Reproduced from the Proceedings of the Royal Society, A, 144: 285-307 (1934) with permission of the Royal Society

108

TWO NEW PROPERTIES OF MATHEMATICAL LIKELIHOOD

Author's Note (CMS 24.284a)*

* Reproduced from Contributions to Mathematical Statistics (1950) by permission of John Wiley & Sons

From a logician's point of view one of the most surprising results obtained by the theory of estimation is that not only the mathematical form of the inferences which can be rigorously drawn concerning the unknown parameters of the populations sampled, from the frequencies observed in a random sample, depends on the particular mathematical specification of this population, but that the logical nature of these inferences depends on this also.

The present paper is designed to illustrate the fact that, if one set of functional conditions is satisfied, there will exist sufficient statistics, while, if a second and distinct limitation is imposed on the problem, estimates may be made exhaustive, and the small sample problem solved with exactitude, by means of ancillary statistics.

Examples of each class are treated in detail, so that the reader may grasp clearly the peculiarities of the likelihood function which each implies. Both classes are of rather common occurrence, but beyond them it would appear that it is not possible to derive exact statements of fiducial probability from the primary inference supplied by the relative likelihood of all possible combinations of parametric values.

The particular contents of the paper are briefly sketched in the summary, page 306.

Two New Properties of Mathematical Likelihood.

By R. A. Fisher, F.R.S.

(Received December 14, 1933.)

1. *Introductory.*

To Thomas Bayes* must be given the credit of broaching the problem of using the concepts of mathematical probability in discussing problems of inductive inference, in which we argue from the particular to the general; or, in statistical phraselogy, argue from the sample to the population, from which, *ex hypothesi*, the sample was drawn. 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.

To the merit of broaching a fundamentally important problem, Bayes added that of perceiving, much more clearly than some of his followers have done, the logical weakness of the form of solution he put forward. Indeed we

* 'Phil. Trans.,' vol. 53, p. 370 (1763).

are told that it was his doubts respecting the validity of the postulate needed for establishing the method of inverse probability that led to his withholding his entire treatise from publication. Actually it was not published until after his death.

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 samples is, as was first shown by Bernoulli,

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

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 had *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

$$\frac{n!}{a!\,(n-a)!}\,x^a\,(1-x)^{n-a}f(x)\,dx; \tag{2}$$

since this sequence of events has occurred for some value of x, the expression (2) 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. Whereas, however, in relation to psychological judgment, likelihood has some resemblance to probability, the two concepts are wholly distinct, in that probability is appropriate to a class of cases in which uncertain inferences are possible from the general to the particular, while likelihood is appropriate to the class of cases arising in the problem of estimation, where we can draw inferences, subject to a different kind of uncertainty, from the particular to the general.

The primary properties of likelihood in relation to the theory of estimation have been previously demonstrated.* In the following sections I propose to exhibit certain further properties arising when the functional properties of the specification of the population fulfil certain special, but practically important, conditions.

2. *The Distribution of Sufficient Statistics.*

The essential feature of statistical estimates which satisfy the criterion of sufficiency is that they by themselves convey the whole of the information, which the sample of observations contains, respecting the value of the parameters of which they are sufficient estimates. This property is manifestly true of a statistic T_1, if for any other estimate T_2 of the same parameter, θ, the simultaneous sampling distribution of T_1 and T_2 for given θ, is such that given T_1, the distribution of T_2 does not involve θ; for if this is so it is obvious that once T_1 is known, a knowledge of T_2, in addition, is wholly irrelevant; and if the property holds for all alternative estimates, the estimate T_1 will contain the whole of the information which the sample supplies.

This remarkable property will be possessed when

$$\frac{\partial}{\partial\theta}\log L,$$

where L is the likelihood of θ for a given sample of observations, is the same function for all samples yielding the same estimate T_1; for on integrating the expression above with respect to θ, it appears that $\log L$ is the sum of two components, one a function only of θ and T_1, and the other dependent on the sample but independent of θ. If

$$f(T_1, T_2, \theta)\, dT_1\, dT_2$$

is the frequency with which samples yield estimates simultaneously in the ranges dT_1 and dT_2, it follows that

$$f(T_1, T_2, \theta) = \phi_1(T_1, \theta) \, . \, \phi_2(T_1, T_2);$$

where the first factor involves T_1 and θ only, and the second does not involve θ. The distribution of T_2 given T_1 will therefore be

$$\phi_2(T_1, T_2)\, dT_2 \Big/ \int \phi_2(T_1, T_2)\, dT_2$$

* Fisher, 'Proc. Camb. Phil. Soc.,' vol. 22, p. 700 (1925).

the integral being taken over all possible values of T_2, and in this expression the parameter θ is seen not to appear.

The condition that $\partial L/\partial\theta$ should be constant over the same sets of samples for all values of θ, which has been shown to establish the existence of a sufficient estimate of θ, thus requires that the likelihood is a function of θ, which, apart from a factor dependent on the sample, is of the same form for all samples yielding the same estimate T. The sufficiency of sufficient statistics may thus be traced to the fact that in such cases the value of T itself alone determines the form of the likelihood as a function of θ. If a conventional value such as unity is given to the maximum likelihood for any sample, the likelihood is thus expressible as a function of θ and T only, if T is the sufficient estimate. We shall use this property in obtaining a general form for the sampling distribution of sufficient statistics.

2.1. It will help if we take an illustrative example of this problem. Let the element of frequency in a distribution be given by

$$\frac{1}{\theta\,!}\,x^{\theta}e^{-x}\,dx$$

where the variate x can take any real value from 0 to ∞, and θ is an unknown parameter greater than -1. Consider the problem of estimating θ from a sample of n values of x.

