text{const. } e^{-4sy_0{}^2x^2}$$

where x is the deviation of the sample median.

The variance of the median is equal to

$$\frac{1}{8sy_0{}^2}$$

which is approximately equal to

$$\frac{1}{4ny_0^2}$$

In the case of the Cauchy distribution under consideration

$$y_0 = \frac{1}{\pi}$$

Therefore, the variance of the median

$$\text{V (median)} = \frac{\pi^2}{4n}$$

Thus the median has increasing precision as the size of the sample is increased. Hence from the median we could elicit a good deal of information from large samples. The median satisfies the criterion of consistency.

5. The Criterion of Efficiency

Among statistics which satisfy the criterion of consistency we shall now consider a particular class, *viz.*, those which satisfy the criterion of Efficiency. In a large and important class of consistent statistics the random sampling distribution tends to the normal (Gaussian) form as the size of the sample is increased, and in such a way that the variance (the square of the standard deviation) falls off inversely to the size of the sample. The criterion of efficiency requires that the fixed value to which the variance of a statistic (of the class of which we are speaking) multiplied by n tends, shall be as small as possible. An efficient statistic is one for which this criterion is satisfied. If we know the variance of any efficient statistic and that of any other statistic under discussion, then the *efficiency* of the latter may be calculated from the ratio of the two values. The efficiency of a statistic represents the fraction, of the relevant information available, actually utilised, in large samples, by the statistic in question.

Let us again go back to our normal population

$$df = \frac{1}{\sqrt{2\pi v}} e^{-\frac{(x-\mu)^2}{2v}} dx$$

and suppose we want to estimate μ.

The variance of the mean of a sample of n is

$$\frac{v}{n}$$

Hence n times the variance is equal to v. If on the other hand we use the median then the variance is

$$\frac{1}{4ny_0^2}$$

where now

$$y_0 = \frac{1}{\sqrt{2\pi v}}$$

so that the variance of the median is

$$\frac{\pi v}{2n}$$

Therefore n times the variance is $\frac{\pi v}{2}$. Hence the median is only $\frac{2}{\pi}$ times as efficient as the mean. It shall appear later that, in the case of a normal population, the mean has the minimum possible variance and is therefore efficient.

This naturally raises the question whether we can find a better statistic than the mean. A certain amount of enlightenment on this question can be obtained by noticing a special property of the mean. The joint distribution of a sample of n observations is given by

$$\left(\frac{1}{\sqrt{2\pi v}}\right)^n e^{-\frac{S(x-\mu)^2}{2v}} dx_1 . dx_2 . \dots dx_n$$

$$S(x-\mu)^2 = n(\bar{x} - \mu)^2 + S(x-\bar{x})^2$$

where $\bar{x}$ is the mean.

Here by using methods already explained the distribution of $\bar{x}$ itself comes out in the normal form

$$\frac{\sqrt{n}}{\sqrt{2\pi v}} e^{\frac{-n(\bar{x}-\mu)^2}{2v}} d\bar{x}$$

If we were to take the mean value as fixed and consider the probability that the sample has a given composition consistent with this constraint, then by dividing the frequency of the sample, by that of the mean value, it appears that given the mean, the probability of the sample having any given composition is independent of μ. Thus any other statistic besides $\bar{x}$ gives us no further information about μ once $\bar{x}$ has been calculated. If T be such a statistic then the joint distribution of $\bar{x}$ and T will reduce to the form

$$f\ (\bar{x},\ \mathrm{T},\ \mu)\ d\bar{x}\ d\mathrm{T} \equiv \phi\ (\bar{x},\ \mu)\ .\ \psi(\mathrm{T},\ \bar{x})\ d\bar{x}\ d\mathrm{T}$$

so that for a given value of $\bar{x}$, T is distributed independently of μ.

Thus we have shown that the mean satisfies a more penetrating criterion, that is, the criterion of Sufficiency.

✓ In general, if θ is any parameter, T_1 a sufficient statistic in estimating that parameter, and T_2 any other statistic, the sampling distribution of simultaneous values of T_1 and T_2 must be such that for any given value of T_1, the distribution of T_2 does not involve θ.

