42

THEORY OF STATISTICAL ESTIMATION

Author's Note (CMS 11.699a)

In this paper an explanation of the theory of estimation is attempted, more compact and businesslike than was possible in 1922 (Paper 18). In particular, clarity has been gained by distinguishing the theory of large samples, in which all estimates are normally distributed, and all efficient estimates are equivalent, more emphatically from the theory of small samples, in which it is necessary to distinguish between different methods of estimation, all of which are efficient in large samples, and which moreover in finite samples give curves of sampling error of various forms, for which the variance is of no special interest.

The transition is effected by recognising that the limiting form in large samples of the distribution of efficient statistics supplies an intrinsic measure of the amount of information, as to any unknown parameter, supplied by data of the kind investigated, and that this measure is applicable equally to the original data, and to any proposed estimates calculated from them. A simple proof is given on page 717 that no gain in intrinsic accuracy can accrue from any process of statistical reduction, so that if there is no loss the maximum precision has been obtained. The method gives also the simple condition for zero loss leading both to the special case of sufficient statistics, and to the possibility of diminishing or obviating such loss altogether, by the use of ancillary information.

From this point of view the paper is a study of the conditions in which information is lost. In an important class of cases it is shown that, by using the method of maximum likelihood, the loss is less than by other efficient methods, and tends to a finite value in large samples. Moreover, for large samples a series of ancillary values is indicated successively reducing the limiting loss of information to quantities of the order of N^{-1}, N^{-2}, $\cdots$, where N is the number of observations, or, in general, a measure of the extent of the observational record. As these quantities are the successive derivatives at

Proceedings of the Cambridge Philosophical Society, 22: 700-725, (1925).

its maximum of the likelihood function, the inference is approached, though it seems not to be drawn in this paper, that exhaustive estimation is achieved by a set of statistics which together are capable of specifying the entire course of the likelihood function.

A correction has been made in the analysis on pages 721 and 722, in the expression for the additional loss of information using efficient statistics, which, unlike that derived from maximising the likelihood, are not linear in the frequencies. The correct formulae are somewhat simpler than those originally given.

Theory of Statistical Estimation. By Mr R. A. Fisher, Gonville and Caius College.

[*Received* 17 March, *read* 4 May, 1925.]

PREFATORY NOTE.

It has been pointed out to me that some of the statistical ideas employed in the following investigation have never received a strictly logical definition and analysis. The idea of a frequency curve, for example, evidently implies an infinite hypothetical population distributed in a definite manner; but equally evidently the idea of an infinite hypothetical population requires a more precise logical specification than is contained in that phrase. The same may be said of the intimately connected idea of random sampling. These ideas have grown up in the minds of practical statisticians and lie at the basis especially of recent work; there can be no question of their pragmatic value. It was no part of my original intention to deal with the logical bases of these ideas, but some comments which Dr Burnside has kindly made have convinced me that it may be desirable to set out for criticism the manner in which I believe the logical foundations of these ideas may be established.

The idea of an infinite hypothetical population is, I believe, implicit in all statements involving mathematical probability. If, in a Mendelian experiment, we say that the probability is one half that a mouse born of a certain mating shall be white, we must conceive of our mouse as one of an infinite population of mice which might have been produced by that mating. The population must be infinite for in sampling from a finite population the fact of one mouse being white would affect the probability of others being white, and this is not the hypothesis which we wish to consider; moreover, the probability may not always be a rational number. Being infinite the population is clearly hypothetical, for not only must the actual number produced by any parents be finite, but we might wish to consider the possibility that the probability should depend on the age of the parents, or their nutritional conditions. We can, however, imagine an unlimited number of mice produced upon the conditions of our experiment, that is, by similar parents, of the same age, in the same environment. The proportion of white mice in this imaginary population appears to be the actual meaning to be assigned to our statement of probability. Briefly, the hypothetical population is the conceptual resultant of the conditions which we are studying. The probability, like other statistical parameters, is a numerical characteristic of that population.

We only need the conception of an infinite hypothetical population, in connection with random sampling. The ultimate logical elucidation of the one idea implies that of the other. Also, the word infinite is to be taken in its proper mathematical sense as denoting the limiting conditions approached by increasing a finite number indefinitely. I imagine that an exact meaning can be given to all the ideas required by some process such as the following.

Imagine a population of N individuals belonging to s classes, the number in class k being $p_k N$. This population can be arranged in order in $N!$ ways. Let it be so arranged and let us call the first n individuals in each arrangement a sample of n. Neglecting the order within the sample, these samples can be classified into the several possible types of sample according to the number of individuals of each class which appear. Let this be done, and denote the proportion of samples which belong to type j by q_j, the number of types being t. Consider the following proposition.

Given any series of proper fractions $P_1, P_2, \ldots, P_s$, such that $S(P_k)=1$, and any series of positive numbers $\eta_1, \eta_2, \ldots, \eta_t$, however small, it is possible to find a series of proper fractions $Q_1, Q_2, \ldots, Q_t$, and a series of positive numbers $\epsilon_1, \epsilon_2, \ldots, \epsilon_s$, and an integer N_0, such that, if

$$N > N_0$$

and $\quad |p_k - P_k| < \epsilon_k$ for all values of k,

then will $\quad |q_j - Q_j| < \eta_j$ for all values of j.

I imagine it possible to provide a rigorous proof of this proposition, but I do not propose to do so. If it be true, we may evidently speak without ambiguity or lack of precision of an infinite population characterised by the proper fractions, P, in relation to the random sampling distribution of samples of a finite size n.

It will be noticed that I provide no definition of *a* random sample, and it is not necessary to do so. What we have to deal with in all cases is a random sampling distribution of samples, and it is only as a typical member of such a distribution that a random sample is ever considered.

When in 1921 the author put forward in the *Phil. Trans.* a paper(1)* on mathematical statistics he was principally concerned, in respect of problems of estimation, with the practical importance of making estimates of high efficiency, i.e. of using statistics which embody a large proportion of the relevant information available in the data, and which ignore, or reject along with the irrelevant information, only a small proportion of that which is relevant. Many of the properties of efficient statistics, such as that even moderate inefficiency of estimation will vitiate tests of goodness of fit, were at that time unknown, and the further discrimination among statistics within the efficient group, a discrimination which is essential to the advance of the theory of small samples, was left in much obscurity. Further work along the lines of the 1921 paper has, however, cleared up the main outstanding difficulties, and seems to make possible a theory of statistical estimation with some approach to logical completeness.

