Reproduced from the Philosophical Transactions of the Royal Society, A, 222: 309-368 (1922)
with permission of the Royal Society

18

ON THE MATHEMATICAL FOUNDATIONS OF THEORETICAL STATISTICS

Author's Note (CMS 10.308a) *

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

This is the first large-scale attack on the problem of estimation. In the author's opinion the frequently stated view that the concept of a best estimate was arbitrary and subjective ignored the guidance afforded by purely mathematical considerations of absolute validity. He had been impressed by the property of sufficiency, first found in 1920 (Paper 12).

This property, when it exists, picks out one particular method of estimation as uniquely superior to all possible alternatives. In such cases the sufficient estimate may be found by the method of maximum likelihood. Consequently, one purpose of the paper is to examine the properties of the likelihood function, here defined, and the properties of the estimates arrived at by maximising this function. Several fruitful ideas, such as the measurement of the amount of information, emerge. These were classified and developed further in 1925 (Paper 42).

On these further points the present paper is only of historical interest, for the author was by no means clear on such points as whether a sufficient statistic, or something with equivalent advantages, could always exist, and the extension of the theory towards an exact treatment of small samples is consequently incomplete. He did not clearly see, for example, that the variance of an estimate does not, in the theory of small samples, supply a satisfactory basis for comparison. The correct treatment is, however, foreshadowed on page 350.

His object was to test the value of the new ideas by applying them to a variety of problems. An older or more judicious writer would not have allowed the course of the argument to be interrupted by such long excursuses as that giving the exact treatment of corrections for grouping; or that in which the efficiency of fitting Pearsonian curves by moments is examined. These were, however, questions for which at the time no analytically competent discussion existed. After all, it is a common weakness of young authors to put too much into their papers.

Philosophical Transactions of the Royal Society of London, A, 222: 309-368, (1922).

IX. *On the Mathematical Foundations of Theoretical Statistics.*

By R. A. Fisher, M.A., *Fellow of Gonville and Caius College, Cambridge, Chief Statistician, Rothamsted Experimental Station, Harpenden.*

Communicated by Dr. E. J. Russell, F.R.S.

Received June 25,—Read November 17, 1921.

Contents.

Definitions.

Centre of Location.—That abscissa of a frequency curve for which the sampling errors of optimum location are uncorrelated with those of optimum scaling. (9.)

Consistency.—A statistic satisfies the criterion of consistency, if, when it is calculated from the whole population, it is equal to the required parameter. (4.)

Distribution.—Problems of distribution are those in which it is required to calculate the distribution of one, or the simultaneous distribution of a number, of functions of quantities distributed in a known manner. (3.)

Efficiency.—The efficiency of a statistic is the ratio (usually expressed as a percentage) which its intrinsic accuracy bears to that of the most efficient statistic possible. It expresses the proportion of the total available relevant information of which that statistic makes use. (4 and 10.)

Efficiency (Criterion).—The criterion of efficiency is satisfied by those statistics which, when derived from large samples, tend to a normal distribution with the least possible standard deviation. (4.)

Estimation.—Problems of estimation are those in which it is required to estimate the value of one or more of the population parameters from a random sample of the population. (3.)

Intrinsic Accuracy.—The intrinsic accuracy of an error curve is the weight in large samples, divided by the number in the sample, of that statistic of location which satisfies * the criterion of sufficiency. (9.)

Isostatistical Regions.—If each sample be represented in a generalized space of which the observations are the co-ordinates, then any region throughout which any set of statistics have identical values is termed an isostatistical region.

Likelihood.—The likelihood that any parameter (or set of parameters) should have any assigned value (or set of values) is proportional to the probability that if this were so, the totality of observations should be that observed.

Location.—The location of a frequency distribution of known form and scale is the process of estimation of its position with respect to each of the several variates. (8.)

Optimum.—The optimum value of any parameter (or set of parameters) is that value (or set of values) of which the likelihood is greatest. (6.)

Scaling.—The scaling of a frequency distribution of known form is the process of estimation of the magnitudes of the deviations of each of the several variates. (8.)

Specification.—Problems of specification are those in which it is required to specify the mathematical form of the distribution of the hypothetical population from which a sample is to be regarded as drawn. (3.)

Sufficiency.—A statistic satisfies the criterion of sufficiency when no other statistic which can be calculated from the same sample provides any additional information as to the value of the parameter to be estimated. (4.)

Validity.—The region of validity of a statistic is the region comprised within its contour of zero efficiency. (10.)

1. The Neglect of Theoretical Statistics.

Several reasons have contributed to the prolonged neglect into which the study of statistics, in its theoretical aspects, has fallen. In spite of the immense amount of fruitful labour which has been expended in its practical applications, the basic principles of this organ of science are still in a state of obscurity, and it cannot be denied that, during the recent rapid development of practical methods, fundamental problems have been ignored and fundamental paradoxes left unresolved. This anomalous state of statistical science is strikingly exemplified by a recent paper (1) entitled "The Funda-

* For sufficiency, read efficiency.

mental Problem of Practical Statistics," in which one of the most eminent of modern statisticians presents what purports to be a general proof of BAYES' postulate, a proof which, in the opinion of a second statistician of equal eminence, " seems to rest upon a very peculiar—not to say hardly supposable—relation." (2.)

Leaving aside the specific question here cited, to which we shall recur, the obscurity which envelops the theoretical bases of statistical methods may perhaps be ascribed to two considerations. In the first place, it appears to be widely thought, or rather felt, that in a subject in which all results are liable to greater or smaller errors, precise definition of ideas or concepts is, if not impossible, at least not a practical necessity. In the second place, it has happened that in statistics a purely verbal confusion has hindered the distinct formulation of statistical problems ; for it is customary to apply the same name, *mean, standard deviation, correlation coefficient*, etc., both to the true value which we should like to know, but can only estimate, and to the particular value at which we happen to arrive by our methods of estimation ; so also in applying the term probable error, writers sometimes would appear to suggest that the former quantity, and not merely the latter, is subject to error.

It is this last confusion, in the writer's opinion, more than any other, which has led to the survival to the present day of the fundamental paradox of inverse probability, which like an impenetrable jungle arrests progress towards precision of statistical concepts. The criticisms of BOOLE, VENN, and CHRYSTAL have done something towards banishing the method, at least from the elementary text-books of Algebra ; but though we may agree wholly with CHRYSTAL that inverse probability is a mistake (perhaps the only mistake to which the mathematical world has so deeply committed itself), there yet remains the feeling that such a mistake would not have captivated the minds of LAPLACE and POISSON if there had been nothing in it but error.

2. THE PURPOSE OF STATISTICAL METHODS.

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

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

When we speak of the *probability* of a certain object fulfilling a certain condition, we imagine all such objects to be divided into two classes, according as they do or do not fulfil the condition. This is the only characteristic in them of which we take cognisance. For this reason probability is the most elementary of statistical concepts. It is a parameter which specifies a simple dichotomy in an infinite hypothetical population, and it represents neither more nor less than the frequency ratio which we imagine such a population to exhibit. For example, when we say that the probability of throwing a five with a die is one-sixth, we must not be taken to mean that of any six throws with that die one and one only will necessarily be a five; or that of any six million throws, exactly one million will be fives; but that of a hypothetical population of an infinite number of throws, with the die in its original condition, exactly one-sixth will be fives. Our statement will not then contain any false assumption about the actual die, as that it will not wear out with continued use, or any notion of approximation, as in estimating the probability from a finite sample, although this notion may be logically developed once the meaning of probability is apprehended.

The concept of a *discontinuous frequency distribution* is merely an extension of that of a simple dichotomy, for though the number of classes into which the population is divided may be infinite, yet the frequency in each class bears a finite ratio to that of the whole population. In *frequency curves*, however, a second infinity is introduced. No finite sample has a frequency curve: a finite sample may be represented by a histogram, or by a frequency polygon, which to the eye more and more resembles a curve, as the size of the sample is increased. To reach a true curve, not only would an infinite number of individuals have to be placed in each class, but the number of classes (arrays) into which the population is divided must be made infinite. Consequently, it should be clear that the concept of a frequency curve includes that of a hypothetical infinite population, distributed according to a mathematical law, represented by the curve. This law is specified by assigning to each element of the abscissa the corresponding element of probability. Thus, in the case of the normal distribution, the probability of an observation falling in the range dx, is

$$\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}dx,$$

in which expression x is the value of the variate, while m, the mean, and σ, the standard deviation, are the two parameters by which the hypothetical population is specified. If a sample of n be taken from such a population, the data comprise n independent facts. The statistical process of the reduction of these data is designed to extract from them all relevant information respecting the values of m and σ, and to reject all other information as irrelevant.

It should be noted that there is no falsehood in interpreting any set of independent measurements as a random sample from an infinite population; foı any such set of numbers are a random sample from the totality of numbers produced by the same matrix of causal conditions: the hypothetical population which we are studying is an aspect of the totality of the effects of these conditions, of whatever nature they may be. The postulate of randomness thus resolves itself into the question, "Of what population is this a random sample?" which must frequently be asked by every practical statistician.