This will evidently be the case, if

$$f(\theta,\ \mathrm{T}_1,\ \mathrm{T}_2)\ .\ d\mathrm{T}_1\ .\ d\mathrm{T}_2$$

be the probability that T_1 and T_2 should fall in the ranges $d\mathrm{T}_1$, $d\mathrm{T}_2$, and if

$$f(\theta,\ \mathrm{T}_1,\ \mathrm{T}_2) = \phi\ (\theta,\ \mathrm{T}_1)\ .\ \phi'\ (\mathrm{T}_1,\ \mathrm{T}_2)$$

If this condition is fulfilled for all possible statistics T_2, then will T_1 be a sufficient statistic.

When a sufficient statistic exists it is equivalent, for all subsequent purposes of estimation, to the original data from which it was derived.

6. The Method of Maximal Likelihood

The form in which the criterion of sufficiency has been presented is not of direct assistance in the solution of problems of estimation. For it is necessary first to know the statistic concerned and its surface of distribution, with an infinite number of other statistics, before its sufficiency can be tested. For the solution of problems of estimation we require a method which for each particular problem will lead us automatically to a sufficient statistic if one exists, and in any case to an efficient statistic. Such a method is provided by the Method of Maximal Likelihood.

If in any distribution involving unknown parameters θ_1, θ_2, θ_3, ..., the chance of an observation falling in the range dx be represented by

$$f\,(x, \theta_1, \theta_2, \ldots) dx,$$

then the chance that in a sample of n, n_1 will fall in the range dx_1, n_2 in the range dx_2, and so on, will be

$$\frac{n\,!}{\Pi\,(n_p!)}\,\Pi\left\{f\,(x_p, \theta_1, \theta_2, \ldots)\,dx_p\right\}^{n_p}$$

The method of maximal likelihood consists simply in choosing that set of values for the parameters which makes this quantity a maximum, and since in this expression the parameters are only involved in the function f, we have to make

$$S (\log f)$$

a maximum for variations of θ_1, θ_2, θ_3, etc. In this form the method is applicable to the fitting of populations involving any number of variates, and equally to discontinuous as to continuous distributions.

I should like to make clear at this stage the distinction between this method and that of Bayes. Bayes put forward, with considerable caution, a method by which such problems could be reduced to the form of problems of probability. His method of doing this depended essentially on postulating *a priori* knowledge not of the particular population of which our observations form a sample, but of an imaginary population of populations from which this population was regarded as having been drawn at random. Clearly, if we have possession of such *a priori* knowledge, our problem is not properly an inductive one at all, for the population under discussion is then regarded merely as a particular case of a general type, of which we already possess exact knowledge, and are therefore in a position to draw exact deductive inferences.

If a sample of n independent observations each of which may be classified unambiguously in two alternative classes as "successes" and "failures" be drawn from a population containing a relative frequency x of successes, then the probability that there shall be a successes in our sample is, as was first shown by Bernoulli,

$$\frac{n!}{a!\,(n-a)!} \quad x^{a}\,(1-x)^{n-a}$$

This is an inference, drawn from the general to the particular, and expressible in terms of probability. We are given the parameter x, which characterizes the population of events of which our observations form a sample and from it can infer the probability of occurrence of samples of any particular kind.

If, however, we have *a priori* knowledge of the probability $f(x)dx$ that x should lie in any specified range dx, or if, in other words, we knew that our population had been chosen at random from the population of populations having various values of x, but in which the distribution of the variate x is specified by the frequency element $f(x)dx$ of known form, then we might argue that the probability of first drawing a population in the range dx, and then drawing from it a sample of n having a successes, must be

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

Since this sequence of events has occurred for some value of x, the expression above must be proportional to the probability, subsequent to the observation of the sample, that x lies in the range dx. The postulate which Bayes considered was that $f(x)$, the frequency density in the hypothetical population of populations, could be assumed *a priori* to be equal to unity.