1. *The problem of estimation.*

Any body of numerical observations, or qualitative data thrown into a numerical form as frequencies, may be interpreted as a random sample of some infinite hypothetical population of possible values. Problems of estimation arise when we know, or are willing to assume, the form of the frequency distribution of the population, as a mathematical function involving one or more unknown parameters, and wish to estimate the values of these parameters by means of the observational record available. A statistic may be defined as a function of the observations designed as an estimate of any such parameter. The primary qualifications of satisfactory statistics may most readily be seen by their behaviour when derived from large samples.

A statistic, to be of any service, must tend to some fixed value

* See numbered list of references on p. 725.

as the number in the sample is increased; more precisely if T 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 $| T - T_\infty |$ exceeds ϵ, tends to zero as n tends to infinity. One example will suffice to illustrate the class of statistics which fail to fulfil this condition.

If the frequency with which the variate x falls into the range dx, be given by

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

where m is the unknown parameter representing the centre of the symmetrical frequency curve of x, then it is not difficult to show that the arithmetic mean of any number of independent values of x, will be distributed in exactly the same distribution as a single value of x. If the observational material consisted of 1000 values of x, we should be able from it to estimate the value of m with some precision; but if we were to replace the actual observations by their mean, our action would be equivalent to discarding 999 of the observations, and retaining one of them chosen at random. Clearly the mean is a useless statistic for our purpose in that it does not tend to a fixed value as the size of the sample is increased.

2. *Consistent statistics.*

When T tends to a limiting value T_∞, the latter will be some determinate function of the unknown parameters. If, therefore, T is to be used for purposes of estimation, it must be equated to one particular parameter, or function of the parameters, and if it is equated to some other function its use will be inconsistent, though perhaps approximately accurate. A statistic is said to be a consistent estimate of any parameter, if when calculated from an indefinitely large sample it tends to be accurately equal to that parameter. The criterion of consistency has been widely used in the development of statistical methods, and too often it has been the only criterion employed. For example the "method of moments" consists merely in evaluating a number of arbitrarily chosen statistics, and equating as many of them as may be necessary to the corresponding series of parametric functions. Estimates of the parameters may be obtained from these equations, but they are often estimates of little value. In the example given above we have shown how little value has the mean, the first moment, in locating a particular curve, one of the Pearsonian types, in fitting which the method of moments has been so extensively used.

In a special group of cases the criterion of consistency is adequate alone to give a complete solution. If the number of frequency classes is only one greater than the number of adjustable parameters, then for each parameter there is only one consistent statistic, and this of course is the one which must be used. Generally, however, there are a great number of possible statistics available, all of them consistent, but by no means all of equal value.

3. *Efficient statistics.*

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. In such cases the characteristics of any particular statistic, for large samples, are completely specified by (i) its bias, and (ii) its variance. The question of bias is only of preliminary interest; if b is the mean value of $T - T_\infty$, then for consistent statistics b must tend to zero with increasing samples. If we wish to make tests of significance for the deviation of T from some hypothetical value, then b must fall off more rapidly than $n^{-\frac{1}{2}}$; if, finally, we wish to use mean values of T from a number of finite samples, then b must be actually zero, or at least a small quantity of an order determined by the number of such samples to be used. In any case a knowledge of the exact form of the distribution of T will enable us to eliminate any disadvantages from which a statistic might seem to suffer by reason of bias.

Such knowledge is, however, of no avail to repair the defects of a statistic in respect of variance. 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.

For example, in estimating the value of the standard deviation of a normal distribution from a sample of n values, two methods have frequently been employed. If

$$ns_1 = \sqrt{\frac{\pi}{2}}\, S\,(|\,x - \bar{x}\,|),$$

where S stands for summation over the sample, s_1 is an estimate of the true value σ, based on the method of the mean error. It has been shown (2) that the mean value of s_1 in random samples is

$$\sigma\sqrt{\frac{n-1}{n}}$$

while the variance of s_1 is

$$\frac{(n-1)\,\sigma^2}{n^2}\left(\frac{\pi}{2}+\sqrt{n\,(n-2)}-n+\sin^{-1}\frac{1}{n-1}\right).$$

If, also, s_2 is given by the equation of the mean square error

$$n{s_2}^2 = S\,(x-\bar{x})^2$$

then the mean value of s_2 is

$$\sigma\sqrt{\frac{2}{n}}\cdot\frac{\frac{n-2}{2}\,!}{\frac{n-3}{2}\,!},$$

(where $x!$ is used as equivalent to $\Gamma\,(x+1)$, whether x is an integer or not); while the variance of s_2 is

$$\frac{\sigma^2}{n}\left[n-1-2\left(\frac{\frac{n-2}{2}\,!}{\frac{n-3}{2}\,!}\right)^2\right].$$

The latter happens to be an efficient statistic, and for large samples the variance reduces to $\frac{\sigma^2}{2n}$ while for large samples the variance of s_1 reduces to

$$\frac{\sigma^2}{2n}\,(\pi-2).$$

Evidently s_1 is not an efficient statistic, but has an efficiency of nearly 88 per cent. From a body of 800 observations it will derive an estimate of about the same value, as that obtained by s_2 from 700 observations. That is to say the behaviour of the two statistics for large samples indicates that about one-eighth of the information available is rejected if s_1 is used, while if s_2 is employed the whole is retained. An exact knowledge of the distribution of s_1 would not enable us to recover the lost information, for if the sample were increased without limit, and consequently the distribution brought infinitely near to the normal form, nevertheless the fraction of information lost tends to a fixed value.

4. *Properties of efficient statistics.*

Some simple properties of efficient statistics may be derived directly from their definition, e.g. their correlational properties (3). The correlation between any two statistics, both efficient estimates of the same parameter, tends to + 1 as the sample is increased.

For if A and B be two such statistics, let the variance of each be $\frac{\sigma^2}{n}$ and the correlation between them be r; also let

$$C = \tfrac{1}{2}(A + B),$$

then C will be a statistic providing a consistent estimate of the same parameter, but the variance of C is

$$\frac{\sigma^2}{n}\left(\frac{1+r}{2}\right),$$

and this by hypothesis must not be less than $\frac{\sigma^2}{n}$ therefore r cannot be less than $+1$; but r cannot be greater than $+1$; therefore $r = +1$.

For large samples therefore all efficient statistics are equivalent, and if in practical work we were only concerned with infinitely large samples, the theory of estimation would not require development beyond the stipulation that statistics should be efficient.