If L is the likelihood of any possible value of θ,

$$\log L = -\,n \log \theta\,! + \theta S\,(\log x) - S\,(x),$$

and this is maximized for variations of θ, when $\theta = T$, where

$$\mathcal{F}\,(T) = \frac{1}{n}\,S\,(\log x),$$

and $\mathcal{F}\,(T)$ is the first differential of the logarithm of the factorial function. This is the equation of estimation by the method of maximum likelihood. It will be observed that apart from a constant factor the likelihood is expressible as a function of θ and T only, that is

$$L = A \exp\{-n \log \theta\,! + n\theta\mathcal{F}\,(T)\}$$

so that T is evidently a sufficient estimate.

2.2. The sampling distribution of our estimate must evidently be derived from that of the mean of the logarithms of the several values of x in the sample. Now the mean value of

$$e^{\frac{it}{n}\log x}$$

is

$$\int \frac{1}{\theta\,!} x^{\theta+\frac{it}{n}} e^{-x}\,dx = \frac{\left(\theta+\frac{it}{n}\right)!}{\theta\,!}.$$

By the familiar process of expanding the multiple integral in a product of single integrals, the mean value over all samples of

$$e^{it\frac{1}{n}\mathrm{S}\,(\log x)} = e^{it\mathcal{F}\,(\mathrm{T})}$$

is

$$\mathrm{M} = \frac{\left\{\left(\theta+\frac{it}{n}\right)!\right\}^n}{\{\theta\,!\}^n}$$

and this is the characteristic function of

$$\mathcal{F}\,(\mathrm{T})$$

from which its distribution may be inferred.

To determine the probability function knowing the characteristic function M (it) we may use the property of the sine integral

$$\int_0^\infty \frac{\sin u}{u}\,du = \frac{\pi}{2};$$

writing kt for u it appears that

$$\frac{1}{\pi}\int_0^\infty \sin kt\,\frac{dt}{t} = \tfrac{1}{2}$$

when k is positive, and $-\frac{1}{2}$ when k is negative; or that

$$\frac{1}{\pi}\int_0^\infty \{\sin\,(x-a)\,t - \sin\,(x-b)\,t\}\,\frac{dt}{t}$$

where $b > a$, is unity when $b > x > a$, and zero when x is less than a or exceeds b.

Consequently the Stieltjes integral

$$\int_a^b df\,(x) = \frac{1}{\pi}\int_0^\infty \frac{dt}{t}\int_{-\infty}^\infty \{\sin\,(x-a)\,t - \sin\,(x-b)\,t\}\,df\,(x);$$

writing

$$\sin\,(x-a)\,t = \frac{1}{2i}\left(e^{it\,(x-a)} - e^{-it\,(x-a)}\right)$$

gives us

$$\int_a^b df\,(x) = \frac{1}{2\pi}\int_0^\infty \frac{dt}{it}\,\{e^{-iat}\,\mathrm{M}\,(it) - e^{iat}\,\mathrm{M}\,(-it) - e^{-ibt}\,\mathrm{M}\,(it) + e^{ibt}\,\mathrm{M}\,(-it)\},$$

where $f(x)$ is the probability function of the variate x, and M is its characteristic function.

We may note that $\mathrm{M}(it)$ and $\mathrm{M}(-it)$ must be conjugate quantities, which may be written $\mathrm{R} \pm i\mathrm{I}$, then

$$-e^{-ibt}\,\mathrm{M}(it) + e^{ibt}\,\mathrm{M}(-it) = 2i\mathrm{R}\sin bt - 2i\mathrm{I}\cos bt$$

so that the integral takes the real form

$$\frac{1}{\pi}\int_0^\infty \frac{dt}{t}\{\mathrm{R}(\sin bt - \sin at) - \mathrm{I}(\cos bt - \cos at)\}.$$

Where the probability function is differentiable so that

$$df(x) = y\,dx$$

then

$$df(x) = y\,dx = \frac{dx}{2\pi}\int_0^\infty \{e^{-ixt}\,\mathrm{M}(it) + e^{ixt}\,\mathrm{M}(-it)\}\,dt$$

$$= \frac{dx}{\pi}\int_0^\infty \{\mathrm{R}\cos(tx) + \mathrm{I}\sin(tx)\}\,dt.$$

2.4. For the sufficient statistic T, the sampling distribution will therefore be given by

$$df = \frac{d\mathcal{F}(\mathrm{T})}{2\pi}\int_{-\infty}^{\infty} e^{-it\mathcal{F}(\mathrm{T})}\,\mathrm{M}(it)\,dt\,;$$

but

$$e^{-it\mathcal{F}(\mathrm{T})}\,\mathrm{M} = \frac{\mathrm{L}(\theta)}{\mathrm{L}\left(\theta + \frac{it}{n}\right)}\,;$$

hence the distribution may be directly inferred from the nature of the likelihood function in the form

$$df = \frac{\underline{\mathcal{F}}(\mathrm{T})\,.\,d\mathrm{T}}{2\pi}\int_{-\infty}^{\infty} \frac{\mathrm{L}(\theta)}{\mathrm{L}\left(\theta + \frac{it}{n}\right)}\,dt$$

or

$$\frac{n}{2\pi}\,\underline{\mathcal{F}}(\mathrm{T})\,d\mathrm{T}\int_{-\infty}^{\infty} \frac{\mathrm{L}(\theta)}{\mathrm{L}(\theta + it)}\,dt$$

where $\underline{\mathcal{F}}(\mathrm{T})$ stands for the second differential coefficient of log (T !).