It will be seen from the above examples that the process of the reduction of data is, even in the simplest cases, performed by interpreting the available observations as a sample from a hypothetical infinite population; this is *a fortiori* the case when we have more than one variate, as when we are seeking the values of coefficients of correlation. There is one point, however, which may be briefly mentioned here in advance, as it has been the cause of some confusion. In the example of the frequency curve mentioned above, we took it for granted that the values of both the mean and the standard deviation of the population were relevant to the inquiry. This is often the case, but it sometimes happens that only one of these quantities, for example the standard deviation, is required for discussion. In the same way an infinite normal population of two correlated variates will usually require five parameters for its specification, the two means, the two standard deviations, and the correlation; of these often only the correlation is required, or if not alone of interest, it is discussed without reference to the other four quantities. In such cases an alteration has been made in what is, and what is not, relevant, and it is not surprising that certain small corrections should appear, or not, according as the other parameters of the hypothetical surface are or are not deemed relevant. Even more clearly is this discrepancy shown when, as in the treatment of such fourfold tables as exhibit the recovery from smallpox of vaccinated and unvaccinated patients, the method of one school of statisticians treats the proportion of vaccinated as relevant, while others dismiss it as irrelevant to the inquiry. (3.)

3. The Problems of Statistics.

The problems which arise in reduction of data may be conveniently divided into three types:—

(1) Problems of Specification. These arise in the choice of the mathematical form of the population.

(2) Problems of Estimation. These involve the choice of methods of calculating from a sample statistical derivates, or as we shall call them statistics, which are designed to estimate the values of the parameters of the hypothetical population.

(3) Problems of Distribution. These include discussions of the distribution of statistics derived from samples, or in general any functions of quantities whose distribution is known.

It will be clear that when we know (1) what parameters are required to specify the population from which the sample is drawn, (2) how best to calculate from the sample estimates of these parameters, and (3) the exact form of the distribution, in different samples, of our derived statistics, then the theoretical aspect of the treatment of any particular body of data has been completely elucidated.

As regards problems of specification, these are entirely a matter for the practical statistician, for those cases where the qualitative nature of the hypothetical population is known do not involve any problems of this type. In other cases we may know by experience what forms are likely to be suitable, and the adequacy of our choice may be tested *a posteriori*. We must confine ourselves to those forms which we know how to handle, or for which any tables which may be necessary have been constructed. More or less elaborate forms will be suitable according to the volume of the data. Evidently these are considerations the nature of which may change greatly during the work of a single generation. We may instance the development by PEARSON of a very extensive system of skew curves, the elaboration of a method of calculating their parameters, and the preparation of the necessary tables, a body of work which has enormously extended the power of modern statistical practice, and which has been, by pertinacity and inspiration alike, practically the work of a single man. Nor is the introduction of the Pearsonian system of frequency curves the only contribution which their author has made to the solution of problems of specification: of even greater importance is the introduction of an objective criterion of goodness of fit. For empirical as the specification of the hypothetical population may be, this empiricism is cleared of its dangers if we can apply a rigorous and objective test of the adequacy with which the proposed population represents the whole of the available facts. Once a statistic, suitable for applying such a test, has been chosen, the exact form of its distribution in random samples must be investigated, in order that we may evaluate the probability that a worse fit should be obtained from a random sample of a population of the type considered. The possibility of developing complete and self-contained tests of goodness of fit deserves very careful consideration, since therein lies our justification for the free use which is made of empirical frequency formulæ. Problems of distribution of great mathematical difficulty have to be faced in this direction.

Although problems of estimation and of distribution may be studied separately, they are intimately related in the development of statistical methods. Logically problems of distribution should have prior consideration, for the study of the random distribution of different suggested statistics, derived from samples of a given size, must guide us in the choice of which statistic it is most profitable to calculate. The fact is, however, that very little progress has been made in the study of the distribution of statistics derived from samples. In 1900 PEARSON (15) gave the exact form of the distribution of χ^2, the Pearsonian test of goodness of fit, and in 1915 the same author published (18) a similar result of more general scope, valid when the observations are regarded as subject to linear constraints. By an easy adaptation (17) the tables of probability derived from this formula may be made available for the more numerous cases in which linear constraints are imposed upon the hypothetical population by the means which we employ in its reconstruction. The distribution of the mean of samples of n from a normal population has long been known, but in 1908 "Student" (4) broke new ground by calculating the distribution of the ratio which the deviation of the mean from its population value bears to the standard deviation calculated from the sample. At the same time he gave the exact form of the distribution in samples of the standard deviation. In 1915 FISHER (5) published the curve of distribution of the correlation coefficient for the standard method of calculation, and in 1921 (6) he published the corresponding series of curves for intraclass correlations. The brevity of this list is emphasised by the absence of investigation of other important statistics, such as the regression coefficients, multiple correlations, and the correlation ratio. A formula for the probable error of any statistic is, of course, a practical necessity, if that statistic is to be of service: and in the majority of cases such formulæ have been found, chiefly by the labours of PEARSON and his school, by a first approximation, which describes the distribution with sufficient accuracy if the sample is sufficiently large. Problems of distribution, other than the distribution of statistics, used to be not uncommon as examination problems in probability, and the physical importance of problems of this type may be exemplified by the chemical laws of mass action, by the statistical mechanics of GIBBS, developed by JEANS in its application to the theory of gases, by the electron theory of LORENTZ, and by PLANCK'S development of the theory of quanta, although in all these applications the methods employed have been, from the statistical point of view, relatively simple.

The discussions of theoretical statistics may be regarded as alternating between problems of estimation and problems of distribution. In the first place a method of calculating one of the population parameters is devised from common-sense considerations: we next require to know its probable error, and therefore an approximate solution of the distribution, in samples, of the statistic calculated. It may then become apparent that other statistics may be used as estimates of the same parameter. When the probable errors of these statistics are compared, it is usually found that, in large samples, one particular method of calculation gives a result less subject to random errors than those given by other methods of calculation. Attacking the problem more thoroughly, and calculating the surface of distribution of any two statistics, we may find that the whole of the relevant information contained in one is contained in the other: or, in other words, that when once we know the other, knowledge of the first gives us no further information as to the value of the parameter. Finally it may be possible to prove, as in the case of the Mean Square Error, derived from a sample of normal population (7), that a particular statistic summarises the whole of the information relevant to the corresponding parameter, which the sample contains. In such a case the problem of estimation is completely solved.

4. Criteria of Estimation.

The common-sense criterion employed in problems of estimation may be stated thus:—That when applied to the whole population the derived statistic should be equal to the parameter. This may be called the *Criterion of Consistency*. It is often the only test applied: thus, in estimating the standard deviation of a normally distributed population, from an ungrouped sample, either of the two statistics—

$$\sigma_1 = \frac{1}{n}\sqrt{\frac{\pi}{2}}\,\mathrm{S}\left(|x-\bar{x}|\right) \qquad \text{(Mean error)}$$

and

$$\sigma_2 = \sqrt{\frac{1}{n}\,\mathrm{S}\,(x-\bar{x})^2} \qquad \text{(Mean square error)}$$

will lead to the correct value, σ, when calculated from the whole population. They both thus satisfy the criterion of consistency, and this has led many computers to use the first formula, although the result of the second has 14 per cent. greater weight (7), and the labour of increasing the number of observations by 14 per cent. can seldom be less than that of applying the more accurate formula.

Consideration of the above example will suggest a second criterion, namely:—That in large samples, when the distributions of the statistics tend to normality, that statistic is to be chosen which has the least probable error.

This may be called the *Criterion of Efficiency*. It is evident that if for large samples one statistic has a probable error double that of a second, while both are proportional to $n^{-\frac{1}{2}}$, then the first method applied to a sample of $4n$ values will be no more accurate than the second applied to a sample of any n values. If the second method makes use of the whole of the information available, the first makes use of only one-quarter of it, and its efficiency may therefore be said to be 25 per cent. To calculate the efficiency of any given method, we must therefore know the probable error of the statistic calculated by that method, and that of the most efficient statistic which could be used. The square of the ratio of these two quantities then measures the efficiency.

The criterion of efficiency is still to some extent incomplete, for different methods of calculation may tend to agreement for large samples, and yet differ for all finite samples. The complete criterion suggested by our work on the mean square error (7) is:—

That the statistic chosen should summarise the whole of the relevant information supplied by the sample.

This may be called the *Criterion of Sufficiency*.

In mathematical language we may interpret this statement by saying that if θ be the parameter to be estimated, θ_1 a statistic which contains the whole of the information as to the value of θ, which the sample supplies, and θ_2 any other statistic, then the surface of distribution of pairs of values of θ_1 and θ_2, for a given value of θ, is such that for a given value of θ_1, the distribution of θ_2 does not involve θ. In other words, when θ_1 is known, knowledge of the value of θ_2 throws no further light upon the value of θ.