As an axiom this supposition of Bayes fails, since the truth of an axiom should be manifest to all who clearly apprehend its meaning, and to many writers, including, it would seem, Bayes himself, the truth of the supposed axiom has not been apparent. It has, however, been frequently pointed out that, even if our assumed form for $f(x)dx$ be somewhat inaccurate, our conclusions, if based on a considerable sample of observations, will not greatly be affected; and, indeed, subject to certain restrictions as to the true form of $f(x)dx$ it may be shown that our errors from this cause will tend to zero as the sample of observations is increased indefinitely. The conclusions drawn will depend more and more entirely on the facts observed, and less and less upon the supposed knowledge *a priori* introduced into the argument. This property of increasingly large samples has been sometimes put forward as a reason for accepting the postulate of knowledge *a priori*. It appears, however, more natural to infer from it that it should be possible to draw valid conclusions from the data alone, and without *a priori* assumptions. If the justification for any particular form of $f(x)$ is merely that it makes no difference whether the form is right or wrong, we may well ask what the expression is doing in our reasoning at all, and whether, if it were altogether omitted, we could not without its aid draw whatever inferences may, with validity, be inferred from the data. In particular we may question whether the whole difficulty has not arisen in an attempt to express, in terms of the single concept of mathematical probability, a form of reasoning which requires for its exact statement different though equally well-defined concepts.

If, then, we disclaim knowledge *a priori*, or prefer to avoid introducing such knowledge as we possess into the basis of an exact mathematical argument, we are left only with the expression

$$\frac{n!}{a!\,(n-a)!}\;x^a(1-x)^{n-a},$$

which, when properly interpreted, must contain the whole of the information respecting x which our sample of observations has to give. This is a known function of x, for which, in 1922, I proposed the term "likelihood," in view of the fact that, with respect to x, it is not a probability, and does not obey the laws of probability, while at the same time it bears to the problem of rational choice among the possible values of x a relation similar to that which probability bears to the problem of predicting events in games of chance. From the point of view adopted in the theory of estimation, it could be shown, in fact, that the value of x, or of any other parameter, having the greatest likelihood, possessed certain unique properties, in which such an estimate is unequivocally superior to all other possible estimates.

7. THE MAXIMAL PRECISION ATTAINABLE

We shall now first show that if T be an estimate of an unknown parameter θ and if in large samples T is distributed normally with variance V, then the limit, as n tends to infinity, of $\frac{1}{nV}$ cannot exceed a value i defined independently of the methods of estimation.

Let $f=f(x, \theta)$ stand for the frequency of a particular kind of observation, ϕ for that of a particular kind of sample and Φ for that of all the kinds of sample which yield a particular value T of the statistic chosen as an estimate.

Then in general

$$\log \phi = \mathrm{S}\ (\log f)$$

where S stands for summation over the sample; next

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

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

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

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

If T is distributed normally about θ with variance V,

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

Hence

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

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

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

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

Hence

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

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

Now consider

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

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

$$\frac{1}{\Phi}\Sigma(\phi u^2)-\frac{1}{\Phi^2}\Sigma^2(\phi u)$$

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

but the variance of u is positive, or, in 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

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

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

$$n\left[\frac{1}{f}\cdot\frac{\partial f}{\partial\theta}\right]^2$$

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

$$\frac{1}{n\mathrm{V}}\leqq i$$

where i is the average value of

$$\left[\frac{1}{f}\cdot\frac{\partial f}{\partial\theta}\right]^2;$$

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

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

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 the 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

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

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

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

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

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

If the co-efficients k are functions of θ, the equation

$$X = 0$$

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

$$S(kf) = 0$$

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

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

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

$$S(k^2m) = n\,S(k^2f)$$

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

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

If, therefore,

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

tends to a finite limit,

$$-\frac{1}{S\,(f\,\frac{\partial k}{\partial \theta})}$$

the sampling variance of our estimate is

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

or, using the condition for consistency,

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

We may now apply the calculus of variations to find the functions k of θ which will minimize the sampling variance. Since the variance must be stationary for variations of each of several values 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

$$\mathrm{S}\,(kf) = 0$$

for all values of θ. The equation of estimation thus obtained,

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

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

$$n\,\mathrm{V} = \frac{1}{\mathrm{S}\{\frac{1}{f}\,.\,(\frac{\partial f}{\partial \theta})^{2}\}}$$

or

$$\frac{1}{n\mathrm{V}} = \mathrm{S}\{\frac{1}{f}\,.\,(\frac{\partial f}{\partial \theta})^{2}\} = i$$

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.