We may illustrate the use of this formula by deriving the limiting forms for extreme values of θ.

Near the limit $\theta \to -1$ the general expression

$$\frac{nd\,(\mathcal{F}(\mathrm{T})\,)}{2\pi}\int_{-\infty}^{\infty} \frac{(\theta + it)\,!^{n}}{\theta\,!^{n}}\,e^{-int\mathcal{F}(\mathrm{T})}\,dt$$

may be rewritten, substituting

$$T = -1 + e^{-g}$$

$$\theta = -1 + e^{-\gamma},$$

and, since, as will be apparent, when γ is large, all important contributions will arise from small values of t, the limiting form of our distribution is

$$\frac{ne^g\,dg}{2\pi}\int_{-\infty}^{\infty}\left(1+\frac{it}{1+\theta}\right)^{-n} e^{inte^g}\,dt.$$

Writing t for te^{γ} we then have

$$\frac{n\,dg}{2\pi}\,e^{(g-\gamma)}\int_{-\infty}^{\infty}(1+it)^{-n}\,e^{inte^{(g-\gamma)}}\,dt$$

and writing z for $1 + it$

$$e^{-nu}\,\frac{n\,du}{2\pi i}\int z^{-n}\,e^{nzu}\,dz$$

where u stands for $e^{g-\gamma}$, and the integral is taken from $1 - i\,\infty$ to $1 + i\,\infty$, or in an open contour passing counter-clockwise to the right of the singularity, $z = 0$. Writing ζ for nuz we have

$$\frac{n\,(nu)^{n-1}\,e^{-nu}\,du}{2\pi i}\int\zeta^{-n}\,e^{\zeta}\,d\zeta,$$

where the integral does not now involve the variate u, and is evidently $2\pi i/(n-1)!$ The distribution now appears as

$$\frac{1}{(n-1)!}\,n^n\,u^{n-1}\,e^{-nu}\,du,$$

in which $e^{g-\gamma}$ may be substituted for u.

The probability integral in this case is given by the χ^2 distribution for $2n$ degrees of freedom. Thus if a sample of 10 had been taken, we have 20 degrees of freedom, and the 5% values are at $\chi^2 = 10\cdot851$ and $31\cdot410$.* Putting $nu = \frac{1}{2}\chi^2$ the 5% values of u are $0\cdot5426$ and $1\cdot5705$, whence those of $g - \gamma$ can be obtained, showing that in 90% of samples g will lie between $\gamma - 0\cdot6110$ and $\gamma + 0\cdot4514$. For given g, therefore, the fiducial probability is 5% that γ exceeds $g + 0\cdot6110$, or falls short of $g - 0\cdot4514$.

At the upper limit where $\theta \rightarrow \infty$ we may write

$$T = g^2, \quad \theta = \gamma^2,$$

* "Statistical Methods for Research Workers," Table III.

and since $g - \gamma$ remains finite for finite values of the probability function

$$\log \frac{\mathrm{T}}{\theta} = \frac{2(g-\gamma)}{\gamma}$$

and tends to zero. The general expression for the distribution, which is

$$\frac{nd\,(\mathcal{F}\,(\mathrm{T})\,)}{2\pi} \int_{-\infty}^{\infty} \frac{(\theta + it)\,!^{n}}{\theta\,!^{n}}\, e^{-int\mathcal{F}\,(\mathrm{T})}\, dt,$$

tends to the limiting form

$$df = \frac{n\,d\mathrm{T}}{2\pi\mathrm{T}} \int_{-\infty}^{\infty} \theta^{int}\, e^{-\frac{nt^2}{2\theta}}\, \mathrm{T}^{-int}\, dt$$

or substituting for T and θ

$$df = \frac{n\,dg}{\pi g} \int_{-\infty}^{\infty} e^{-2int\,(g-\gamma)/\gamma}\, e^{-nt^2/2\gamma^2}\, dt$$

$$= \sqrt{\frac{2n}{\pi}}\, e^{-2n(g-\gamma)^2}\, dg$$

showing that g tends in the limit to be normally distributed about the population value γ with variance $1/4n$. The 5% points of the distribution of g are therefore $\gamma \pm 1{\cdot}645/\sqrt{4n}$, and for a given g the 5% points of the fiducial distribution of γ are $g \pm 1{\cdot}645/\sqrt{4n}$.*

2.5. The interest of this form lies in the possibility of generalizing it for all sufficient statistics. For, let the equation of maximum likelihood have a solution

$$\phi\,(\mathrm{T}) = \mathrm{A}$$

where A is a symmetric function of the observations not involving the parameter θ. The expression for $\partial/\partial\theta \log \mathrm{L}$ must have been of the form

$$\mathrm{C}\,\{\mathrm{A}\psi'\,(\theta) - \phi\,(\theta)\,.\,\psi'\,(\theta)\}$$

where the possible factor C, if not a constant, must be a function of the observations which is expressible as a function of A, if the likelihood is to be expressible as a function of θ and T only.

The expression for log L then must be of the form

$$\mathrm{CA}\psi\,(\theta) - \mathrm{C}\int \phi\,(\theta)\, d\psi\,(\theta) + \mathrm{B}$$

* "Statistical Methods for Research Workers," Table I.

where B is a function of the observations only. That C a symmetric function of all the observations must be merely the number n in the sample appears from the fact that log L is the sum of expressions involving each observation singly. Hence

$$\mathrm{CA} = \mathrm{S}\,(\mathrm{X}), \qquad \mathrm{B} = \mathrm{S}\,(\mathrm{X}_1),$$

where X, X_1, are functions of the individual observations x. The likelihood is now the product

$$e^{-n\int \phi\,(\theta)\,d\psi\,(\theta)}\, e^{\psi\,(\theta)\,.\,\mathrm{S}\,(\mathrm{X})}\, e^{\mathrm{S}\,(\mathrm{X}_1)}$$

and

$$\frac{\mathrm{L}\,(\psi)}{\mathrm{L}\,(\psi + it)} = e^{-it\mathrm{S}\,(\mathrm{X})}\, e^{n\mathrm{F}_1\,(\psi+it)\,-n\mathrm{F}_1\,(\psi)}$$

where $\mathrm{F}_1\,(\psi)$ is written for $\int \phi\, d\psi$.