It may be shown that a statistic which fulfils the criterion of sufficiency will also fulfil the criterion of efficiency, when the latter is applicable. For, if this be so, the distribution of the statistics will in large samples be normal, the standard deviations being proportional to $n^{-\frac{1}{2}}$. Let this distribution be

$$df = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-r^2}}\, e^{-\frac{1}{1-r^2}\left\{\frac{\overline{\theta_1-\theta}^2}{2\sigma_1^2} - \frac{2r\overline{\theta_1-\theta}\,\overline{\theta_2-\theta}}{2\sigma_1\sigma_2} + \frac{\overline{\theta_2-\theta}^2}{2\sigma_2^2}\right\}}\, d\theta_1\, d\theta_2,$$

then the distribution of θ_1 is

$$df = \frac{1}{\sigma_1\sqrt{2\pi}}\, e^{-\frac{\overline{\theta_1-\theta}^2}{2\sigma_1^2}}\, d\theta_1,$$

so that for a given value of θ_1 the distribution of θ_2 is

$$df = \frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-r^2}}\, e^{-\frac{1}{2\,\overline{1-r^2}}\left\{\frac{r\,\overline{\theta_1-\theta}}{\sigma_1} - \frac{\overline{\theta_2-\theta}}{\sigma_2}\right\}^2}\, d\theta_2\,;$$

and if this does not involve θ, we must have

$$r\sigma_2 = \sigma_1\,;$$

showing that σ_1 is necessarily less than σ_2, and that the efficiency of θ_2 is measured by r^2, when r is its correlation in large samples with θ_1.

Besides this case we shall see that the criterion of sufficiency is also applicable to finite samples, and to those cases when the weight of a statistic is not proportional to the number of the sample from which it is calculated.

5. Examples of the Use of the Criterion of Consistency.

In certain cases the criterion of consistency is sufficient for the solution of problems of estimation. An example of this occurs when a fourfold table is interpreted as representing the double dichotomy of a normal surface. In this case the dichotomic ratios of the two variates, together with the correlation, completely specify the four fractions into which the population is divided. If these are equated to the four fractions into which the sample is divided, the correlation is determined uniquely.

In other cases where a small correction has to be made, the amount of the correction is not of sufficient importance to justify any great refinement in estimation, and it is sufficient to calculate the discrepancy which appears when the uncorrected method is applied to the whole population. Of this nature is SHEPPARD'S correction for grouping, and it will illustrate this use of the criterion of consistency if we derive formulæ for this correction without approximation.

Let ξ be the value of the variate at the mid point of any group, a the interval of grouping, and x the true value of the variate at any point, then the k^{th} moment of an infinite grouped sample is

$$\sum_{p=-\infty}^{p=\infty} \int_{\xi-\frac{1}{2}a}^{\xi+\frac{1}{2}a} \xi^k f(x)\, dx,$$

in which $f(x)\,dx$ is the frequency, in any element dx, of the ungrouped population, and

$$\xi = \left(p+\frac{\theta}{2\pi}\right)a,$$

p being any integer.

Evidently the k^{th} moment is periodic in θ, we will therefore equate it to

$$\text{A}_0+\text{A}_1 \sin\theta+\text{A}_2 \sin 2\theta\ldots$$

$$+\text{B}_1 \cos\theta+\text{B}_2 \cos 2\theta\ldots.$$

Then

$$\text{A}_0 = \frac{1}{2\pi} \sum_{p=-\infty}^{p=\infty} \int_0^{2\pi} d\theta \int_{\xi-\frac{1}{2}a}^{\xi+\frac{1}{2}a} \xi^k f(x)\, dx$$

$$\text{A}_s = \frac{1}{\pi} \sum_{p=-\infty}^{p=\infty} \int_0^{2\pi} \sin s\theta\, d\theta \int_{\xi-\frac{1}{2}a}^{\xi+\frac{1}{2}a} \xi^k f(x)\, dx,$$

$$\text{B}_s = \frac{1}{\pi} \sum_{p=-\infty}^{p=\infty} \int_0^{2\pi} \cos s\theta\, d\theta \int_{\xi-\frac{1}{2}a}^{\xi+\frac{1}{2}a} \xi^k f(x)\, dx.$$

But

$$\theta = \frac{2\pi}{a}\xi-2\pi p,$$

therefore

$$d\theta = \frac{2\pi}{a}d\xi,$$

$$\sin s\theta = \sin\frac{2\pi}{a}s\xi,$$

$$\cos s\theta = \cos\frac{2\pi}{a}s\xi,$$

hence

$$\text{A}_0 = \frac{1}{a}\int_{-\infty}^{\infty} d\xi \int_{\xi-\frac{1}{2}a}^{\xi+\frac{1}{2}a} \xi^k f(x)\, dx = \frac{1}{a}\int_{-\infty}^{\infty} f(x)\, dx \int_{x-\frac{1}{2}a}^{x+\frac{1}{2}a} \xi^k\, d\xi.$$

Inserting the values 1, 2, 3 and 4 for k, we obtain for the aperiodic terms of the first four moments of the grouped population

$$
\begin{aligned}
{}_1\mathrm{A}_0 &= \int_{-\infty}^{\infty} x f(x)\,dx,\\
{}_2\mathrm{A}_0 &= \int_{-\infty}^{\infty} \left(x^2+\frac{a^2}{12}\right) f(x)\,dx,\\
{}_3\mathrm{A}_0 &= \int_{-\infty}^{\infty} \left(x^3+\frac{a^2x}{4}\right) f(x)\,dx,\\
{}_4\mathrm{A}_0 &= \int_{-\infty}^{\infty} \left(x^4+\frac{a^2x^2}{2}+\frac{a^4}{80}\right) f(x)\,dx.
\end{aligned}
$$

If we ignore the periodic terms, these equations lead to the ordinary SHEPPARD corrections for the second and fourth moment. The nature of the approximation involved is brought out by the periodic terms. In the absence of high contact at the ends of the curve, the contribution of these will, of course, include the terms given in a recent paper by PEARSON (8) ; but even with high contact it is of interest to see for what degree of coarseness of grouping the periodic terms become sensible.

Now

$$
\begin{aligned}
\mathrm{A}_\mathrm{S} &= \frac{1}{\pi}\sum_{p=-\infty}^{p=\infty}\int_0^{2\pi} \sin s\theta\,d\theta \int_{\xi-\frac{1}{2}a}^{\xi+\frac{1}{2}a} \xi^k f(x)\,dx,\\
&= \frac{2}{a}\int_{-\infty}^{\infty} \sin\frac{2\pi s\xi}{a}\,d\xi \int_{\xi-\frac{1}{2}a}^{\xi+\frac{1}{2}a} \xi^k f(x)\,dx,\\
&= \frac{2}{a}\int_{-\infty}^{\infty} f(x)\,dx \int_{x-\frac{1}{2}a}^{x+\frac{1}{2}a} \xi^k \sin\frac{2\pi s\xi}{a}\,d\xi.
\end{aligned}
$$

But

$$\frac{2}{a}\int_{x-\frac{1}{2}a}^{x+\frac{1}{2}a} \xi \sin\frac{2\pi s\xi}{a}\,d\xi = -\frac{a}{\pi s}\cos\frac{2\pi sx}{a}\cos\pi s,$$

therefore

$${}_1\mathrm{A}_\mathrm{S} = (-)^{s+1}\frac{a}{\pi s}\int_{-\infty}^{\infty} \cos\frac{2\pi sx}{a}\, f(x)\,dx\,;$$

similarly the other terms of the different moments may be calculated.

For a normal curve referred to the true mean

$$
\begin{aligned}
{}_1\mathrm{A}_\mathrm{S} &= (-)^{s+1}\frac{2\epsilon}{s}\,e^{-\frac{s^2\sigma^2}{2\epsilon^2}},\\
{}_1\mathrm{B}_\mathrm{S} &= 0,
\end{aligned}
$$

in which

$$a = 2\pi\epsilon.$$

The error of the mean is therefore

$$-2\epsilon\left(e^{-\frac{\sigma^2}{2\epsilon^2}}\sin\theta-\tfrac{1}{2}e^{-\frac{4\sigma^2}{2\epsilon^2}}\sin 2\theta+\tfrac{1}{3}e^{-\frac{9\sigma^2}{2\epsilon^2}}\sin 3\theta-\ldots\right).$$

To illustrate a coarse grouping, take the group interval equal to the standard deviation: then

$$\epsilon = \frac{\sigma}{2\pi},$$

and the error is

$$-\frac{\sigma}{\pi} e^{-2\pi^2} \sin\theta$$

with sufficient accuracy. The standard error of the mean being $\frac{\sigma}{\sqrt{n}}$, we may calculate the size of the sample for which the error due to the periodic terms becomes equal to one-tenth of the standard error, by putting

$$\frac{\sigma}{10\sqrt{n}} = \frac{\sigma}{\pi} e^{-2\pi^2},$$

whence

$$n = \frac{\pi^2}{100} e^{4\pi^2} = 13{,}790 \times 10^{12}$$

For the second moment

$$\mathrm{B}_s = (-)^s 4\left(\sigma^2 + \frac{\epsilon^2}{s^2}\right) e^{-\frac{s^2\sigma^2}{2\epsilon^2}},$$

and, if we put

$$\frac{\sqrt{2}\sigma^2}{10\sqrt{n}} = 4\sigma^2 e^{-2\pi^2},$$

there results

$$n = \tfrac{1}{800} e^{4\pi^2} = 175 \times 10^{12}$$

The error, while still very minute, is thus more important for the second than for the first moment.