For example, take the Mendelian case considered earlier, where the frequency distribution in the four classes were taken to be

$$\frac{2+\theta}{4},\quad \frac{1-\theta}{4},\quad \frac{1-\theta}{4},\quad \frac{\theta}{4}$$

where $\theta = p^2$

Then

$$\frac{\partial f}{\partial \theta} = \tfrac{1}{4},\quad -\tfrac{1}{4},\quad -\tfrac{1}{4},\quad \tfrac{1}{4}$$

and

$$\frac{1}{f}\left(\frac{\partial f}{\partial \theta}\right)^2 = \tfrac{1}{4}\left[\frac{1}{2+\theta},\quad \frac{1}{1-\theta},\quad \frac{1}{1-\theta},\quad \frac{1}{\theta}\right]$$

Therefore

$$i = \frac{1+2\theta}{2\theta\,(1-\theta)\,(2+\theta)}$$

which is the information each seedling observed contributes relevant to the estimation of the value of θ.

If we take the Normal distribution

$$fdx = \frac{1}{\sqrt{2\pi v}} \, e^{-\frac{(x-\mu)^2}{2v}} \, dx$$

then

$$\mathrm{E}\left[n\left(\frac{1}{f}\frac{\partial f}{\partial \mu}\right)^2 \right] = \frac{n}{v} = \frac{1}{\mathrm{V}}$$

where V is the variance of the mean of the sample. This shows that the mean contains all the information about the parameter that the sample can supply.

Let us next turn to the Cauchy distribution

$$fdx = \frac{1}{\pi} \cdot \frac{dx}{1+(x-\mu^2)}$$

$$\mathrm{E}\left[n\left(\frac{1}{f}\frac{\partial f}{\partial \mu}\right)^2 \right] = n\int f\left(\frac{\partial}{\partial \mu} log\ f\right)^2 dx$$

$$= n\int_{-\pi/2}^{+\pi/2} \frac{d\theta}{\pi} \cdot \frac{4tan^2\theta}{sec^4\theta} \quad \text{where } x-\mu = \tan\theta$$

$$= \frac{n}{2}$$

i.e., amount of information is $n/2$, and the least possible variance is $\frac{2}{n}$. We have already found that the variance for the median is $\frac{\pi^2}{4n}$. Hence the efficiency of the median is $\frac{8}{\pi^2}$.

8. Frequency Space

Suppose there are s classes and the probability of an individual falling in the kth class is p_k. If n is the size of the sample then the expected frequency in the kth class is $n.p_k$. Of course p_k is a function of the parameter θ to be estimated. Any sample of n observations is distributed among the s classes, the observed frequency in the kth class being a_k. Then

$$S\,(a_k) = n$$

The following representation by means of hyperspace geometry will be useful to us. The sample in which the observed frequency in the kth class is a_k is represented by a point with co-ordinates $(a_1, a_2, \ldots a_k, \ldots a_s)$. In this way every sample is represented by some point of an s dimensional space. Since a_1, a_2, etc. are integral, only the lattice points of this space can represent samples. Corresponding to any given value of θ if

$$m_1, m_2, \ldots m_k, \ldots m_s$$

be the expectations in the different classes, then we get the Expectation point $(m_1, m_2, \ldots m_k, \ldots m_s)$. For different values of θ we have different expectation points, their locus constituting what we may call the curve of Expectation. When n is increased indefinitely the lattice reduces to a continuum about the expectation point. The frequency with which our sample is observed is

$$\frac{n\,!}{a_1\,!\,a_2\,!\ldots a_k\,!\ldots a_s\,!}\;p_1^{a_1}\cdot p_2^{a_2}\ldots p_k^{a_k}\ldots p_s^{a_s}$$

with the restriction

$$\mathrm{S}\,(a_k)=n$$

Supposing now we choose the size of our sample at random from a Poisson series with mean N, so that the chance of getting a sample of size n is

$$e^{-\mathrm{N}}\cdot\frac{\mathrm{N}^n}{n\,!}$$

Hence the frequency with which the sample $(a_1,a_2\ldots a_k,\ldots a_s)$ will be observed becomes

$$e^{-\mathrm{N}}\cdot\frac{\mathrm{N}^n}{n\,!}\cdot\frac{n\,!}{a_1\,!\,a_2\,!\ldots a_k\,!\ldots a_s\,!}\;p_1^{a_1}\cdot p_2^{a_2}\ldots p_k^{a^k}\ldots p_s^{a^s}$$

Also as N is the expected size of the sample

$$\mathrm{N}\,p_k=m_k$$

Hence the frequency becomes

$$e^{-\mathrm{S}\,(m_k)}\cdot\frac{m_1^{a_1}m_2^{a_2}\ldots m_k^{a_k}\ldots m_s^{a_s}}{a_1\,!\,a_2\,!\ldots a_k\,!\ldots a_s\,!}$$

$$=\prod_1^s e^{-m_k}\cdot\frac{m_k^{a_k}}{a_k\,!}$$

The a's are now all independently distributed in different Poisson distributions. If now m_k is increased without limit, then the Poisson distribution tends to Normal with the same mean and variance.