But the frequency function of the variate X was given by

$$e^{-\mathrm{F}_1\,(\psi)}\, e^{\mathrm{X}\psi}\, e^{\mathrm{X}_1}\, \frac{dx}{d\mathrm{X}}\, d\mathrm{X},$$

hence its characteristic function is

$$\mathrm{M}\,(it) = e^{\mathrm{F}_1\,(\psi+\mathrm{it})-\mathrm{F}_1\,(\psi)}$$

while that of S (X) is the nth power of this expression, hence the probability that S (X) lies between S_0 and S_1 is

$$\int df = \frac{1}{2\pi}\int_0^\infty \frac{dt}{it}\{e^{-i\mathrm{S}_0t}\,\mathrm{M}^n\,(it) - e^{i\mathrm{S}_0t}\,\mathrm{M}^n\,(-it) - e^{-i\mathrm{S}_1t}\,\mathrm{M}^n\,(it) + e^{i\mathrm{S}_1t}\,\mathrm{M}^n\,(-it)\}$$

$$= \frac{1}{2\pi}\int_0^\infty \frac{dt}{it}\left\{\frac{\mathrm{L}\,(\psi,\,\mathrm{S}_0)}{\mathrm{L}\,(\psi+it,\,\mathrm{S}_0)} - \frac{\mathrm{L}\,(\psi,\,\mathrm{S}_0)}{\mathrm{L}\,(\psi-it,\,\mathrm{S}_0)} - \frac{\mathrm{L}\,(\psi,\,\mathrm{S}_1)}{\mathrm{L}\,(\psi+it,\,\mathrm{S}_1)} + \frac{\mathrm{L}\,(\psi,\,\mathrm{S}_1)}{\mathrm{L}\,(\psi-it,\,\mathrm{S}_1)}\right\}$$

this being the general expression for the probability of any sufficient statistic falling within assigned limits; S_1 and S_0 being the limits of the known function $n\phi$ (T) of the sufficient estimate T.

2.6. The property that where a sufficient statistic exists, the likelihood, apart from a factor independent of the parameter to be estimated, is a function only of the parameter and the sufficient statistic, explains the principal result obtained by Neyman and Pearson in discussing the efficacy of tests of significance. Neyman and Pearson introduce the notion that any chosen test of a hypothesis H_0 is more powerful than any other equivalent test, with regard to an alternative hypothesis H_1, when it rejects H_0 in a set of samples having an assigned aggregate frequency ε when H_0 is true, and the greatest possible aggregate frequency when H_1 is true.

If any group of samples can be found within the region of rejection whose probability of occurrence on the hypothesis H_1 is less than that of any other group of samples outside the region, but is not less on the hypothesis H_0, then the test can evidently be made more powerful by substituting the one group for the other. Consequently, for the most powerful test possible the ratio of the probabilities of occurrence on the hypothesis H_0 to that on the hypothesis H_1 is less in all samples in the region of rejection than in any sample outside it. For samples involving continuous variation the region of rejection will be bounded by contours for which this ratio is constant. The regions of rejection will then be required in which the likelihood of H_0 bears to the likelihood of H_1, a ratio less than some fixed value defining the contour.

The test of significance is termed uniformly most powerful with regard to a class of alternative hypotheses if this property holds with respect to all of them. This evidently requires that the contours defined by the ratio of the likelihood of H_1 and H_0 shall be the same as those defined by the ratios of the likelihood of any two hypotheses in the class. If, therefore, T′ is a statistic defining these contours, and θ_1, θ_2, ..., are variable parameters defining the hypothetical populations, the likelihood of any hypothesis must be expressed in the form

$$L = Af(T', \theta_1, \theta_2, \ldots)$$

where A is a factor independent of the parameters.

The method of estimation by maximum likelihood, when applied to the form above, will yield equations for θ_1, θ_2, .., etc.

$$\phi_1(T', \theta_1, \theta_2, \ldots) = 0,$$
$$\phi_2(T', \theta_1, \theta_2, \ldots) = 0;$$

where

$$\phi_s = \frac{\partial}{\partial\theta_s} \log f,$$

and the solutions of these will give estimates of θ_1, θ_2, ..., which we may designate T_1, T_2, ..., in the form

$$T_1 = \psi_1(T')$$
$$T_2 = \psi_2(T'), \text{ etc.}$$

It is evident, at once, that such a system is only possible when the class of hypotheses considered involves only a single parameter θ, or, what comes to the same thing, when all the parameters entering into the specification of the population are definite functions of one of their number. In this case, the regions defined by the uniformly most powerful test of significance are those defined by the estimate of maximum likelihood, T. For the test to be uniformly most powerful, moreover, these regions must be independent of θ, showing that the statistic must be of the special type distinguished as sufficient. Such sufficient statistics have been shown to contain all the information which the sample provides relevant to the value of the appropriate parameter θ. It is inevitable therefore that if such a statistic exits it should uniquely define the contours best suited to discriminate among hypotheses differing only in respect of this parameter; and it is surprising that Neyman and Pearson should lay it down as a preliminary consideration that "the testing of statistical hypotheses cannot be treated as a problem in estimation." When tests are considered only in relation to sets of hypotheses specified by one or more variable parameters, the efficacy of the tests can be treated directly as the problem of estimation of these parameters. Regard for what has been established in that theory, apart from the light it throws on the results already obtained by their own interesting line of approach, should also aid in treating the difficulties inherent in cases in which no sufficient statistic exists.