For the third moment

$$\mathrm{A}_s = (-)^s \frac{6\sigma^4 s}{\epsilon}\left\{1 + \frac{\epsilon^2}{s^2\sigma^2} - \frac{\epsilon^4}{3s^4\sigma^4}(\pi^2 s^2 - 6)\right\} e^{-\frac{s^2\sigma^2}{2\epsilon^2}};$$

putting

$$\frac{\sqrt{15}\sigma^3}{10\sqrt{n}} = 12\pi\sigma^3 e^{-2\pi^2},$$

$$n = \frac{1}{960\pi^2} e^{4\pi^2} = 14{\cdot}7 \times 10^{12}$$

While for the fourth moment

$$\mathrm{B}_s = (-)^{s+1} \frac{8\sigma^6 s^4}{\epsilon^2}\left\{1 - (\pi^2 s^2 - 3)\frac{\epsilon^4}{s^4\sigma^4} - (\pi^2 s^2 - 6)\frac{\epsilon^6}{s^6\sigma^6}\right\} e^{-\frac{s^2\sigma^2}{2\epsilon^2}},$$

so that, if we put,

$$\frac{\sqrt{96}\sigma^4}{10\sqrt{n}} = 32\pi^2\sigma^4 e^{-2\pi^2},$$

$$n = \frac{3}{3200\pi^4} e^{4\pi^2} = 1{\cdot}34 \times 10^{12}$$

In a similar manner the exact form of SHEPPARD'S correction may be found for other curves; for the normal curve we may say that the periodic terms are exceedingly minute so long as a is less than σ, though they increase very rapidly if a is increased beyond this point. They are of increasing importance as higher moments are used, not only absolutely, but relatively to the increasing probable errors of the higher moments. The principle upon which the correction is based is merely to find the error when the moments are calculated from an infinite grouped sample; the corrected moment therefore fulfils the criterion of consistency, and so long as the correction is small no greater refinement is required.

Perhaps the most extended use of the criterion of consistency has been developed by PEARSON in the "Method of Moments." In this method, which is without question of great practical utility, different forms of frequency curves are fitted by calculating as many moments of the sample as there are parameters to be evaluated. The parameters chosen are those of an infinite population of the specified type having the same moments as those calculated from the sample.

The system of curves developed by PEARSON has four variable parameters, and may be fitted by means of the first four moments. For this purpose it is necessary to confine attention to curves of which the first four moments are finite; further, if the accuracy of the fourth moment should increase with the size of the sample, that is, if its probable error should not be infinitely great, the first eight moments must be finite. This restriction requires that the class of distribution in which this condition is not fulfilled should be set aside as "heterotypic," and that the fourth moment should become practically valueless as this class is approached. It should be made clear, however, that there is nothing anomalous about these so-called "heterotypic" distributions except the fact that the method of moments cannot be applied to them. Moreover, for that class of distribution to which the method can be applied, it has not been shown, except in the case of the normal curve, that the best values will be obtained by the method of moments. The method will, in these cases, certainly be serviceable in yielding an approximation, but to discover whether this approximation is a good or a bad one, and to improve it, if necessary, a more adequate criterion is required.

A single example will be sufficient to illustrate the practical difficulty alluded to above. If a point P lie at known (unit) distance from a straight line AB, and lines be drawn at random through P, then the distribution of the points of intersection with AB will be distributed so that the frequency in any range dx is

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

in which x is the distance of the infinitesimal range dx from a fixed point 0 on the line, and m is the distance, from this point, of the foot of the perpendicular PM. The distribution will be a symmetrical one (Type VII.) having its centre at $x = m$ (fig. 1). It is therefore a perfectly definite problem to estimate the value of m (to find the best value of m) from a random sample of values of x. We have stated the problem in its simplest possible form: only one parameter is required, the middle point of the distribution.

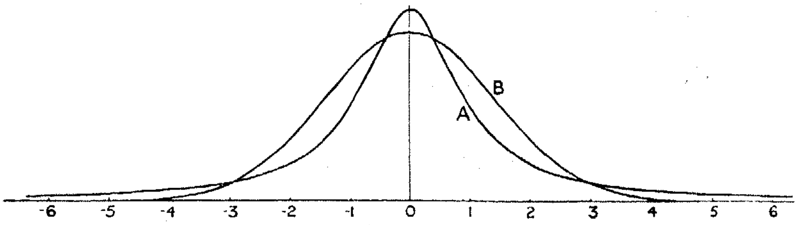

Fig. 1. Symmetrical error curves of equal intrinsic accuracy.

$$\text{A} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad df = \frac{1}{\pi}\,\frac{dx}{1+x^2}.$$

$$\text{B} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad d_j = \frac{1}{2\sqrt{\pi}}\,e^{-\frac{x^2}{4}}$$

By the method of moments, this should be given by the first moment, that is by the mean of the observations: such would seem to be at least a good estimate. It is, however, entirely valueless. The distribution of the mean of such samples is in fact the same, identically, as that of a single observation. In taking the mean of 100 values of x, we are no nearer obtaining the value of m than if we had chosen any value of x out of the 100. The problem, however, is not in the least an impracticable one: clearly from a large sample we ought to be able to estimate the centre of the distribution with some precision; the mean, however, is an entirely useless statistic for the purpose. By taking the median of a large sample, a fair approximation is obtained, for the standard error of the median of a large sample of n is $\frac{\pi}{2\sqrt{n}}$, which, alone, is enough to show that by adopting adequate statistical methods it must be possible to estimate the value for m, with increasing accuracy, as the size of the sample is increased.

This example serves also to illustrate the practical difficulty which observers often find, that a few extreme observations appear to dominate the value of the mean. In these cases the rejection of extreme values is often advocated, and it may often happen that gross errors are thus rejected. As a statistical measure, however, the rejection of observations is too crude to be defended: and unless there are other reasons for rejection than mere divergence from the majority, it would be more philosophical to accept these extreme values, not as gross errors, but as indications that the distribution of errors is not normal. As we shall show, the only Pearsonian curve for which the mean is the best statistic for locating the curve, is the normal or gaussian curve of errors. If the curve is not of this form the mean is not necessarily, as we have seen, of any value whatever. The determination of the true curves of variation for different types of work is therefore of great practical importance, and this can only be done by different workers recording their data in full without rejections, however they may please to treat the data so recorded. Assuredly an observer need be exposed to no criticism, if after recording data which are not probably normal in distribution, he prefers to adopt some value other than the arithmetic mean.

6. Formal Solution of Problems of Estimation.

The form in which the criterion of sufficiency has been presented is not of direct assistance in the solution of problems of estimation. For it is necessary first to know the statistic concerned and its surface of distribution, with an infinite number of other statistics, before its sufficiency can be tested. For the solution of problems of estimation we require a method which for each particular problem will lead us automatically to the statistic by which the criterion of sufficiency is satisfied. Such a method is, I believe, provided by the Method of Maximum Likelihood, although I am not satisfied as to the mathematical rigour of any proof which I can put forward to that effect. Readers of the ensuing pages are invited to form their own opinion as to the possibility of the method of the maximum likelihood leading in any case to an insufficient statistic. For my own part I should gladly have withheld publication until a rigorously complete proof could have been formulated; but the number and variety of the new results which the method discloses press for publication, and at the same time I am not insensible of the advantage which accrues to Applied Mathematics from the co-operation of the Pure Mathematician, and this co-operation is not infrequently called forth by the very imperfections of writers on Applied Mathematics.

If in any distribution involving unknown parameters $\theta_1, \theta_2, \theta_3, \dots$, the chance of an observation falling in the range dx be represented by

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

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

$$\frac{n!}{\Pi\,(n_p!)}\,\Pi\,\{f(x_p, \theta_1, \theta_2, \dots)\, dx_p\}^{n_p}.$$

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

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

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

In order to make clear the distinction between this method and that of BAYES, we will apply it to the same type of problem as that which BAYES discussed, in the hope of making clear exactly of what kind is the information which a sample is capable of supplying. This question naturally first arose, not with respect to populations distributed in frequency curves and surfaces, but with respect to a population regarded as divided into two classes only, in fact in problems of *probability*. A certain proportion, p, of an infinite population is supposed to be of a certain kind, *e.g.*, "successes," the remainder are then "failures." A sample of n is taken and found to contain x successes and y failures. The chance of obtaining such a sample is evidently

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

Applying the method of maximum likelihood, we have

$$\mathrm{S}(\log f) = x\log\hat{p}+y\log(1-\hat{p})$$

whence, differentiating with respect to p, in order to make this quantity a maximum,

$$\frac{x}{p} = \frac{y}{1-\hat{p}}, \quad \text{or} \quad \hat{p} = \frac{x}{n}.$$

The question then arises as to the accuracy of this determination. This question was first discussed by BAYES (10), in a form which we may state thus. After observing this sample, when we know p, what is the *probability* that p lies in any range dp? In other words, what is the frequency distribution of the values of p in populations which are selected by the restriction that a sample of n taken from each of them yields x successes. Without further data, as BAYES perceived, this problem is insoluble. To render it capable of mathematical treatment, BAYES introduced the *datum*, that among the populations upon which the experiment was tried, those in which p lay in the range dp were equally frequent for all equal ranges dp. The probability that the value of p lay in any range dp was therefore assumed to be simply dp, before the sample was taken. After the selection effected by observing the sample, the probability is clearly proportional to

$$p^x(1-p)^y\,dp.$$

After giving this solution, based upon the particular datum stated, BAYES adds a *scholium* the purport of which would seem to be that in the absence of all knowledge save that supplied by the sample, it is reasonable to assume this particular *a priori* distribution of p. The *result*, the *datum*, and the *postulate* implied by the *scholium*, have all been somewhat loosely spoken of as BAYES' Theorem.