If A is an efficient statistic, and B is an inefficient estimate of the same parameter with efficiency equal to E, then in large samples the correlation between A and B tends to a limit $r = +\sqrt{E}$. For if from them we compound a new statistic C, such that

$$(1 + E - 2r\sqrt{E})\,C = (1 - r\sqrt{E})\,A + (E - r\sqrt{E})\,B,$$

then C will be an estimate of the same parameter with variance

which reduces to $$\frac{\sigma^2}{n}\cdot\frac{1 - r^2}{(1 - r^2) + (r - \sqrt{E})^2},$$

and if r does not tend to the limiting value $+\sqrt{E}$, this will be less than the variance of A, which is impossible; hence $r = +\sqrt{E}$. It should be noted that in making a new statistic with variance as low as that of A, when $r = +\sqrt{E}$, the above equation for C degenerates into $C = A$. In other words if we have an efficient statistic and an inefficient estimate of the same parameter, the best use we can make of these two values, at any rate with large samples, will be to ignore the latter entirely. Any compound of the two will be less efficient than A.

For example, if a quantity x be normally distributed with variance σ^2, then it is well known that the mean of a sample of n is also distributed normally with variance σ^2/n. The mean in this case is an efficient statistic; the median is a second statistic which may be used to locate the curve. If

$$\phi(a) = \sqrt{\frac{2}{\pi}}\int_0^a e^{-\frac{1}{2}t^2}\,dt$$

then if a is the central value of a sample of n values (n being odd) it appears that the probability that a shall fall into the range da is

$$\frac{n!}{\left(\frac{n-1}{2}!\right)^2}\cdot\frac{1}{\sqrt{2\pi}}\,e^{-\frac{a^2}{2\sigma^2}}\,\frac{da}{\sigma}\,.\,2^{-(n-1)}\left\{1-\phi^2\left(\frac{a}{\sigma}\right)\right\}^{\frac{n-1}{2}}$$

When n is large, $\phi\left(\frac{a}{\sigma}\right)$ in this expression must be small, and may be replaced by

$$\frac{a}{\sigma}\sqrt{\frac{2}{\pi}},$$

and the factor involving ϕ by

$$e^{-\frac{(n-1)a^2}{\pi\sigma^2}},$$

so that the variance of a for large samples multiplied by n tends to the limit

$$\frac{\pi}{2}\sigma^2.$$

The efficiency of the median in locating the normal curve is therefore

$$E=\frac{2}{\pi}=63{\cdot}66 \text{ per cent.};$$

from this value may be deduced the correlation, in large samples, between the mean and the median derived from the same sample

$$r=\sqrt{E}={\cdot}7979.$$

The median thus utilises about 64 per cent. of the information provided by the sample, its correlation with the mean of the same sample is about ·8, but any value obtained by combining the values of the median and the mean will result in an estimate inferior to that given by the mean.

A further consequence of this relation between the efficiency of a statistic and its correlation with any efficient statistic, is that if A be any efficient statistic, and B any inefficient statistic, then **the correlation of A with the difference B-A will tend to zero.** We may thus divide the error of B into two parts, which in large samples at least will be independent; the first part is equal to the error in A, and is the error of random sampling properly so called; the second $B-A$ is not properly speaking an error of random sampling, but an error of estimation. It is the property of efficient statistics that, when applied to large samples, they shall have no errors of estimation of order comparable with the errors of random sampling.

In all tests of significance an observed deviation is compared with the random sampling variation to be anticipated; in tests of goodness of fit in particular the "expectation" from which the deviations are measured is usually the product of a process of estimation, the basis of which is the actual sample of observations with which the expectation is to be compared. If, therefore, the process of estimation employed involves errors of the same order as the errors of random sampling the test of goodness of fit will be vitiated; the apparent discrepancy between observation and hypothesis will in fact involve errors of estimation of the same order as the errors of random sampling to which it is to be compared. The effects of such errors upon tests of goodness of fit have been shown in more detail in (3).

5. *Derivation of efficient statistics.*

To discover the efficiency of any statistic it is necessary that we should have found at least one statistic efficient for the estimation of the same parameter, and should know the variance in large samples of the latter. We shall see that the method of maximum likelihood will always provide a statistic which, if normally distributed in large samples with variance falling off inversely to the sample number, will be an efficient statistic. The variance in large samples of such solutions may be obtained directly from the equations by which they were obtained(1).

For example, if we have a number of observations drawn from a population, of which the distribution is given by

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

and wish from the observations to obtain an estimate of the value of m, we may write down in terms of m the actual probability of such a sample as ours occurring. This probability will be

$$\pi^{-n}\, dx_1 dx_2 \ldots dx_n \{1 + (x_1 - m)^2\}^{-1} \{1 + (x_2 - m)^2\}^{-1} \ldots \{1 + (x_n - m)^2\}^{-1}.$$

The likelihood of any value of m, in relation to such a sample, is defined as a quantity, of which the maximum value is unity, and which shall be proportional to the above probability. It is therefore independent of the elements $dx_1 \ldots dx_n$ which enter into the probability, but which do not involve m. *Likelihood* in this sense is not a synonym for probability, and is a quantity which does not obey the laws of probability; it is a property of the values of the parameters, which can be determined from the observations without antecedent knowledge. An exact knowledge of the likelihood of different values of m tells us nothing whatever about the probability that m will fall in any given range.

If we write for simplicity,

$$S(\log df) = L,$$

then

$$\frac{\partial L}{\partial m} = 0$$

is the equation of maximum likelihood, the solution of which gives an estimate of m, which we shall write $\hat{m}$. In the example before us this reduces to

$$S\left\{\frac{2(x-m)}{1+(x-m)^2}\right\} = 0.$$

To find the value of the variance of $\hat{m}$ derived from a large sample, it is only necessary to differentiate a second time;

$$\frac{\partial^2 L}{\partial m^2} = S\left\{\frac{2(x-m)^2-2}{\{1+(x-m)^2\}^2}\right\},$$

and for large samples the value of the right-hand side divided by n, tends to the limiting value $-\frac{1}{2}$. If $V(\hat{m})$ is the variance of $\hat{m}$, we therefore equate

$$-\frac{n}{2} = \frac{-1}{V(\hat{m})}$$

or

$$V(\hat{m}) = \frac{2}{n}.$$

Knowing this value it is easy to determine the efficiency of any other proposed statistic; in particular, since the equations of maximum likelihood do not always lend themselves to direct solution, it is of importance that, starting with an inefficient estimate, we can, by a single process of approximation, obtain an efficient estimate.