3. *A Second Class of Parameters for which Estimation need Involve no Loss of Information.*

In the case of sufficient statistics the likelihood function is, apart from a constant factor, the same for all sets of observations which yield the same estimate by the method of maximum likelihood. A second case, of somewhat wider practical application, occurs when, although the sets of observations which provide the same estimate differ in their likelihood functions, and therefore in the nature and quantity of the information they supply, yet when samples alike in the information they convey exist for all values of the estimate and occur with the same frequency for corresponding values of the parameter..

The nature of the correspondence may be stated as follows: If $x_1, \ldots, x_n$ stands for a sample of n values of a variate x, the distribution of which is conditioned by a parameter, θ, then for any value of θ, there will be a definite probability

$$p(x, \theta)$$

of the occurrence of a variate less in value than x.

If, therefore, we take any other value of the parameter, say ϕ, there will, with continuous variates, always exist a series of observational values, y, corresponding to the original series x, such that

$$p(y, \phi) = p(x, \theta).$$

The samples x and y will, however, only correspond in the sense required for our present purpose if corresponding to any possible value, θ, a value, ϕ, can be found so that the relationship above holds for all values of x. If, in fact, the equation were solved for y, in the form

$$y = f(x, \theta, \phi)$$

it is required that f shall be of the form

$$f(x, \theta, \phi) = \mathrm{F}(x, \Omega) \qquad (3)$$

where Ω is a function of θ and ϕ, independent of the observations, and such that for any possible values of θ and Ω there exists a corresponding value of ϕ. Stated symmetrically it is required that some function of x and y can be equated to a function of θ and ϕ.

The typical case of such a relationship occurs in parameters of location. If the distribution of the variate x involves a parameter θ, such that the frequency with which x falls in any element dx of its range is a function of $(x - \theta)$, then θ may be called a parameter of location. In such a case the functional relationship (3) may be written

$$x - y = \theta - \phi$$

and is clearly of the form required.

Let us take an example in which there is no sufficient estimate, and in which the loss of information in estimating the unknown parameter even by the method of maximum likelihood is considerable. The distribution of x is a double exponential curve, the probability of x falling in the range dx, being

$$\tfrac{1}{2}e^{-|x-\theta|}\,dx.$$

The logarithm of the likelihood is

$$-\mathrm{S}\,|x - \theta|,$$

and this increases when θ is increased only if more observations are greater than those less than θ. The likelihood is therefore maximized if the number of observations is odd, by equating θ to the median observation; if the number of observations is even the likelihood is constant when θ has any value between the two central observations.

For a sample of an odd number, $n = 2s + 1$ of observations, the sampling distribution of the median is determinate, and the loss of information, if we use the median as an estimate, unsupplemented by the ancillary information which the sample contains, may be calculated. For, if the central observation lies at a distance u from the centre of the distribution, u being supposed positive, then the s highest values observed must each have fallen in a region comprising only

$$\tfrac{1}{2}e^{-u}$$

of the total frequency, while the s lowest values have fallen in the remaining region comprising

$$1 - \tfrac{1}{2}e^{-u}$$

of the total. Finally, the probability of the median itself falling in the range du is

$$\tfrac{1}{2}e^{-u}\,du,$$

so that compounding the independent probabilities into which the event has been analysed we have

$$df = \frac{(2s+1)\,!}{(s\,!)^2} \cdot (\tfrac{1}{2}e^{-u})^s\,(1 - \tfrac{1}{2}e^{-u})^s\,\tfrac{1}{2}e^{-u}\,du$$

as the probability of the median having a positive sampling error, u. As s is increased without limit we may write

$$u\sqrt{n} = t,$$

and the distribution tends to the limit

$$df = \frac{1}{\sqrt{2\pi}}\,e^{-\frac{1}{2}t^2}\,dt.$$

The amount of information derivable from a large sample of n thus tends to equality with n, as the size of the sample is increased. Since the information supplied by the independent observations is additive,* each must supply one unit, and a sample of $2s + 1$ observations must contain $2s + 1$ units of information. The quantity elicited by using the median, *i.e.*, by replacing the $2s + 1$ observations from the distribution

$$df = \tfrac{1}{2}\,e^{-|x-\theta|}\,dx;$$

* Fisher, 'Proc. Camb. Phil. Soc.' vol. 22, p. 700 (1925).