* For know p , read know $\hat{p}$.

The postulate would, if true, be of great importance in bringing an immense variety of questions within the domain of probability. It is, however, evidently extremely arbitrary. Apart from evolving a vitally important piece of knowledge, that of the exact form of the distribution of values of p, out of an assumption of complete ignorance, it is not even a unique solution. For we might never have happened to direct our attention to the particular quantity p: we might equally have measured probability upon an entirely different scale. If, for instance,

$$\sin\theta = 2p-1,$$

the quantity, θ, measures the degree of probability, just as well as p, and is even, for some purposes, the more suitable variable. The chance of obtaining a sample of x successes and y failures is now

$$\frac{n!}{2^n x!\, y!}(1+\sin\theta)^x(1-\sin\theta)^y;$$

applying the method of maximum likelihood,

$$S(\log f) = x\log(1+\sin\theta)+y\log(1-\sin\theta)-n\log 2,$$

and differentiating with respect to θ,

* $$\frac{x\cos\theta}{1+\sin\theta} = \frac{y\cos\theta}{1-\sin\theta}, \qquad \text{whence} \qquad \sin\theta = \frac{x-y}{2n},$$

an exactly equivalent solution to that obtained using the variable p. But what *a priori* assumption are we to make as to the distribution of θ? Are we to assume that θ is equally likely to lie in all equal ranges $d\theta$? In this case the *a priori* probability will be $d\theta/\pi$, and that after making the observations will be proportional to

$$(1+\sin\theta)^x(1-\sin\theta)^y\, d\theta.$$

But if we interpret this in terms of p, we obtain

$$p^x(1-p)^y\frac{dp}{\sqrt{p(1-p)}} = p^{x-\frac{1}{2}}(1-p)^{y-\frac{1}{2}}\,dp,$$

a result inconsistent with that obtained previously. In fact, the distribution previously assumed for p was equivalent to assuming the special distribution for θ,

$$df = \frac{\cos\theta}{2}\,d\theta,$$

the arbitrariness of which is fully apparent when we use any variable other than p.

In a less obtrusive form the same species of arbitrary assumption underlies the method

* For $2n$, read n.

known as that of inverse probability. Thus, if the same observed result A might be the consequence of one or other of two hypothetical conditions X and Y, it is assumed that the probabilities of X and Y are in the same ratio as the probabilities of A occurring on the two assumptions, X is true, Y is true. This amounts to assuming that before A was observed, it was known that our universe had been selected at random from an infinite population in which X was true in one half, and Y true in the other half. Clearly such an assumption is entirely arbitrary, nor has any method been put forward by which such assumptions can be made even with consistent uniqueness. There is nothing to prevent an irrelevant distinction being drawn among the hypothetical conditions represented by X, so that we have to consider two hypothetical possibilities X_1 and X_2, on both of which A will occur with equal frequency. Such a distinction should make no difference whatever to our conclusions; but on the principle of inverse probability it does so, for if previously the relative probabilities were reckoned to be in the ratio x to y, they must now be reckoned $2x$ to y. Nor has any criterion been suggested by which it is possible to separate such irrelevant distinctions from those which are relevant.

There would be no need to emphasise the baseless character of the assumptions made under the titles of inverse probability and BAYES' Theorem in view of the decisive criticism to which they have been exposed at the hands of BOOLE, VENN, and CHRYSTAL, were it not for the fact that the older writers, such as LAPLACE and POISSON, who accepted these assumptions, also laid the foundations of the modern theory of statistics, and have introduced into their discussions of this subject ideas of a similar character. I must indeed plead guilty in my original statement of the Method of the Maximum Likelihood (9) to having based my argument upon the principle of inverse probability; in the same paper, it is true, I emphasised the fact that such inverse probabilities were relative only. That is to say, that while we might speak of one value of p as having an inverse probability three times that of another value of p, we might on no account introduce the differential element dp, so as to be able to say that it was three times as probable that p should lie in one rather than the other of two equal elements. Upon consideration, therefore, I perceive that the word probability is wrongly used in such a connection: probability is a ratio of frequencies, and about the frequencies of such values we can know nothing whatever. We must return to the actual fact that one value of p, of the frequency of which we know nothing, would yield the observed result three times as frequently as would another value of p. If we need a word to characterise this relative property of different values of p, I suggest that we may speak without confusion of the *likelihood* of one value of p being thrice the likelihood of another, bearing always in mind that likelihood is not here used loosely as a synonym of probability, but simply to express the relative frequencies with which such values of the hypothetical quantity p would in fact yield the observed sample.

The solution of the problems of calculating from a sample the parameters of the hypothetical population, which we have put forward in the method of maximum likelihood, consists, then, simply of choosing such values of these parameters as have the maximum likelihood. Formally, therefore, it resembles the calculation of the mode of an inverse frequency distribution. This resemblance is quite superficial : if the scale of measurement of the hypothetical quantity be altered, the mode must change its position, and can be brought to have any value, by an appropriate change of scale ; but the optimum, as the position of maximum likelihood may be called, is entirely unchanged by any such transformation. Likelihood also differs from probability* in that it is not a differential element, and is incapable of being integrated : it is assigned to a particular point of the range of variation, not to a particular element of it. There is therefore an absolute measure of probability in that the unit is chosen so as to make all the elementary probabilities add up to unity. There is no such absolute measure of likelihood. It may be convenient to assign the value unity to the maximum value, and to measure other likelihoods by comparison, but there will then be an infinite number of values whose likelihood is greater than one-half. The sum of the likelihoods of admissible values will always be infinite.

Our interpretation of BAYES' problem, then, is that the likelihood of any value of p is proportional to

$$p^x (1-p)^y,$$

and is therefore a maximum when

$$p = \frac{x}{n},$$

which is the best value obtainable from the sample ; we shall term this the *optimum* value of p. Other values of p for which the likelihood is not much less cannot, however, be deemed unlikely values for the true value of p. We do not, and cannot, know, from the information supplied by a sample, anything about the probability that p should lie between any named values.

The reliance to be placed on such a result must depend upon the frequency distribution of x, in different samples from the same population. This is a perfectly objective statistical problem, of the kind we have called problems of distribution ; it is, however, capable of an approximate solution, directly from the mathematical form of the likelihood.

When for large samples the distribution of any statistic, θ_1, tends to normality, we

* It should be remarked that likelihood, as above defined, is not only fundamentally distinct from mathematical probability, but also from the logical "probability" by which Mr. KEYNES (21) has recently attempted to develop a method of treatment of uncertain inference, applicable to those cases where we lack the statistical information necessary for the application of mathematical probability. Although, in an important class of cases, the likelihood may be held to measure the degree of our rational belief in a conclusion, in the same sense as Mr. KEYNES' "probability," yet since the latter quantity is constrained, somewhat arbitrarily, to obey the addition theorem of mathematical probability, the likelihood is a quantity which falls definitely outside its scope.

may write down the chance for a given value of the parameter θ, that θ_1 should lie in the range $d\theta_1$ in the form

$$\Phi = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\theta_1-\theta)^2}{2\sigma^2}} d\theta_1.$$

The mean value of θ_1 will be the true value θ, and the standard deviation is σ, the sample being assumed sufficiently large for us to disregard the dependence of σ upon θ.

The likelihood of any value, θ, is proportional to

$$e^{-\frac{(\theta_1-\theta)^2}{2\sigma^2}},$$

this quantity having its maximum value, unity, when

$$\theta = \theta_1;$$

for

$$\frac{\partial}{\partial\theta}\log\Phi = \frac{\theta_1-\theta}{\sigma^2}.$$

Differentiating now a second time

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

Now Φ stands for the total frequency of all samples for which the chosen statistic has the value θ_1, consequently $\Phi = S'(\phi)$, the summation being taken over all such samples, where ϕ stands for the probability of occurrence of a certain specified sample. For which we know that

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

the summation being taken over the individual members of the sample.

If now we expand $\log f$ in the form

$$\log f(\theta) = \log f(\theta_1) + \overline{\theta-\theta_1}\frac{\partial}{\partial\theta}\log f(\theta_1) + \frac{\overline{\theta-\theta_1}^2}{\underline{|2}}\frac{\partial^2}{\partial\theta^2}\log f(\theta_1) + \ldots,$$

or

$$\log f = \log f_1 + a\,\overline{\theta-\theta_1} + \frac{b}{2}\overline{\theta-\theta_1}^2 + \ldots,$$

we have

$$\log\phi = C + \overline{\theta-\theta_1}\,S(a) + \tfrac{1}{2}\overline{\theta-\theta_1}^2\,S(b) + \ldots;$$

now for optimum statistics

$$S(a) = 0,$$

and for sufficiently large samples $S(b)$ differs from $n\bar{b}$ only by a quantity of order $\sqrt{n}\,\sigma_b$; moreover, $\theta-\theta_1$ being of order $n^{-\frac{1}{2}}$, the only terms in $\log\phi$ which are not reduced without limit, as n is increased, are

$$\log\phi = C + \tfrac{1}{2}n\bar{b}\,\overline{\theta-\theta_1}^2;$$

hence

$$\phi \propto e^{\frac{1}{2} n \bar{b}\, \overline{\theta - \theta_1}^2}.$$

Now this factor is constant for all samples which have the same value of θ_1, hence the variation of Φ with respect to θ is represented by the same factor, and consequently

$$\log \Phi = C' + \tfrac{1}{2} n \bar{b}\, \overline{\theta - \theta_1}^2 ;$$

whence

$$-\frac{1}{\sigma_{\theta_1}^2} = \frac{\partial^2}{\partial \theta^2} \log \Phi = n \bar{b},$$

where

$$b = \frac{\partial^2}{\partial \theta^2} \log f(\theta_1),$$

θ_1 being the optimum value of θ.