Then

$$x_k = \frac{a_k - m_k}{\sqrt{m_k}}$$

tends to be normally distributed with unit variance. Thus we get a Normalised isotropic distribution. The sample points (x_1, $x_2, \ldots x_s$) now form a globular cluster of normally distributed points about the expectation point, and

$$\chi^2 = S\,(x^2)$$

is simply the distance of the sample point from the expectation point.

The problem of estimation is to find from the sample point the most appropriate point on the curve of expectation. Thus every method of estimation is virtually equivalent to dividing up space into what may be called equistatistical regions such that every sample point on the same region leads to the same estimate. The criterion of consistency then simply states that the equistatistical region leading to any estimate of θ should actually cut the curve of expectation at the point corresponding to this value of θ. Efficient statistics have the peculiarity that the equistatistical region corresponding to such a statistic cuts the curve of expectation at right angles in the transformed space. The maximal likelihood solution is unique in that, in addition, its equistatistical region is linear. The equistatistical regions for minimum χ^2 are not linear and touch the maximal likelihood regions on the curve of expectation.

9. Amount of Information from Finite Samples

If now we want to apply our ideas to small samples, then instead of $\frac{1}{V}$ we must consider the quantity

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

which we may regard as the amount of information extracted by our method of estimation. It has already been shown that

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

so that the amount extracted can never exceed the amount supplied. We have now to consider the condition under which the whole of the available information can be extracted. The condition is that

$$\frac{1}{\phi}\,\frac{\partial\phi}{\partial\theta}$$

should be constant over any equistatistical region for all values of θ. In this case the equation of maximal likelihood will provide a sufficient statistic.

For if this is the case

$$\frac{\partial}{\partial\theta}(\log\phi)$$

depends, apart from θ, only on the set to which the sample belongs; in other words, it is a function of θ and T only (where T is the maximal likelihood estimate). Thus if f is the frequency with which any sample, or group of samples having the same T, occurs, then

$$\frac{1}{f}\cdot\frac{\partial f}{\partial\theta}=\psi(\theta,\mathrm{T}).$$

Now let the frequency of samples such that T lies in the range dT, and a second statistic, T′, lies in the range dT′, be F(θ, T, T′)dTdT′, then since the above equation will be true for all values of θ, we shall integrate it with respect to θ and obtain

$$\log f=\int\psi(\theta,\mathrm{T})\,d\theta+C$$

where C does not involve θ, but does involve the sample readings. If now we find the joint distribution of T and T′, it will come in the form

$$\xi(\theta, \mathrm{T}).\ \eta(\mathrm{T}, \mathrm{T}').\ d\mathrm{T}.\ d\mathrm{T}'$$

so that F is of the form

$$\xi(\theta, \mathrm{T}).\ \eta(\mathrm{T}, \mathrm{T}')$$

Hence it is demonstrated that T is a sufficient statistic.

10. Ancillary Information

When no sufficient statistic exists, then the original data cannot be replaced by a single statistic without loss of accuracy. It is of interest to see what can be done by calculating in addition to our estimate an ancillary statistic which shall be available in combination with our estimate in future calculations.

We have traced the loss of information, in cases in which no sufficient statistic exists, to the fact that $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta}$ is not constant over the equistatistical surfaces. By using the maximal likelihood equation $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta} = 0$, we do in fact impose this constancy over our chosen region, but only for the value $\theta = \hat{\theta}$, where $\hat{\theta}$ is tha estimate at which we actually arrive. Since, in general, this will not be exactly the true value of θ, $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta}$ will not generally be constant over our chosen equi-statistical region, if $\frac{\partial^2}{\partial\theta^2}\log\phi$ shows any variation among samples leading to the same estimate.