by a single observation from the distribution

$$df = \frac{(2s+1)\,!}{(s\,!)^2\,2^{2s+1}}\,e^{-s|u-\theta|}\,(2-e^{-|u-\theta|})^s\,e^{-|u-\theta|}\,du,$$

may be calculated from the mean value of

$$\left(\frac{d}{d\theta}\log\frac{df}{du}\right)^2,$$

or of

$$\left(s-\frac{se^{-|u-\theta|}}{2-e^{-|u-\theta|}}+1\right)^2.$$

When s exceeds unity the average values may be evaluated from the consideration that

$$\int_0^\infty \frac{(2s+1)\,!}{(s-1)\,!\;(s+1)\,!\;2^{2s+1}}\,e^{-(s+1)u}\,(2-e^{-u})^{s-1}\,e^{-u}\,du$$

represents the probability that at least $s+2$ observations have positive, and only $s-1$ observations negative, deviations, and may therefore be equated to

$$\tfrac{1}{2}-\frac{(2s+1)\,!}{s\,!\;(s+1)\,!\;2^{2s+1}}.$$

The mean value of $e^{-|u-\theta|}/(2-e^{-|u-\theta|})$ is therefore found to be

$$\frac{s+1}{s}\left(1-\frac{(2s+1)\,!}{s\,!\;(s+1)\,!\;2^{2s}}\right).$$

Similarly, the mean value of $e^{-2|u-\theta|}/(2-e^{-|u-\theta|})^2$ is

$$\frac{(s+1)\,(s+2)}{s\,(s-1)}\left(1-\frac{(2s+1)\,!}{s\,!\;s+1\,!\;2^{2s}}-\frac{(2s+1)\,!}{(s-1)\,!\;(s+2)\,!\;2^{2s}}\right).$$

The amount of information provided by the median of $2s+1$ observations is therefore

$$(s+1)^2-2s\,(s+1)\;.\;\frac{s+1}{s}+s^2\,.\,\frac{(s+1)\,(s+2)}{s\,(s-1)}$$

$$+\;2\,(s+1)^2\,.\,\frac{(2s+1)\,!}{s\,!\;(s+1)\,!\;2^{2s}}-\frac{s\,(s+1)\,(s+2)}{s-1}\,.\,\frac{(2s+1)\,!}{s\,!\;(s+1)\,!\;2^{2s}}$$

$$-\;\frac{s\,(s+1)\,(s+2)}{s-1}\;\frac{(2s+1)\,!}{(s-1)\,!\;(s+2)\,!\;2^{2s}}$$

or

$$\frac{(s+1)(2s+1)}{(s-1)}\left\{1-\frac{(2s)\,!}{(s\,!)^2\,2^{2s-1}}\right\}.$$

In the special case, $s = 1$, the general method fails, and a direct integration yields the value

$$12\,(\log 2 - \tfrac{1}{2}),$$

which is the limit to which the general expression tends as $s \rightarrow 1$.

The median is an efficient estimate in the sense of the theory of large samples, for the ratio of the amount of information supplied to the total available tends to unity as the sample is increased. Nevertheless, the absolute amount lost increases without limit. As s increases, this amount lost,

$$2s+1-\frac{(s+1)(2s+1)}{s-1}\left\{1-\frac{2s\,!}{(s\,!)^2\,2^{2s-1}}\right\},$$

may be replaced by

$$\frac{2\,(2s+1)}{s-1}\left(\frac{s+1}{\sqrt{\pi\,(s+\frac{1}{4})}}-1\right)$$

approximately, or by $4\,(\sqrt{s/\pi}-1)$. Thus with $s = 314$, for a sample of 629 observations, the loss of information is near to 36 units, or the value of about 36 observations.

It is a matter of no great practical urgency, but of some theoretical importance, to consider the process of interpretation by which this loss can be recovered. Evidently, the simple and convenient method of relying on a single estimate will have to be abandoned. The loss of information has been traced to the fact that samples yielding the same estimate will have likelihood functions of different forms, and will therefore supply different amounts of information. When these functions are differentiable successive portions of the loss may be recovered by using as ancillary statistics, in addition to the maximum likelihood estimate, the second and higher differential coefficients at the maximum. In general we can only hope to recover the total loss, by taking into account the entire course of the likelihood function.

In our particular problem the curve of likelihood is a succession of exponential arcs, having n discontinuities at the values of the n observations of the sample, the exponent changing by -2, as each observation is passed in a positive direction. For the same value of our estimate, the median observation, this function will have very different forms according to the length of the intervals which separate the median from its successive neighbours. Any samples, however, in which these $n-1$ intervals are the same will have the same likelihood function. More explicitly the likelihood of the parameter having a value ϕ as judged from the series of observations $y_1, \ldots, y_n$ will be equal to the likelihood of its value being θ as judged from the series $x_1, \ldots, x_n$, if

$$x - y = \theta - \phi$$

for each pair of observations in the pair of samples.

We may specify the configuration of a sample by a series of positive non-decreasing numbers $a_1, \ldots, a_s$ representing the positive deviations from the median of the s largest observations, and a second series of positive non-decreasing numbers $a'_1, \ldots, a'_s$ representing the excesses of the median over the s smallest observations, so that if T is the median value the n observations are represented by $\mathrm{T} - a'_s, \ldots, \mathrm{T} - a'_1, \mathrm{T}, \mathrm{T} + a_1, \ldots, \mathrm{T} + a_s$.

The probability of occurrence of any series of observations, the true centre of the distribution being θ,

$$\mathrm{L}\, dx_1, \ldots, dx_n$$

may now be written

$$n\,!\ \mathrm{L} \,.\, \frac{\partial\,(x_1 \ldots x_n)}{\partial\,(\mathrm{T}, a_1 \ldots a_s, a'_1 \ldots a'_s)}\, d\mathrm{T}\, da_1 \ldots da_s\, da'_1 \ldots da'_s$$
$$= n\,!\ \mathrm{L} \,.\, d\mathrm{T} da_1 \ldots da_s\, da'_1 \ldots da'_s$$

where, if, for example, θ lies between $\mathrm{T} - a'_p$ and $\mathrm{T} - a'_{p-1}$

$$\mathrm{L} = \frac{1}{2^n}\, e^{-(2p-1)(\mathrm{T}-\theta) - \mathrm{S}'_1(a+a') + 2(a'_1 + \ldots + a'_{p-1})}.$$