The formula

$$-\frac{1}{\sigma_\theta^2} = n \overline{\frac{\partial^2}{\partial \theta^2} \log f}$$

supplies the most direct way known to me of finding the probable errors of statistics. It may be seen that the above proof applies only to statistics obtained by the method of maximum likelihood.*

For example, to find the standard deviation of

$$\hat{p} = \frac{x}{n}$$

* A similar method of obtaining the standard deviations and correlations of statistics derived from large samples was developed by PEARSON and FILON in 1898 (16). It is unfortunate that in this memoir no sufficient distinction is drawn between the *population* and the *sample*, in consequence of which the formulæ obtained indicate that the likelihood is always a maximum (for continuous distributions) when the *mean* of each variate in the sample is equated to the corresponding mean in the population (16, p. 232, " $A_r = 0$ "). If this were so the mean would always be a sufficient statistic for location ; but as we have already seen, and will see later in more detail, this is far from being the case. The same argument, indeed, is applied to all statistics, as to which nothing but their *consistency* can be truly affirmed.

The probable errors obtained in this way are those appropriate to the method of maximum likelihood, but not in other cases to statistics obtained by the method of moments, by which method the examples given were fitted. In the ' Tables for Statisticians and Biometricians ' (1914), the probable errors of the constants of the Pearsonian curves are those proper to the method of moments ; no mention is there made of this change of practice, nor is the publication of 1898 referred to.

It would appear that shortly before 1898 the process which leads to the correct value, of the probable errors of *optimum* statistics, was hit upon and found to agree with the probable errors of statistics found by the method of moments for *normal* curves and surfaces ; without further enquiry it would appear to have been assumed that this process was valid in all cases, its directness and simplicity being peculiarly attractive. The mistake was at that time, perhaps, a natural one ; but that it should have been discovered and corrected without revealing the inefficiency of the method of moments is a very remarkable circumstance.

In 1903 the correct formulæ for the probable errors of statistics found by the method of moments are given in ' Biometrika ' (19) ; references are there given to SHEPPARD (20), whose method is employed, as well as to PEARSON and FILON (16), although both the method and the results differ from those of the latter.

in samples from an infinite population of which the true value is p,

$$\log f = x \log p + y \log(1-p),$$

$$\frac{\partial}{\partial p}\log f = \frac{x}{p} - \frac{y}{1-p},$$

$$\frac{\partial^2}{\partial p^2}\log f = -\frac{x}{p^2} - \frac{y}{\overline{1-p}^2}.$$

Now the mean value of x is pn, and of y is $(1-p)n$, hence the mean value of $\frac{\partial^2}{\partial p^2}\log f$ is

$$-\left(\frac{1}{p}+\frac{1}{1-p}\right)n;$$

therefore

$$\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n},$$

the well-known formula for the standard error of p.

7. Satisfaction of the Criterion of Sufficiency.

That the criterion of sufficiency is generally satisfied by the solution obtained by the method of maximum likelihood appears from the following considerations.

If the individual values of any sample of data are regarded as co-ordinates in hyperspace, then any sample may be represented by a single point, and the frequency distribution of an infinite number of random samples is represented by a density distribution in hyperspace. If any set of statistics be chosen to be calculated from the samples, certain regions will provide identical sets of statistics; these may be called *isostatistical* regions. For any particular space element, corresponding to an actual sample, there will be a particular set of parameters for which the frequency in that element is a maximum; this will be the optimum set of parameters for that element. If now the set of statistics chosen are those which give the optimum values of the parameters, then all the elements of any part of the same isostatistical region will contain the greatest possible frequency for the same set of values of the parameters, and therefore any region which lies wholly within an isostatistical region will contain its maximum frequency for that set of values.

Now let θ be the value of any parameter, $\hat{\theta}$ the statistic calculated by the method of maximum likelihood, and θ_1 any other statistic designed to estimate the value of θ, then for a sample of given size, we may take

$$f(\theta, \hat{\theta}, \theta_1)\, d\hat{\theta}\, d\theta_1$$

to represent the frequency with which $\hat{\theta}$ and θ_1 lie in the assigned ranges $d\hat{\theta}$ and $d\theta_1$.

The region $d\hat{\theta}\, d\theta_1$ evidently lies wholly in the isostatistical region $d\hat{\theta}$. Hence the equation

$$\frac{\partial}{\partial\theta}\log f(\theta, \hat{\theta}, \theta_1) = 0$$

is satisfied, irrespective of θ_1, by the value $\theta = \hat{\theta}$. This condition is satisfied if

$$f(\theta, \hat{\theta}, \theta_1) = \phi(\theta, \hat{\theta}) \,.\, \phi'(\hat{\theta}, \theta_1);$$

for then

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

and the equation for the optimum degenerates into

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

which does not involve θ_1.

But the factorisation of f into factors involving $(\theta, \hat{\theta})$ and $(\hat{\theta}, \theta_1)$ respectively is merely a mathematical expression of the condition of sufficiency; and it appears that any statistic which fulfils the condition of sufficiency must be a solution obtained by the method of the optimum.

It may be expected, therefore, that we shall be led to a sufficient solution of problems of estimation in general by the following procedure. Write down the formula for the probability of an observation falling in the range dx in the form

$$f(\theta, x)\, dx,$$

where θ is an unknown parameter. Then if

$$\mathrm{L} = \mathrm{S}(\log f),$$

the summation being extended over the observed sample, L differs by a constant only from the logarithm of the likelihood of any value of θ. The most likely value, $\hat{\theta}$, is found by the equation

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

and the standard deviation of $\hat{\theta}$, by a second differentiation, from the formula

$$\frac{\partial^2 \mathrm{L}}{\partial\theta^2} = -\frac{1}{\sigma_{\hat{\theta}}^2};$$

this latter formula being applicable only where $\hat{\theta}$ is normally distributed, as is often the case with considerable accuracy in large samples. The value $\sigma_{\hat{\theta}}$ so found is in these cases the least possible value for the standard deviation of a statistic designed to estimate the same parameter ; it may therefore be applied to calculate the efficiency of any other such statistic.

When several parameters are determined simultaneously, we must equate the second differentials of L, with respect to the parameters, to the coefficients of the quadratic terms in the index of the normal expression which represents the distribution of the corresponding statistics. Thus with two parameters,

$$\frac{\partial^2 \mathrm{L}}{\partial \theta_1{}^2} = -\frac{1}{1-r^2_{\hat{\theta}_1\hat{\theta}_2}} \cdot \frac{1}{\sigma^2_{\hat{\theta}_1}}, \qquad \frac{\partial^2 \mathrm{L}}{\partial \theta_2{}^2} = -\frac{1}{1-r^2_{\hat{\theta}_1\hat{\theta}_2}} \cdot \frac{1}{\sigma^2_{\hat{\theta}_2}},$$

$$\frac{\partial^2 \mathrm{L}}{\partial \theta_1\, \partial \theta_2} = +\frac{1}{1-r^2_{\hat{\theta}_1\hat{\theta}_2}} \cdot \frac{r}{\sigma_{\hat{\theta}_1}\sigma_{\hat{\theta}_2}},$$

* or, in effect, $\sigma_{\hat{\theta}}^2$ is found by dividing the Hessian determinant of L, with respect to the parameters, into the corresponding minor.

The application of these methods to such a series of parameters as occur in the specification of frequency curves may best be made clear by an example.

8. The Efficiency of the Method of Moments in Fitting Curves of the Pearsonian Type III.

Curves of Pearson's Type III. offer a good example for the calculation of the efficiency of the Method of Moments. The chance oi an observation falling in the range dx is

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

By the method of moments the curve is located by means of the statistic μ_1, its dimensions are ascertained from the second moment μ_2, and the remaining parameter p is determined from β_1. Considering first the problem of location, if a and p were known and we had only to determine m, we should take, according to the method of moments,

$$\mu_1 = m_\mu + a\,(p+1),$$

where m_μ represents the estimate of the parameter m, obtained by using the method of moments. The variance of m_μ is, therefore,

$$\sigma^2_{m_\mu} = \sigma^2_{\mu_1} = \frac{\mu_2}{n} = \frac{a^2\,(p+1)}{n}.$$

If, on the other hand, we aim at greater accuracy, and make the likelihood of the sample a maximum for variations of m, we have

$$\mathrm{L} = -n \log a - n \log (p\,!) + p\mathrm{S}\left(\log \frac{x-m}{a}\right) - \mathrm{S}\left(\frac{x-m}{a}\right),$$

* The expression, $x\,!$, is used here and throughout as equivalent to the Gaussian $\Pi\,(x)$, or to $\Gamma\,(x+1)$, whether x is an integer or not.