For example, if m_1, the median of the above distribution, be chosen as starting point, it is easy to show that the variance of m_1 in large samples is

$$\frac{\pi^2}{4n}$$

so that its efficiency is $8/\pi^2$. The median will differ from the maximum likelihood solution by errors of estimation, of which the variance will be

$$\frac{\pi^2-8}{4n}.$$

It is sufficient for our purpose that the error of estimation is of the order $n^{-\frac{1}{2}}$. If now we evaluate

$$S\left\{\frac{2(x-m_1)}{1+(x-m_1)^2}\right\}$$

from the observations, and calculate a new estimate m_2 from the equation

$$m_2 = m_1 + \frac{2}{n} S \left\{ \frac{2(x - m_1)}{1 + (x - m_1)^2} \right\},$$

it is easy to see that the error of estimation of m_2 will be of the order n^{-1}, and therefore that m_2 will be an efficient statistic.

6. *Intrinsic accuracy of error curves.*

The variance of efficient statistics from a distribution of any form affords us a measure of an important property of the distribution itself. The fact that from a large sample of n it is possible to obtain an estimate of the value of a parameter with variance $2/n$, shows that regarded as an error curve the above distribution is intrinsically of the same accuracy as, for example, a normal error curve with variance 2. We may thus obtain a measure of the intrinsic accuracy of an error curve, and so compare together curves of entirely different form. If the variance of an efficient estimate derived from a large sample of n is A/n, then the intrinsic accuracy of the distribution is defined as $1/A$.

If a frequency curve is defined by

$$df = y dx$$

where y is a function of a parameter θ, then the intrinsic accuracy of the curve, as a means of estimating θ, is

$$- \int y \frac{\partial^2}{\partial \theta^2} (\log y) . dx$$

over the whole range of possible values. Since

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

while the integral of the first term over all possible values must vanish, the intrinsic accuracy may equally be written

$$\int \frac{1}{y} \left(\frac{\partial y}{\partial \theta} \right)^2 dx,$$

over all possible values; in this form it is clearly seen to be necessarily positive.

What we have spoken of as the intrinsic accuracy of an error curve may equally be conceived as the amount of information in a single observation belonging to such a distribution. If for instance two independent observations were available from the same or different distributions, the distribution of the *pair* of values would be

$$df = yy' \, dx \, dx',$$

and the intrinsic accuracy of such a pair would be

$$-\iint yy'\left(\frac{\partial^2}{\partial\theta^2}\log y+\frac{\partial^2}{\partial\theta^2}\log y'\right)dx\,dx',$$

which, with the identities,

$$\int ydx=1,\quad \int y'dx'=1,$$

reduces to $\quad -\int y\,\frac{\partial^2}{\partial\theta^2}\log y\,.\,dx-\int y'\,\frac{\partial^2}{\partial\theta^2}\log y'\,.\,dx';$

the amount of information provided by a combination of two or more independent observations is thus merely the sum of the amounts of information in each piece separately.

It is a common case for a sample of n observations to be distributed into a finite number of classes, the numbers "expected" in each class being functions of one or more unknown parameters, if p is the probability of an observation falling into any one class, the amount of information in the sample is

$$S\left\{\frac{1}{m}\left(\frac{\partial m}{\partial\theta}\right)^2\right\}$$

where $m=np$, is the expectation in any one class. The variance of an efficient statistic derived from a large sample may, of course, be calculated from this expression.

7. *Efficiency of the maximum likelihood solution.*

We shall now prove that when an efficient statistic as defined above exists one may be found by the method of maximum likelihood.

If f stand for the probability that any particular type of observation should occur, and ϕ for the probability that any particular type of sample should occur, then

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

when C is a constant which does not involve the parameters, the summation extending over all observations.

As regards the variation of ϕ with varying θ, it is to be noted that

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

will tend to a fixed limit for large samples. Set $(-A)$ for this limiting value. Then since

$$\frac{\partial}{\partial\theta}\log\phi=0,$$

when $\theta = \hat{\theta}$, and $\hat{\theta}$ is the solution of the equations of maximum likelihood, it follows that

$$\frac{\partial}{\partial\theta} \log \phi = -nA(\theta - \hat{\theta})$$

if $\theta - \hat{\theta}$ is a small quantity of order $n^{-\frac{1}{2}}$.

Now if T is any statistic used as an estimate of θ, the probability, Φ, of T having any assigned value, will be the sum of the probabilities of those samples which yield the said value T, that is

$$\Phi = S(\phi)$$

when S stands for summation over all the samples which yield the same value for the statistic T. Also, if T is in large samples normally distributed with variance σ^2,

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

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

The problem of making σ^2 as small as possible, is the problem of so grouping the several sorts of samples under the same values of the estimate T, that the second differential coefficient of $\log \Phi$ shall be a negative quantity as large as possible. Now

$$\frac{1}{\sigma^2} = -\frac{\partial^2}{\partial\theta^2} \log \Phi = \frac{S^2(\phi')}{S^2(\phi)} - \frac{S(\phi'')}{S(\phi)}$$

and $S(\phi') \div S(\phi)$ is the mean value, within the group, of $-nA(\theta - \hat{\theta})$, while $S(\phi'') \div S(\phi)$ is the mean value of

$$-nA + n^2A^2(\theta - \hat{\theta})^2,$$

consequently
$$\frac{1}{n\sigma^2} = A - nA^2V'(\hat{\theta})$$

when $V'(\hat{\theta})$ is the variance of $\hat{\theta}$ within the group. If $T = \hat{\theta}$, then $\hat{\theta}$ will be constant within the group, so that the variance of $\hat{\theta}$ in random samples will be $1/nA$. For any statistic T which has the same value in sets of samples for which the variance of $\hat{\theta}$ is of order n^{-1} the value of $1/n\sigma^2$ will be reduced, for the variance within the group is necessarily a positive quantity, and consequently the variance of any such statistic will be greater than that of $\hat{\theta}$.

8. *Efficiency of weighting.*

The effects of the familiar process of weighting observations may be well shown in terms of efficiency. If w is the weight of a normally distributed observation x, so that its variance is $1/w$, and if a number of such observations be combined with false weights w', then the variance of the weighted mean will be

$$\frac{1}{S^2(w')}.S\left(\frac{w'^2}{w}\right),$$

and, when $w' = w$, this reduces to $1/S(w)$, the minimum value.

The efficiency is therefore

$$\frac{S^2(w')}{S(w).S\left(\frac{w'^2}{w}\right)}.$$

The loss of weight is

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

If now the inaccuracy in weighting is due to a slight variation in w, and we have chosen weights w' equal to the mean values of w, then

$$w = w' + \epsilon,$$

$$S\left(\frac{w'^2}{w}\right) = S(w') - S(\epsilon) + S\left(\frac{\epsilon^2}{w'}\right) - \dots,$$

$$S^2(w') \div S\left(\frac{w'^2}{w}\right) = S(w') + S(\epsilon) - S\left(\frac{\epsilon^2}{w'}\right) + \dots,$$

whence the mean loss of weight is, approximately

$$S\left(\frac{1}{w'}V(w)\right),$$

when $V(w)$ is the variance of w.