Given the configuration of the sample, therefore, the probability that T lies in a range $d\mathrm{T}$, between the limits $\theta + a'_{p-1}$ and $\theta + a'_p$ is

$$df = \frac{1}{\mathrm{A}}\, e^{2(a'_1 + \ldots + a'_{p-1})}\, e^{-(2p-1)(\mathrm{T}-\theta)}\, d\mathrm{T}$$

of which the integral between these limits is

$$\frac{1}{(2p-1)\,\mathrm{A}}\, e^{2(a'_1 + \ldots + a'_{p-1})} \left(e^{-(2p-1)a'_{p-1}} - e^{-(2p-1)a'_p}\right),$$

and A is equal to the sum of all such integrals

$$1 - e^{-a'_1} + \tfrac{1}{3}\left(e^{-a'_1} - e^{2a'_1 - 3a'_2}\right) + \tfrac{1}{5}\left(e^{2a'_1 - 3a'_2} - e^{2a'_1 + 2a'_2 - 5a'_3}\right) + \ldots$$
$$+ 1 - e^{-a_1} + \tfrac{1}{3}\left(e^{-a_1} - e^{2a_1 - 3a_2}\right) + \tfrac{1}{5}\left(e^{2a_1 - 3a_2} - e^{2a_1 + 2a_2 - 5a_3}\right) + \ldots$$

Apart from the details of the analysis, however, it is apparent that *if attention is confined to samples having a given configuration* the sampling distribution of

T for a given θ is found from the likelihood of θ for a given T, the probability curve in the first case being the mirror image of the likelihood curve in the second.

To evaluate the amount of information supplied by this distribution we must evaluate the mean square of

$$\frac{d}{d\theta} \log \frac{df}{d\mathrm{T}}.$$

Now, if T lies between $\theta + a'_{p-1}$ and $\theta + a'_p$,

$$\log \frac{df}{d\mathrm{T}} = -(2p-1)(\mathrm{T}-\theta)$$

so that in this case

$$\left(\frac{d}{d\theta} \log \frac{df}{dt}\right)^2 = (2p-1)^2,$$

and the amount of information supplied by our estimate, *in conjunction with a specification of the configuration of the sample from which it was obtained*, is

$$\frac{1}{\mathrm{A}}\{1 - e^{-a'_1} + 3\,(e^{-a'_1} - e^{2a'_1-3a'_2}) + 5\,(\ldots$$
$$+ 1 - e^{-a_1} + 3\,(e^{-a_1} - e^{2a_1-3a_2}) + 5\,(\qquad\qquad \}.$$

This value will differ greatly from sample to sample. Thus, if a_1 and a'_1 were both large, so that the median lies in a considerable range otherwise unoccupied by observations, the amount of information approaches unity; at the other extreme if a_s and a'_s were both so small that e^{-a_s} is near to unity, then

$$\mathrm{A} \rightarrow 2/(2s+1),$$

and the amount of information rises to $(2s+1)^2$, or n^2.

To find the average value of the amount of information derivable from the median, in conjunction with the configuration of the sample, we may note that the probability for a given configuration that $s+p$ observations shall exceed, and $s-p+1$ fall short of the true value is

$$\frac{1}{(2p-1)\,\mathrm{A}}\,(e^{2a'_1+2a'_2+\ldots-(2p-3)\,a'_{p-1}} - e^{2a'_1+2a'_2+\ldots-(2p-1)\,a'_p})$$

and that the amount of information is obtained by multiplying this probability by $(2p-1)^2$ and adding for all values of p.

The average information for all configurations may, therefore, be found from the total probability for all configurations that exactly $s+p$ observations shall exceed the true value; since the probability of exceeding θ is $\frac{1}{2}$ independently for each observation, the probability is

$$\frac{1}{2^n}\frac{n\,!}{(s+p)\,!\;(s-p+1)\,!}$$

and this, multiplied by $(2p-1)^2$, and added for all values of p, will give the average amount of information. The probabilities are the terms of the expansion of

$$(\tfrac{1}{2}+\tfrac{1}{2})^n,$$

and $(2p-1)$ is twice the deviation from the mean corresponding to each value of p. The variance of the binomial is well known to be exactly $\frac{1}{4}n$, and the average amount of information used is consequently found to be exactly n, equal to the total amount known to be contained on the average in the sample.

The process of taking account of the distribution of our estimate in samples of the particular configuration observed has therefore recovered the whole of the information available. This process will seldom be so convenient as the use of an estimate by itself, without reference to the configuration, for instead of replacing the n observations by a single value, we now have to take account of all their values individually. Actually, indee1, in this case only the central group of values matters greatly, but in general the theoretical process illustrated here uses the available information exhaustively, only at the expense of abandoning the convenience of disregarding all properties of the sample beyond the best estimate it can provide. The reduction of the data is sacrificed to its complete interpretation.

The frequency distribution, which makes this complete interpretation possible, is the mirror image of the likelihood function. Thus if T_1 is the estimate (the median) derived from the actual sample observed, and $L\,(\theta-T_1)$ is the likelihood derived from this sample of any value of θ, then the sampling distribution of T for any value of θ, in samples of the same configuration is given by

$$df \propto L\,(\theta-T)\,dT.$$

This is an extremely simple derivation of the sampling distribution of the estimate of maximum likelihood from the form of the likelihood function.

4. *The Simultaneous Estimation of Location and Scaling.*

In a very frequent class of cases not only the origin but the scalo of the distribution is also represented by a parameter to be estimated from the observations. The frequency element is then of the form

$$f\,(\xi)\,d\xi,$$

where

$$\xi = \frac{x - \theta_1}{\theta_2}.$$

In such cases it is obvious that the sample of values ξ in relation to any values α_1 and α_2 of the parameters corresponds in the sense of section 3 to the sample of values of x in relation to the values $\theta_1 + \alpha_1\theta_2$ and $\theta_2\alpha_2$, and a double series of samples exists corresponding to any sample observed.