* For L, read -L.

and the equation to determine m is

$$\frac{\partial \mathrm{L}}{\partial m} = -p\mathrm{S}\left(\frac{1}{x-\hat{m}}\right) + \frac{n}{a} = 0\,; \qquad (1)$$

the accuracy of the value so obtained is found from the second differential.

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

of which the mean value is

$$-\frac{n}{a^2(p-1)},$$

whence

$$\sigma_{\hat{m}}^{\ 2} = \frac{a^2(p-1)}{n}.$$

We now see that the efficiency of location by the method of moments is

$$\frac{p-1}{p+1} = 1 - \frac{2}{p+1}.$$

Efficiencies of over 80 per cent. for location are therefore obtained if p exceeds 9 ; for $p = 1$ the efficiency of location vanishes, as in other cases where the curve makes an angle with the axis at the end of its range.

Turning now to the problem of scaling, we have, by the method of moments,

$$\mu_2 = a^2(p+1),$$

whence, knowing p, a is obtained. Since

$$\sigma^2_{\mu_2} = \frac{\beta_2-1}{n}\mu_2^{\ 2},$$

we must have

$$\sigma^2_{a_\mu} = \frac{\beta_2-1}{4n}a^2 = \frac{4+3\beta_1}{8n}a^2 = \frac{p+4}{2(p+1)n}a^2\,;$$

on the other hand, from the value of L, we find the equation

$$\frac{\partial \mathrm{L}}{\partial a} = -\frac{n}{a}(p+1) + \frac{1}{a^2}\mathrm{S}(x-m) = 0, \qquad (2)$$

to be solved for m and a as a simultaneous equation with (1) ; whence

$$\frac{\partial^2 \mathrm{L}}{\partial m\,\partial a} = -\frac{n}{a^2},$$

and

$$\frac{\partial^2 \mathrm{L}}{\partial a^2} = \frac{n}{a^2}(p+1) - \frac{2}{a^3}\mathrm{S}(x-m),$$

of which the mean value is

$$-\frac{n(p+1)}{a^2}.$$

The variance of a, determined from this pair of simultaneous equations, is found by dividing

*

$$-\frac{\partial^2 L}{\partial m^2} = +\frac{n}{a^2(p-1)}$$

by the determinant

*

*

$$\begin{vmatrix} +\frac{n}{a^2(p-1)} & +\frac{n}{a^2} \\ +\frac{n}{a^2} & +\frac{n(p+1)}{a^2} \end{vmatrix}$$

which reduces to

*

$$+\frac{2}{p-1}\cdot\frac{n^2}{a^4},$$

whence

$$\sigma_a{}^2 = \frac{a^2}{2n},$$

and the efficiency of scaling by the method of moments is

$$\frac{p+1}{p+4} = 1 - \frac{3}{p+4}.$$

Efficiency of over 80 per cent. for scaling are, therefore, obtained when p exceeds 11. The efficiency of scaling does not, however, vanish for any possible value of p, though it tends to zero, as p approaches its limiting value, -1.

Lastly, p is found by the method of moments by putting

$$\frac{4}{p+1} = \beta_1.$$

Now

$$\sigma_{\beta_1}{}^2 = \frac{\beta_1}{n}(4\beta_4 - 24\beta_2 + 36 + 9\beta_1\beta_2 - 12\beta_3 + 35\beta_1),$$

and for curves of Type III,

$$\begin{aligned} \beta_2 &= 3 + \tfrac{3}{2}\beta_1, \\ \beta_3 &= 2\beta_1\beta_2 + 4\beta_1 = \beta_1(3\beta_1 + 10), \\ \beta_4 &= \tfrac{5}{2}(\beta_3 + 2\beta_2) = \tfrac{5}{2}(3\beta_1{}^2 + 13\beta_1 + 6), \end{aligned}$$

hence

$$\sigma_{\beta_1}{}^2 = \frac{3\beta_1}{n}(5\beta_1 + 4)(\beta_1 + 4),$$

$$= \frac{\beta_1{}^2}{n}\cdot\frac{6(p+2)(p+6)}{p+1},$$

* The signs of all terms have been changed.

whence it follows, since n is large, that

$$\sigma_{p_\mu}^2 = \frac{\overline{p+1}^2}{n} \cdot \frac{6(p+2)(p+6)}{p+1} = \frac{6}{n}(p+1)(p+2)(p+6).$$

From the value of L,

$$\frac{\partial L}{\partial p} = -n\frac{d}{dp}\log(p!) + S\left(\log\frac{x-m}{a}\right),$$

which equation solved for m, a and p as a simultaneous equation with (1) and (2), will yield the set of values for the parameters which has the maximum likelihood. To find the variance of the value of p, so obtained, observe that

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

of which the mean value is

$$-\frac{n}{ap},$$

*

$$\frac{\partial^2 L}{\partial a\,\partial p} = -\frac{n}{a^2},$$

and

$$\frac{\partial^2 L}{\partial p^2} = -n\frac{d^2}{dp^2}\log(p!).$$

The variance of p, derived from this set of simultaneous equations, is therefore found by dividing the minor of $\frac{\partial^2 L}{\partial p^2}$, namely

$$\frac{2}{p-1}\cdot\frac{n^2}{a^4},$$

by the determinant

$$\frac{n^3}{a^4}\begin{vmatrix} \frac{1}{p-1} & 1 & \frac{1}{p} \\ 1 & p+1 & 1 \\ \frac{1}{p} & 1 & \frac{d^2}{dp^2}\log(p!) \end{vmatrix} = \frac{n^3}{a^4}\cdot\frac{1}{p-1}\left\{2\frac{d^2}{dp^2}\log(p!) - \frac{2}{p} + \frac{1}{p^2}\right\};$$

hence

$$\sigma_p^2 = \frac{2}{n\left\{2\frac{d^2}{dp^2}\log(p!) - \frac{2}{p} + \frac{1}{p^2}\right\}}.$$

When p is large,

$$2\frac{d^2}{dp^2}\log(p!) - \frac{2}{p} + \frac{1}{p^2} = \tfrac{1}{3}\left(\frac{1}{p^3} - \frac{1}{5p^5} + \frac{1}{7p^7}\dots\right),$$

* For $-\frac{n}{a^2}$, read $-\frac{n}{a}$.

so that, approximately,

$$\sigma_{p}{}^{2} = \frac{6}{n}(p^{3}+\tfrac{1}{5}p)\,;$$

for large values of p, the efficiency of the method of moments is, therefore, approximately

$$\frac{p^{3}+\frac{1}{5}p}{\overline{p+1}\,\overline{p+2}\,\overline{p+6}}\,.$$

Efficiencies of over 80 per cent. occur when p exceeds 38·1 ($\beta_1 = 0{\cdot}102$); evidently the method of moments is effective for determining the form of the curve only when it is relatively close to the normal form. For small values of p, the above approximation for the efficiency is not adequate. The true values can easily be obtained from the recently published tables of the Trigamma* function (11). The following values are obtained for the integral values of p from 0 to 5.

p	0	1	2	3	4	5
Efficiency . . .	0	0·0274	0·0871	0·1532	0·2159	0·2727

An interesting point which may be resolved at this stage of the enquiry is to find the variance of m, when a and p are not known, derived from the above set of simultaneous equations; that is to say, to calculate the accuracy with which the limiting point of the curve is determined; such determinations are often stated as the result of fitting curves of limited range, but their probable errors are seldom, if ever, evaluated. To obtain the greatest possible accuracy with which such a point can be determined we must divide the minor of $\frac{\partial^2 \mathrm{L}}{\partial m^2}$, namely,

$$\frac{n^{2}}{a^{2}}\left\{\overline{p+1}\,\frac{d^{2}}{dp^{2}}\log(p\,!)-1\right\}\cdot$$

by

$$\frac{n^{3}}{a^{4}}\cdot\frac{1}{p-1}\left\{2\frac{d^{2}}{dp^{2}}\log(p\,!)-\frac{2}{p}+\frac{1}{p^{2}}\right\},$$

whence

$$\sigma_{m}{}^{2}=\frac{a^{2}}{n}\,\frac{\overline{p-1}\left\{\overline{p+1}\,\frac{d^{2}}{dp^{2}}\log(p\,!)-1\right\}}{2\frac{d^{2}}{dp^{2}}\log(p\,!)-\frac{2}{p}+\frac{1}{p^{2}}}\,.$$

The position of the limiting point will, when p is at all large, evidently be determined with much less accuracy than is the position, as a whole, of a curve of known form and size. Let n' be a multiplier such that the position of the extremity of a curve calculated

* It is sometimes convenient to write $\digamma(x)$ for $\frac{d^{2}}{dx^{2}}\log(x\,!)$.