9. *Small samples; Sufficient statistics.*

It is now possible to approach the more general problem of the estimation of statistics from finite samples, when the distributions of the statistics considered will not generally be of the normal form, nor will the errors of random sampling be small quantities. The different possible efficient statistics will no longer be equivalent, and it will be necessary to discriminate among them. In previous work on this subject(1) two circumstances seemed to point to fruitful lines of development. In the first place attention was called to a class of statistics possessing very remarkable properties, which contain in themselves the whole of the relevant information available in the data. These statistics were therefore distinguished by the term sufficient. In the second place it was suggested that the idea of intrinsic accuracy might be applied to the random sampling distributions of statistics when these were not normal, so as to afford a method of comparing their relative values.

As an example of a sufficient statistic consider the mean of the Poisson Series. A variate, confined to whole numbers, is distributed in a Poisson Series if the probability of its taking any particular value, x, is

$$e^{-m} \cdot \frac{m^x}{x!}.$$

The parameter m may be estimated from the mean of the observed sample. This is evidently the solution of the equation of maximum likelihood. If $\bar{x}$ is the mean of a sample of n, the distribution of $n\bar{x}$ may readily be proved to be given by the Poisson Series

$$e^{-nm} \frac{(nm)^{n\bar{x}}}{(n\bar{x})!}.$$

Now the probability of drawing in order any particular sample $x_1, x_2, \ldots, x_n$ is

$$e^{-nm} \frac{m^{n\bar{x}}}{x_1!\, x_2! \ldots x_n!},$$

and this may be divided into two factors,

$$e^{-nm} \frac{(nm)^{n\bar{x}}}{(n\bar{x})!} \cdot \frac{(n\bar{x})!}{n^{n\bar{x}}\, x_1!\, x_2! \ldots x_n!},$$

of which the first represents the probability that the actual total $n\bar{x}$ should have been scored, and the second the probability, given this total, that the partition of it among the n observations should be that actually observed. In the latter factor, m, the parameter sought, does not appear. Now when the mean is known any further information which the sample has to give must depend on the observed partition; but the probability of any particular partition is wholly independent of the value of m. Consequently no statistic calculated from the sample can give any information whatever respecting the value of m, beyond that supplied by the value of the mean.

In general, if θ is any parameter, T_1 a statistic sufficient 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 $f(\theta, T_1, T_2)\, dT_1\, dT_2$ be the probability that T_1 and T_2 should fall in the ranges dT_1, dT_2, and if

$$f(\theta, T_1, T_2) = \phi(\theta, T_1) \cdot \phi'(T_1, T_2).$$

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. For example the mean of a normal sample is a sufficient statistic, and the mean possesses the property that it can be combined with the mean of a second sample from the same population to find the mean of the combined sample. If $f(x_1, \ldots, x_n)$ be for any distribution a sufficient estimate of some parameter, then

$$f(x_1, \ldots, x_n) = \phi \{f(x_1, \ldots, x_p), f(x_{p+1}, \ldots, x_n)\};$$

this circumstance much limits the functions which can possibly be sufficient statistics.

For example, the function

$$\frac{1}{k} \log S(e^{kx}) - \frac{1}{k} \log n,$$

is a function which might, for the right distribution, be a sufficient statistic. As k is made to increase without limit the above function tends to be simply the greatest value observed in the sample; just as the mean of a number of means is the mean of the aggregate, so the greatest of a series of greatest observations, will be the greatest of the aggregate.

When sufficient statistics exist it has been shown that they will be solutions of the equations of maximum likelihood.

10. *Intrinsic accuracy of error curves of statistics.*

The fact that sufficient statistics do not always exist renders it necessary to explore the possibilities of comparing statistics by means of the intrinsic accuracy of their random sampling distributions.

We may, in fact, give an extended meaning to the word efficiency by the definition

The efficiency of a statistic is the ratio of the intrinsic accuracy of its random sampling distribution to the amount of information in the data from which it has been derived.

This definition is in accordance with the definition previously given of efficiency for the case of large samples with normally distributed statistics. For if $\frac{\sigma^2}{n}$ is the variance in large samples of an efficient statistic, the intrinsic accuracy of the original distribution will be $1/\sigma^2$, and the whole information in the data will be n/σ^2. Moreover if in large samples any statistic has variance σ^2/En, its intrinsic accuracy will be En/σ^2, and its efficiency, by either definition, will be E.

The extended definition has the advantage of applying to finite samples and to other cases where the distribution is not normal.

As an example of the calculation of the efficiency of statistics derived from finite samples, consider the median of an odd number, $2s + 1$, of observations the error curve of which is given by

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

It is easy to see that the frequency distribution of the median will be given by

$$df = \frac{(2s + 1)!}{(s!)^2 \, \pi^{2s+1}} \left(\frac{\pi^2}{4} - \theta^2\right)^s \frac{dx}{1 + (x - m)^2},$$

where $\tan \theta = x - m$, and θ lies between $\pm \frac{1}{2}\pi$.

To find the intrinsic accuracy of the distribution, we differentiate the logarithm of the above expression with respect to the unknown parameter, m; then since

$$\frac{d\theta}{dm} = - \cos^2\theta$$

we have

$$\frac{2s\theta \cos^2 \theta}{\dfrac{\pi^2}{4} - \theta^2} + \sin 2\theta.$$

The intrinsic accuracy will be the average value of the square of this quantity, or

$$\frac{(2s + 1)!}{(s!)^2 \, \pi^{2s+1}} \int_{-\frac{\pi}{2}}^{+\frac{\pi}{2}} \left\{2s\theta \cos^2 \theta + \left(\frac{\pi^2}{4} - \theta^2\right) \sin 2\theta\right\}^2 \left(\frac{\pi^2}{4} - \theta^2\right)^{s-2} d\theta.$$