The samples will have all the same configurations in the sense that supposing any two observations of the sample, such as the lowest and the lowest but one in value, have values a, $a + b$ then the other members of the sample will be

$$a + bt_p \qquad p = 1, \ldots, n - 2,$$

where the $n - 2$ values of t_p specify the configuration, and are the same for all samples of which the configuration is the same.

The frequency element

$$\mathrm{L}dx_1, \ldots dx_n,$$

giving the frequency with which the n observations fall within assigned values, may then be replaced by

$$\mathrm{L}\frac{\partial(x_1, \ldots x_n)}{\partial(a, b, t, \ldots t_{n-2})}\, da\, db\, dt, \ldots dt_{n-2},$$

where the Jacobian is simply

$$\begin{vmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & 1 & 0 & \ldots & 0 \\ 1 & t_1 & b & \ldots & 0 \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ 1 & t_{n-2} & 0 & \ldots & b \end{vmatrix}$$

or b^{n-2}. The simultaneous frequency distribution of a and b is therefore given by

$$df \propto \mathrm{L}b^{n-2}\, da\, db.$$

Now, it is evident that the estimates of θ_1 and θ_2 from such a sample will be

$$\mathrm{T}_1 = a + \lambda b,$$
$$\mathrm{T}_2 = \mu b,$$

where λ and μ depend only on the configuration of the sample. Hence

$$\frac{\partial(\mathrm{T}_1, \mathrm{T}_2)}{\partial(a, b)} = \begin{vmatrix} 1 & 0 \\ \lambda & \mu \end{vmatrix} = \mu$$

and the distribution of these estimates in samples of the same configuration will be

$$df \propto \mathrm{LT}_2^{n-2}\, d\mathrm{T}_1\, d\mathrm{T}_2, \qquad (4)$$

where in L, $\mathrm{T}_1 + u_p\mathrm{T}_2$ is substituted for x_p, $p = 1, \ldots, n$, the n values of u being known for the configuration observed.

If, therefore, we choose to take into account not merely the sampling distribution of our estimates for samples of all configurations, distributions which will involve, apart from the parameters of the population, only these two statistics, but rather the special simultaneous distribution for the particular configuration observed, we may obtain this special distribution directly from the form of the likelihood function.

Since, moreover, the whole course of the likelihood function is taken into account, it is, from this point of view evident that no information can be lost. An independent analytical proof of this is as follows; it is equally applicable to information in respect of θ_1 and of θ_2.

The information respecting θ_1 contained in a single observation from the distribution (4) is numerically equal to the average value of

$$\left(\frac{\partial}{\partial\theta_1}\log \mathrm{L}\right)^2$$

for all values of $(\mathrm{T}_1 - \theta_1)$ from $-\infty$ to ∞, or, otherwise, to the average value of

$$\left\{\frac{\partial}{\partial\theta_1}\mathrm{S}(\log f)\right\}^2$$

where $f(x - \theta_1)$ is the frequency of an observation falling in the range dx. The average for all values of T_1 is, for any particular observation, the average for all values of x. Now the average value of

$$\frac{\partial}{\partial\theta_1}\log f$$

is zero, for

$$\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta_1}\log f \,.\, f\, dx = \int_{-\infty}^{\infty}\frac{\partial f}{\partial\theta_1}\,.\, dx$$

which is zero, since the total frequency is unity, independent of θ_1. But the average value for all values of $(\mathrm{T}_1 - \theta_1)$ and for all configurations including variations of T_2, is the average value for all possible samples. We may apply this principle to the expression

$$S^2\left(\frac{\partial}{\partial\theta}\log f\right)$$

when all the values of x are independent. Then the average value of the square of the sums of n terms, independent and all having a mean value zero, is n times the mean square of each of them, or n times the mean value of

$$\left(\frac{\partial}{\partial\theta_1}\log f\right)^2$$

for all values of x from $-\infty$ to ∞, which is, by definition, the amount of information supplied by a sample of n observations. Hence the average amount of information respecting θ_1 supplied by (4) for all configurations is the entirety of that supplied by the data.

With respect to θ_2, we require the average value of

$$S^2\left(\frac{\partial}{\partial\theta_2}\log f\right)$$

for all values of T_2 from 0 to ∞. The average of this for all configurations and for all values of T_1, again reduces to the mean value of

$$n\left(\frac{\partial}{\partial\theta_2}\log f\right)^2$$

for all values of x from $-\infty$ to ∞, and so to the average amount of information contained in a sample of n observations.

Summary.

(I) Reasons are given for the use of mathematical likelihood in problems of inductive inference.

(II) When a statistic exists, satisfying the criterion of sufficiency, the likelihood function involves only that statistic.

(III) An example is given of a sufficient statistic, and its sampling distribution is expressed in terms of the likelihood function.

(IV) This property is generalized for all cases of simple estimation, where a sufficient statistic exists.

(V) It is shown that these cases and only these supply tests of significance of the kind termed by Neyman and Pearson " uniformly most powerful " with regard to a class of alternative hypothesis.

(VI) Where no sufficient statistic exists the precision of estimation may in general be enhanced by the use of ancillary statistics. A class of cases is defined and illustrated in which the totality of the ancillary information supplied by the observations may be utilized.

(VII) This process gives a very simple derivation of sampling distributions, in which there is no loss of information, even for small samples.