from nn' observations will be determined with the same accuracy as the position, as a whole, of a curve of known form and size, can be determined from a sample of n observations when n is large. Then

$$n' = \frac{\overline{p+1}\,\frac{d^2}{dp^2}\log(p\,!)-1}{2\,\frac{d^2}{dp^2}\log(p\,!)-\frac{2}{p}+\frac{1}{p^2}};$$

but, when p is large,

$$\overline{p+1}\,\frac{d^2}{dp^2}\log(p\,!)-1 = \frac{1}{2p}\left(1-\frac{2}{3p}+\frac{1}{3p^2}\cdots\right)$$

and

$$2\frac{d^2}{dp^2}\log(p\,!)-\frac{2}{p}+\frac{1}{p^2} = \frac{1}{3p^3}\left(1-\frac{1}{5p^2}\cdots\right);$$

therefore

$$\begin{aligned} n' &= \tfrac{3}{2}p^2\left(1-\frac{2}{3p}\,p+\frac{8}{15p^2}\cdots\right) \\ &= \tfrac{3}{2}p^2-p+\tfrac{4}{5}. \end{aligned}$$

For large values of p the probable error of the determination of the end-point may be found approximately by multiplying the probable error of location by

$$(p-\tfrac{1}{3})\sqrt{\tfrac{3}{2}}.$$

As p grows smaller, n' diminishes until it reaches unity, when $p = 1$. For values of p less than 1 it would appear that the end-point had a smaller probable error than the probable error of location, but, as a matter of fact, for these values location is determined by the end-point, and as we see from the vanishing of $\sigma_{\hat{m}}$, whether or not p and a are known, when $p = \mathbf{1}$, the weight of the determination from this point onwards increases more rapidly than n, as the sample increases. (See Section 10.)

The above method illustrates how it is possible to calculate the variance of any function of the population parameters as estimated from large samples; by comparing this variance with that of the same function estimated by the method of moments, we may find the efficiency of that method for any proposed function. The above examination, in which the determinations of the locus, the scale, and the form of the curve are treated separately, will serve as a general criterion of the application of the method of moments to curves of Type III. Special combinations of the parameters will, however, be of interest in special cases. It may be noted here that by virtue of equation (2) the function of $m + a\,(p+1)$ is the same, whether determined by moments or by the method of the optimum:

$$m_\mu + a_\mu\,(p_\mu+1) = \hat{m}+\hat{a}\,(\hat{p}+1).$$

The efficiency of the method of moments in determining this function is therefore 100 per cent. That this function is the abscissa of the mean does not imply 100 per cent. efficiency of location, for the centre of location of these curves is not the mean (see p. 340).

9. Location and Scaling of Frequency Curves in General.

The general problem of the location and scaling of curves may now be treated more generally. This is the problem which presents itself with respect to *error curves* of assumed form, when to find the best value of the quantity measured we must *locate* the curve as accurately as possible, and to find the probable error of the result of this process we must, as accurately as possible, estimate its *scale*.

The *form* of the curve may be specified by a function ϕ, such that

$$df \propto e^{\phi(\xi)}\, d\xi, \text{ when } \xi = \frac{x-m}{a}.$$

In this expression ϕ specifies the form of the curve, which is unaltered by variations of a and m.

When a sample of n observations has been taken, the likelihood of any combination of values of a and m is

$$\mathrm{L} = \mathrm{C} - n \log a + \mathrm{S}(\phi),$$

whence

$$\frac{\partial \mathrm{L}}{\partial m} = \mathrm{S}\left(\frac{d\phi}{d\xi} \cdot \frac{d\xi}{dm}\right) = -\frac{1}{a}\mathrm{S}(\phi'),$$

since

$$\frac{\partial \xi}{\partial m} = -\frac{1}{a};$$

also

$$\frac{\partial \mathrm{L}}{\partial a} = -\frac{1}{a}\mathrm{S}(\xi\phi') - \frac{n}{a},$$

since

$$\frac{\partial \xi}{\partial a} = -\frac{\xi}{a}.$$

Differentiating a second time,

$$\frac{\partial^2 \mathrm{L}}{\partial m^2} = \frac{1}{a^2}\mathrm{S}(\phi''):$$

therefore

$$\sigma_m^2 = -\frac{a^2}{n\overline{\phi''}}.$$

This expression enables us to compare the accuracy of error curves of different form, when the location is performed in each case by the method which yields the minimum error.

Example :—The curve

$$df = \frac{1}{\pi}\frac{d\xi}{1+\xi^2}$$

referred to in Section 5 has an infinite standard deviation, but it is not on that account an error curve of zero accuracy, for

$$\phi = -\log(1+\xi^2), \qquad \phi' = -\frac{2\xi}{1+\xi^2}, \qquad \phi'' = -\frac{2(1-\xi^2)}{(1+\xi^2)^2}.$$

Now

$$\frac{1}{\pi}\int_{-\infty}^{\infty}\frac{1-\xi^2}{(1+\xi^2)^3}\,d\xi = \tfrac{1}{4},$$

hence

$$\overline{\phi''} = -\tfrac{1}{2} \qquad \text{and} \qquad \sigma_{\hat{m}}^2 = \frac{2a^2}{n}$$

The quantity,

$$-\frac{\overline{\phi''}}{a^2} = \frac{1}{2a^2},$$

which is the factor by which n is multiplied in calculating the weight of the estimate made from n measurements, may be called the *intrinsic accuracy* of an *error curve*. In the above example we see that errors distributed so that

$$df = \frac{a}{\pi}\frac{dx}{a^2+x^2}$$

have the same intrinsic accuracy as errors distributed according to the normal curve

$$df = \frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}\,dx,$$

provided

$$\sigma^2 = 2a^2.$$

Fig. 1 illustrates two such curves of equal intrinsic accuracy.

Returning now to the general problem in which

$$\mathrm{L} = \mathrm{C} - n\log a + \mathrm{S}\,(\phi),$$

we have

$$\frac{\partial^2 \mathrm{L}}{\partial m\,\partial a} = \frac{1}{a^2}\,\mathrm{S}\,(\phi' + \xi\phi'') = \frac{1}{a^2}\,\mathrm{S}\,(\xi\phi'')$$

and

$$\frac{\partial^2 \mathrm{L}}{\partial a^2} = \frac{1}{a^2}\,\mathrm{S}\,(2\xi\phi' + \xi^2\phi'') + \frac{n}{a^2} = \frac{1}{a^2}\,\mathrm{S}\,(\xi^2\phi'' - 1).$$

The latter expression will directly give the accuracy with which a is determined only if

$$\frac{\partial^2 \mathrm{L}}{\partial m\,\partial a} = 0,$$

and we can always arrange that this shall be so by subtracting from ξ the quantity

$$\frac{\overline{\xi\phi''}}{\overline{\phi''}}.$$

Thus in a Type III. curve where, referred to the end of the range,

$$\overline{\xi\phi''} = -1, \qquad \overline{\phi''} = -\frac{1}{p-1}.$$

instead of

$$\phi = p \log \xi - \xi$$

we must write

$$\phi = p \log \overline{\xi+p-1} - \overline{\xi+p-1}\,;$$

then

$$\phi' = \frac{p}{\xi+p-1} - 1, \qquad \phi'' = -\frac{p}{\overline{\xi+p-1}^{\,2}},$$

hence

$$\frac{\partial^2 \mathrm{L}}{\partial a^2} = \frac{1}{a^2}\,\mathrm{S}\,(\xi^2\phi''-1)$$

$$= \frac{1}{a^2}\,\mathrm{S}\left(-p + \frac{2p\,\overline{p-1}}{\xi+p-1} - \frac{p\,\overline{p-1}^{\,2}}{\overline{\xi+p-1}^{\,2}} - 1\right),$$

of which the mean value is

$$\frac{n}{a^2}(-p+2\overline{p-1}-\overline{p-1}-1) = -\frac{2n}{a^2},$$

hence

$$\sigma_a{}^2 = \frac{a^2}{2n}.$$

For one particular point of origin, therefore, the variations of the abscissa are uncorrelated with those of a; this point may be termed the *centre of location.*

Example:—To determine the centre of location of the curve of Type IV.,

$$df \propto e^{-\nu \tan^{-1}\xi}\,(1+\xi^2)^{-\frac{r+2}{2}}.$$

Here

$$\phi = -\nu \tan^{-1}\xi - \frac{r+2}{2}\log \overline{1+\xi^2},$$

$$\phi' = -(\nu + \overline{r+2}\,\xi)\,\overline{1+\xi^2}^{\,-1},$$

$$\phi'' = \overline{r+2}\;\overline{1+\xi^2}^{\,-1} + 2\,(\nu\xi - \overline{r+2})\,\overline{1+\xi^2}^{\,-2}\,;$$

from these we find

$$\overline{\phi''} = -\frac{\overline{r+1}\;\overline{r+2}\;\overline{r+4}}{\overline{r+4}^{\,2}+\nu^2},$$

$$\overline{\xi\phi''} = \frac{\overline{r+1}\;\overline{r+2}\;\nu}{\overline{r+4}^{\,2}+\nu^2},$$

so that

$$\frac{\overline{\xi\phi''}}{\overline{\phi''}} = -\frac{\nu}{r+4}.$$

The centre of location, therefore, at the distance from the mode,

$$= -\frac{\nu a}{r+4}.$$

Click here to view part 2 of paper