The definite integral may be expressed in terms of the Bessel functions of π and 2π, in the form

$$\frac{1}{2} + \frac{3s\,(2s + 1)}{2\,(s - 1)\,\pi^2} + \frac{(s + \frac{1}{2})!}{2s - 1} \cdot \left(\frac{2}{\pi}\right)^{s+\frac{1}{2}} \left\{\frac{2s}{s - 1} J_{s-\frac{3}{2}}(\pi) - 2J_{s+\frac{1}{2}}(\pi)\right\}$$
$$\frac{(s + \frac{1}{2})!}{2s - 1} \cdot \left(\frac{1}{\pi}\right)^{s+\frac{1}{2}} \left\{\frac{2s}{s - 1} J_{s-\frac{3}{2}}(2\pi) - \frac{2s + 3}{2} J_{s+\frac{1}{2}}(2\pi)\right\}.$$

The Bessel functions are easily evaluated, for $J_{\frac{1}{2}} = 0$, for both values of the argument; while the values of $J_{\frac{3}{2}}$ are $\sqrt{2}/\pi$ and $- 1/\pi$, the others being thence obtainable by the recurrence formula

$$J_{n+1} = \frac{2n}{z} J_n - J_{n-1}.$$

Thus for $s = 2$, we find the intrinsic accuracy of the median of five observations to be

$$\frac{1}{2} + \frac{15}{\pi^2} - \frac{855}{32\pi^4}.$$

The following table shows the numerical values.

Number in sample	s	Accuracy	Increase for two observations	Efficiency
1	0	·50000	·59064	100·00 %
3	1	1·09064	·65488	72·71 %
5	2	1·74552	·69490	69·82 %
7	3	2·44042	·72121	69·73 %
9	4	3·16164	·73945	70·26 %
11	5	3·90109		70·93

The efficiency appears to have a minimum between 5 and 7 observations, and is not approaching its limiting value, 81·06 per cent., very rapidly. This is perhaps to be anticipated since for large samples it falls short of its limiting efficiency by $1{\cdot}19/s$, and the discrepancy in the above table is considerably less than this.

Statistics which are efficient for large samples may, of course, have comparatively low efficiencies for finite samples, and in certain cases the efficiency may tend to its limiting value so slowly that even samples of over 100 are not very efficiently treated. The median, for example, is an efficient statistic for locating the centre of the double exponential curve,

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

when the sample is increased without limit, but owing to the discontinuity at the apex, its efficiency approaches its limiting value somewhat slowly. The intrinsic accuracy of the original distribution is unity, and that of the median of $(2s+1)$ observations may be shown to be

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

The numerical values are:

Number in sample	s	Accuracy	Efficiency %	Loss of information
1	0	1	100	0
3	1	2·3178	77·26	·6822
5	2	3·7500	75·00	1·2500
7	3	5·2500	75·00	1·7500
9	4	6·7969	75·52	2·2031
19	9	14·940	78·63	4·060
33	16	26·932	81·61	6·068
51	25	42·844	84·01	8·156
73	36	62·709	85·90	10·291
99	49	86·544	87·42	12·456
129	64	114·36	88·65	14·64
163	81	146·16	89·67	16·84
201	100	181·95	90·52	19·05

The case is unusual in that the loss of information does not tend to a fixed limit, but increases ultimately as $4\sqrt{s/\pi} - 4$; the cause of this exceptional behaviour lies in the fact that at the apex of the curve the second differential coefficient with respect to the unknown, m, is infinite. The example stresses the importance of investigating the actual behaviour of statistics from finite samples, instead of relying wholly upon their calculated behaviour in infinitely large samples.

We can now prove in general that the efficiency can never exceed unity, and derive the condition that there shall be no loss of information.

If the probability that any statistic, T, should take a particular value is Φ, then the intrinsic accuracy of the distribution of T is

$$S\left\{\frac{\Phi'^2}{\Phi}\right\},$$

the summation being taken over all possible values. If now every possible sample, having frequency ϕ gave a different value of T, then the intrinsic accuracy would be

$$S\left\{\frac{\phi'^2}{\phi}\right\}$$

and would be independent of T. If, however, a number of different samples give the same value of T, then the effect of this amalgamation will be to decrease the intrinsic accuracy by the amount

$$S\left(\frac{\phi'^2}{\phi}\right) - \frac{\Phi'^2}{\Phi},$$

$$= \quad S\left\{\phi\left(\frac{\phi'}{\phi} - \frac{\Phi'}{\Phi}\right)^2\right\}.$$

This quantity is never negative, so that the intrinsic accuracy of T can never be greater than when every possible sample yields a different value of T. This is obvious because, in such a case, the actual sample can be reconstructed without ambiguity from the value of T, and so the value of T, which is merely a kind of shorthand statement of the original sample, must contain the whole of the information provided by the sample.

The condition that there shall be no loss of information when different samples give the same value of T is that the sets of possible samples for which T is constant shall be those for which

$$\frac{\phi'}{\phi} = \frac{\partial L}{\partial \theta}$$

is constant. If these sets are the same for all values of θ, then the equation of maximum likelihood

$$\frac{\partial L}{\partial \theta} = 0$$

will provide a sufficient statistic.

For if this is the case $\partial L/\partial\theta$ depends, apart from θ, only on the set to which the sample belongs; in other words it is a function of θ and $\hat{\theta}$ only. Thus if f is the frequency with which any sample, or group of samples having the same $\hat{\theta}$, occurs, then

$$\frac{1}{f}\cdot\frac{\partial f}{\partial \theta} = \phi\,(\theta, \hat{\theta}),$$

now let the frequency of samples such that $\hat{\theta}$ lies in the range $d\hat{\theta}$ and a second statistic, T, lies in the range dT, be $f\,(\theta, \hat{\theta}, T)\, d\hat{\theta}\, dT$, then since the above equation will be true for all values of θ, we shall integrate it with respect to θ and obtain

$$\log f = \int \phi\,(\theta, \hat{\theta})\, d\theta + C$$

where C does not involve θ, and is a function therefore of $\hat{\theta}$ and T only. Hence f is of the form

$$\phi\,(\theta, \hat{\theta}) \,.\, \phi'\,(\hat{\theta}, T)$$

whatever statistic may be taken as T, and so $\hat{\theta}$ must be a sufficient statistic.

11. *Minimal loss of accuracy.*

When the sets of samples which for one value of θ have the same value of $\partial L/\partial\theta$, have no longer the same value for other values of θ, there exists no sufficient statistic, and some loss of information will necessarily ensue upon the substitution of a single estimate for the original data upon which it was based.