The expected loss of information due to such variation has for large samples been evaluated in the form

$$\frac{S\left\{\frac{1}{m}\left(m''-\frac{m'^2}{m}\right)^2\right\}}{S\left(\frac{m'^2}{m}\right)}-\frac{1}{n}S\left(\frac{m'^2}{m}\right)$$

$$-\frac{S^2\left\{\frac{m'}{m}\left(m''-\frac{m'^2}{m}\right)\right\}}{S^2\left(\frac{m'^2}{m}\right)}$$

where

$$m'=\frac{\partial m}{\partial\theta}$$

and

$$m''=\frac{\partial^2 m}{\partial\theta^2}.$$

Since the loss is due to using estimates for different samples having different values of the second differential coefficient of log ϕ, or of the logarithm of the likelihood at its maximum, we may examine the effect of using this second differential coefficient itself as ancillary information respecting our sample. The effect of this is that the variance of $\frac{1}{\phi}\frac{\partial\phi}{\partial\theta}$, which measures the information lost, will now have to be measured not over the entire equistatistical region, but only in a zone of that region for which $\frac{\partial^2 (\log\phi)}{\partial\theta^2}$ is constant. A higher degree of precision still would be obtained by using also the third, fourth,......... differential coefficients, supposing the likelihood function to be differentiable.

The utilisation of the information recovered by ancillary statistics is an interesting process. Suppose different large samples furnished maximal likelihood estimates T_1, $T_2,\ldots,T_k,\ldots T_r$ having second differential coefficients A_1, $A_2,\ldots,A_k\ldots.A_r$ then from these reduced data we can reconstruct the likelihood function in the form

$$\frac{1}{2}\sum_1^r A_k(\theta-T_k)^2$$

of which the maximum is given by

$$\sum_1^r A_k(\theta-T_k)=0$$

of

$$\hat{\theta}=\frac{\sum_1^r A_kT_k}{\sum_1^r A_k}$$

It appears in fact that the ancillary information furnished by the second differential coefficient is simply equivalent to giving our different estimates correct weights in place of the average weights appropriate to the sizes of the samples from which they were drawn. Such correct weights will in fact recover the whole of the information lost in the limit for large samples, although some information will still be lost from finite samples.

In other cases, as explained in " Two New Properties of Mathematical Likelihood," it is possible using the configuration of the sample as ancillary information to recover the whole of the lost information; these cases constitute a second group of solutions besides those in which sufficient statistics exist, in which estimates may be made exhaustive.

In general, the problem of recognising the character of configuration to be used in this way, is the Problem of the Nile, which I stated in the Mathematical Conference of the Tricentenary Celebrations at Harvard in 1936 :—

" The agricultural land of a pre-dynastic Egyptian village is of unequal fertility. Given the height to which the Nile will rise, the fertility of every portion of it is known with exactitude, but the height of the flood affects different parts of the territory unequally. It is required to divide the area, between the several households of the village, so that the yields of the lots assigned to each shall be in pre-determined proportion, whatever may be the height to which the river rises.''

In many cases a solution of this problem can be perceived intuitively. At present, however, no general analytical approach has been put forward. Consequently it cannot yet be said that exhaustive estimation is possible in all cases.

BIBLIOGRAPHY

1. " A Mathematical Examination of the Methods of Determining the Accuracy of an Observation by the Mean Error, and by the Mean Square Error." *Monthly Notices of R. A. S.*, LXXX, 758 (1920).

2. " On the Mathematical Foundations of Theoretical Statistics." *Phil. Trans., Series A*, Vol. 222, pp. 309-368 (1922).

3. " The conditions under which χ^2 measures the discrepancy between observation and hypothesis." *Journal R. S. S.*, Vol. LXXXVII, Part 3, pp. 442-450 (1924).

4. "Theory of Statistical Estimation." *Proc. Cam. Phil. Soc.*, Vol. 22, pp. 700-725 (1925).

5. " Two New Properties of Mathematical Likelihood." *Proc. Royal Soc.* A., Vol. 144, pp. 285-307 (1934).

6. " Probability, likelihood, and quantity of information in the logic of uncertain inference." *Proc. Royal Soc.* A, Vol. 146, pp. 1-8 (1934).

7. " The Logic of Inductive Inference." *J. R. S. S.*, Vol. XCVIII, pp. 39-82 (1935).

8. " Uncertain Inference." *Proc. American Acad. Arts and Sc.*, Vol. 71, No. 4, pp. 245-258 (1936).