The extent of this loss, in large samples, for which presumably it will be greatest, may now be calculated. If the sample consist of observed numbers $x_1, \ldots\, x_s$ in categories in which the expectations are $m_1, \ldots, m_s$, then

$$L = S\,(x \log m),$$

$$\partial L/\partial\theta = S\left(x\,\frac{m'}{m}\right),$$

$$\partial^2 L/\partial\theta^2 = S\left\{x\left(\frac{m''}{m} - \frac{m'^2}{m^2}\right)\right\},$$

if now $\partial L/\partial\hat{\theta} = 0$, then to a first approximation

$$\frac{\partial L}{\partial \theta} = (\theta - \hat{\theta})\,\frac{\partial^2 L}{\partial \theta^2},$$

and the variance of $\partial L/\partial\theta$ in a set of samples for which $\hat{\theta}$ is constant, will be given by the variance of $\partial^2 L/\partial\theta^2$ within the set multiplied by $(\theta - \hat{\theta})^2$, or the total loss of information will be given by the general variance within such sets multiplied by $V(\hat{\theta})$.

Now the random sampling distribution of the values of x will be the multinomial distribution, and the simplest method of regarding this distribution is to consider each value of x independently distributed in a Poisson Series about a mean value m; the whole being subject to the restriction that

$$S(x) = S(m) = n.$$

In such a system any quantity $S(kx)$ is easily seen to have a mean value $S(km)$; its variance, if there were no restriction, would be $S(k^2 m)$, and in introducing any linear restriction we have only to remove that portion of the variance produced by varying in the prohibited direction. Two restrictions are here necessary, for the first

$$S(x) = n$$

we have to deduct $\qquad S^2(km) \div n.$

The second restriction arises from the fact that we require the variance within the groups for which $\partial L/\partial\hat{\theta}$ is constant, since

$$\frac{\partial L}{\partial \theta} = S\left(x . \frac{m'}{m}\right),$$

the deduction will be $\quad S^2(km') \div S\left(\dfrac{m'^2}{m}\right).$

Writing $\qquad k = \dfrac{m''}{m} - \dfrac{m'^2}{m^2},$

and remembering that $V(\hat{\theta}) = 1/S\left(\dfrac{m'^2}{m}\right)$, we have for the loss of information in large samples,

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

The method of calculation by differences has the advantage that if, by the estimation of other parameters, further restrictions upon variation are introduced, we may choose such parameters that their maximum likelihood estimates will be in large samples uncorrelated with each other and with $\hat{\theta}$, and the whole effect of the restrictions will be further to reduce the loss of accuracy by terms of the form

$$\frac{S^2\left\{\frac{1}{m}\frac{\partial m}{\partial \phi}\left(m''-\frac{m'^2}{m}\right)\right\}}{S\left\{\frac{1}{m}\left(\frac{\partial m}{\partial \phi}\right)^2\right\}.S\left(\frac{m'^2}{m}\right)},$$

without further examination of the restrictions.

12. *Example of loss of intrinsic accuracy.*

In the curve $$df=\frac{1}{\pi}.\frac{dx}{1+(x-\theta)^2}$$

we may write $$m=\frac{n}{\pi}.\frac{dx}{1+(x-\theta)^2};$$

then for the determination of θ,

$$S\left(\frac{m'^2}{m}\right)=\frac{n}{\pi}\int_{-\infty}^{\infty}\frac{4t^2\,dt}{(1+t^2)^3}=\frac{n}{2},$$

$$S\left\{\frac{1}{m}\left(m''-\frac{m'^2}{m}\right)^2\right\}=\frac{n}{\pi}\int_{-\infty}^{\infty}\frac{4\,(t^2-1)^2}{(1+t^2)^5}\,dt=\frac{7n}{8},$$

$$S\left\{\frac{m'}{m}\left(m''-\frac{m'^2}{m}\right)\right\}=0.$$

The loss of information is therefore

$$\tfrac{7}{4}-\tfrac{1}{2}=\tfrac{5}{4},$$

and since the intrinsic accuracy of the original distribution is $\frac{1}{2}$, the loss on statistical reduction is equivalent in large samples to $2\frac{1}{2}$ observations. For small samples the loss will presumably be less since it vanishes for samples of one.

In the location of the centre of this curve, therefore, we see that the mean is a statistic which throws away the greater part of the information available; the median is an inefficient statistic which makes use of a fraction, approaching in large samples the limit $8/\pi^2$ of the information. The solution of the equation of maximum likelihood, like other efficient statistics makes use of all but an amount which tends to a finite limit as the sample is increased. The amount lost may differ in different efficient statistics; it will be least for the solution of the equation of maximum likelihood.

13. *Loss of accuracy with other statistics.*

With efficient statistics other than the maximum likelihood solution, the loss of accuracy will be somewhat greater, though still tending to a finite limit for large samples. The variance of $\partial L/\partial\theta$ for sets of samples yielding the same statistic will be due to two independent causes. The first depends upon the sampling error, $T-\theta$, and upon the fact that $\partial^2 L/\partial\theta^2$ is not the same for all samples yielding the same statistic. Since all efficient statistics tend to equivalence with increasing samples, this portion will be the same for all efficient statistics. The second portion is approximately independent of the sampling error, and depends upon the deviation of T from the maximum likelihood solution $\hat{\theta}$.

For example, if $\chi^2 = S\left\{\dfrac{(x-m)^2}{m}\right\}$,

the equation for minimum χ^2,

$$S\left(\frac{x^2-m^2}{m^2}\,\frac{\partial m}{\partial\theta}\right)=0,$$

gives an efficient estimate of θ; for the maximum likelihood equation is

$$S\left(\frac{x-m}{m}\,\frac{\partial m}{\partial\theta}\right)=0,$$

and the ratio $\dfrac{x+m}{m}$ tends to the constant value, 2, for large samples. Now, since

$$(x^2-m^2)=2m\,(x-m)+(x-m)^2,$$

the deviation in $\partial L/\partial\theta$ will be

$$\tfrac{1}{2}S\left\{\frac{(x-m)^2}{m^2}\,\frac{\partial m}{\partial\theta}\right\},$$

and it is the variance of this quantity for samples of equal sampling error which gives the loss of information. By the same device as before we evaluate the variance of $S\,\{k\,(x-m)^2\}$ in the form

* $$2S\,(k^2m^2)-\frac{2}{n}S^2\,(km^2)\underline{-2\,\frac{S^2\,(km'^2)}{S^2\left(\dfrac{m'^2}{m}\right)}},$$

or, substituting $k=\dfrac{1}{2}\,\dfrac{m'}{m^2}$,

* we have $$\tfrac{1}{2}S\left(\frac{m'^2}{m^2}\right)-\frac{1}{2}\,\underline{\frac{S^2\left(\dfrac{m'^3}{m^2}\right)}{S^2\left(\dfrac{m'^2}{m}\right)}}.$$

* Delete the underlined terms.

This quantity remains finite as the size of the sample is increased without limit, but increases without limit as the number of classes is increased. Consequently, as one might expect, the method of minimising χ^2 breaks down for fine grouping.

For example, suppose we have 5 classes only, in a population distributed according to the binomial distribution

$$(p+q)^4.$$

Let p be calculated from numbers observed in the 5 possible classes in a large sample. Then

$$m_1 = np^4, \quad m_2 = 4np^3q, \text{ etc.},$$

and the intrinsic accuracy is

$$S\left(\frac{m'^2}{m}\right) = \frac{4n}{pq}.$$

The loss of information due to using the minimum χ^2 solution is calculated from

$$\tfrac{1}{2}S\left(\frac{m'^2}{m^2}\right) = \frac{5}{p^2q^2}(3p^2 - 2pq + 3q^2)$$

and

* $$S\left(\frac{m'^3}{m^2}\right) = -\frac{4(p-q)}{p^2q^2}(p^4 - 2p^3q + 18p^2q^2 - 2pq^3 + q^4)\,n,$$

and is in fact

† $$\frac{5}{p^2q^2}(3p^2 - 2pq + 3q^2) \underline{- \frac{(p-q)^2}{2p^2q^2}(p^4 - 2p^3q + 18p^2q^2 - 2pq^3 + q^4)^2}.$$

This is least when $p = q$, and is then equal to

$$\frac{20}{pq},$$

or equivalent to the loss of 5 observations.

In approaching the maximum likelihood solution by successive approximations we have seen that starting with an inefficient statistic, a single process of approximation will in ordinary cases give an efficient statistic differing from the maximum likelihood solution, by a quantity which with increasing samples decreases as n^{-1}. The loss of information of such efficient statistics is therefore finite for large samples, for the additional variance of $\partial L/\partial\theta$ will be

$$L''^2\,V(T - \hat{\theta}),$$

and L'' increases proportionately to the sample. If the process of approximation be repeated a statistic is obtained differing from the maximum likelihood solution only to the order of $n^{-\frac{3}{2}}$, and for such a statistic the loss of accuracy, beyond that suffered by the maximum likelihood solution will tend to zero for large samples.

* Delete this line.

† Delete the underlined expression.

The practical procedure of fitting will thus not ordinarily require more than a second process of approximation.

14. *Theoretical existence of fully accurate statistics.*

In the manner in which we have developed the theory it would appear that the loss of information inherent in the process of replacing a quantity of observational material by a single value, arose from the circumstance that the groups of samples, which ought to give us the same estimate, change with the value of the unknown parameter, and so that no way of grouping can be the best for all values of the parameter. The method of maximum likelihood takes that way of grouping which is most accurate for the particular value which is equal to the estimate arrived at. The loss of information with which we are concerned is the difference in accuracy between the solution of the equations of maximum likelihood, and another statistic which might conceivably be arrived at by chance, but which cannot be specified without knowledge of the true value of the parameter.

If, for example, the quantity θ in the following expression *happened* to be equal to the value of m in the population specified by

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

and we used the equation of estimation

$$S\left\{\frac{2(x-\theta)}{1+(x-\theta)^2}\right\} = nf(T),$$

where it will be noticed that the left-hand side is $\partial L/\partial\theta$, then $\partial L/\partial\theta$ is constant for samples giving the same T; the form of f can be ascertained from the condition of consistency, for the equation will only be consistent if

$$f(m) = \frac{1}{\pi}\int_{-\infty}^{\infty}\frac{2(x-\theta)}{1+(x-\theta)^2}\cdot\frac{dx}{1+(x-m)^2} = \frac{2(m-\theta)}{(m-\theta)^2+4}.$$

The equation for T will therefore be

$$S\left\{\frac{2(x-\theta)}{1+(x-\theta)^2}\right\} = \frac{2n(T-\theta)}{(T-\theta)^2+4};$$

this equation is nearly equivalent to the equation we have given for improving an approximate value, in this case θ. If, however, θ were in fact equal to the true value of the parameter, the statistic T would be distributed in random samples with an intrinsic accuracy equal to the maximum possible. We cannot, however, utilise this fact, for if we could rely upon our putting the right value of θ into this equation, we could choose θ as our estimate and so avoid errors of random sampling altogether.

15. *Use of ancillary statistics.*

Since 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.

If our two statistics specify the values of $\partial L/\partial\theta$ and $\partial^2 L/\partial^2\theta$ for some central value of θ, such as $\hat{\theta}$, then the variance of $\partial L/\partial\theta$ over the sets of samples for which both statistics are constant, will be that of

$$\tfrac{1}{2}(\theta-\hat{\theta})^2\frac{\partial^3 L}{\partial\theta^3},$$

which will ordinarily be of order n^{-1} at least. With the aid of such an ancillary statistic the loss of accuracy tends to zero for large samples.

The function of the ancillary statistic is analogous to providing a true, in place of an approximate, weight for the value of the estimate. If a number of large samples were available, and if

$$M = S(L)$$

when the summation is taken over all the samples, then M will be the logarithm of the likelihood of θ from the combined sample; but necessarily

$$M' = S(L'),\quad M'' = S(L'')$$

and if $\hat{\theta}$ be the value of θ for which M' vanishes, and $\hat{\theta}_p$ the value for which L_p' vanishes, then with large samples, when $\theta = \hat{\theta}$,

$$L_p' = (\hat{\theta}-\hat{\theta}_p)\,L_p''.$$

Hence $\hat{\theta}$ is given by the equation

$$S\{L''(\hat{\theta}-\hat{\theta}_p)\} = 0,$$

or

$$M''\hat{\theta} = S(L_p''\hat{\theta}_p),$$

where

$$M'' = S(L'').$$

If we had ignored the ancillary statistic and taken as weights the mean value

$$-\overline{L''} = +\frac{1}{V(\hat{\theta})},$$

the loss of weight in the combined value,

$$S\left(\frac{1}{w'}V(w)\right),$$

is the sum of the contributions to the loss of weight from the several samples. Each contribution is equal to the sampling variance of L'' multiplied by $V(\hat{\theta})$, and this is just the quantity we have found as measuring the loss of information.

REFERENCES.

(1) Fisher, R. A. (1921). "The mathematical foundations of theoretical statistics." *Phil. Trans.* A., vol. 222, pp. 309–368.

(2) —— (1920). "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.* vol. 80, pp. 758–770.

(3) —— (1924). "The conditions under which χ^2 measures the discrepancy between observation and hypothesis." *J.R.S.S.* vol. 87, pp. 442–450.