Introduction to Fisher (1922) On the Mathematical Foundations of Theoretical Statistics

Seymour Geisser
University of Minnesota

1. General Remarks

This rather long and extraordinary paper is the first full account of Fisher's ideas on the foundations of theoretical statistics, with the focus being on estimation. The paper begins with a sideswipe at Karl Pearson for a purported general proof of Bayes' postulate. Fisher then clearly makes a distinction between parameters, the objects of estimation, and the statistics that one arrives at to estimate the parameters. There was much confusion between the two since the same names were given to both parameters and statistics, e.g., mean, standard deviation, correlation coefficient, etc., without an indication of whether it was the population or sample value that was the subject of discussion. This formulation of the parameter value was certainly a critical step for theoretical statistics [see, e.g., Geisser (1975), footnote on p. 320 and Stigler (1976)]. In fact, Fisher attributed the neglect of theoretical statistics not only to this failure in distinguishing between parameter and statistic but also to a philosophical reason, namely, that the study of results subject to greater or lesser error implies that the precision of concepts is either impossible or not a practical necessity. He sets out to remedy the situation, and remedy it he did. Indeed, he did this so convincingly that for the next 50 years or so almost all theoretical statisticians were completely parameter bound, paying little or no heed to inference about observables.

Fisher states that the purpose of statistical methods is to reduce a large quantity of data to a few that are capable of containing as much as possible of the relevant information in the original data. Because the data will generally supply a large number of "facts," many more than are sought, much information in the data is irrelevant. This brings to the fore the Fisherian dictum that statistical analysis via the reduction of data is the process of extracting

the relevant information and excluding the irrelevant information. A way of accomplishing this is by modeling a hypothetical population specified by relatively few parameters.

Hence, the critical problems of theoretical statistics in 1920, according to Fisher, were (1) specification, choice of the hypothetical parametric distribution; (2) estimation, choice of the statistics for estimating the unknown parameters of the distribution; (3) sampling distributions, the exact or approximate distributions of the statistics used to estimate the parameters. For a majority of statisticians, these have been and still are the principal areas of statistical endeavor, 70 years later. The two most important additions to this view are that the parametric models were, at best, merely approximations of the underlying process generating the observations, and in view of this, much greater emphasis should be placed on observable inference rather than on parametric inference.

2. Foundational Developments

In this paper, Fisher develops a number of concepts relevant to the estimation of parameters. Some were previously introduced but not generally developed, and others appear for the first time. Here, also, the richness of Fisher's *lingua statistica* emerges, yielding poignant appelatives for his concepts, vague though some of them are. This activity will continue throughout all his future contributions. First he defines consistency: A statistic is consistent if, when calculated from the whole population, it is equal to the parameter describing the probability law. This is in contradistinction to the usual definition which entails a sequence of estimates, one for each sample size, that converges in probability to the appropriate parameter. While Fisher consistency is restricted to repeated samples from the same distribution, it does not suffer from the serious defect of the usual definition. That flaw was formally pointed out later by Fisher (1956): Suppose one uses an arbitrary value A for an estimator for $n < n_1$, where n is as large as one pleases, and for $n > n_1$ uses an asymptotically consistent estimator T_n. The entire sequence, now corrupted by A for $n < n_1$ and then immaculately transformed to T_n thereafter, remains a useless, but perfectly well-defined, consistent estimator for any n. Fisher is not to be trifled with!

Indicating that many statistics for the same parameter can be Fisher-consistent, in particular, the sample standard deviation and sample mean deviation for the standard deviation of a normal population, he goes on to suggest a criterion for efficiency. It is a large sample definition. Among all estimators for a parameter that are Fisher-consistent and whose distributions are asymptotically normal, the one with the smallest variance is efficient. Later, he shows that when the asymptotic distribution of the method of moments estimator is normal for the location of a uniform distribution while that

of the "optimum" estimator is double exponential, he realizes that the variance does not necessarily provide a satisfactory basis for comparison, especially for small samples. Thus, he also recognizes that his large sample definition of intrinsic accuracy (a measure of relative efficiency) should not be based on variances and a definition appropriate for small samples is required. In later papers, e.g., Fisher (1925), vague concepts of intrinsic accuracy will be replaced by the more precise amount of information per observation. At any rate, the large sample criterion is incomplete and needs to be supplemented by a sufficiency criterion. The "remarkable" property of this concept was previously pointed out when introduced for a special case without giving it a name [Fisher (1920)]. A statistic, then, is sufficient if it contains all the information in the sample regarding the parameter to be estimated; that is, given a sufficient statistic, the distribution of any other statistic does not involve the parameter. This compelling concept of his, including the factorization result, is still in vogue. Assuming a sufficient statistic and any other statistic whose joint distribution is asymptotically bivariate normal with both means being the parameter estimated, he then "demonstrates" that the sufficient statistic has an asymptotic variance smaller than that of the other statistic by a clever conditioning argument that exploits the correlation between the statistics. Hence, he claims that a sufficient* statistic satisfies the criterion of (large sample) efficiency. This "proof" of course could only apply to those statistics whose asymptotic bivariate distribution with the sufficient statistic was normal.

He comments further on the method of moments estimation procedure. While ascribing great practical utility to it, he also exposes some of its shortcomings. In particular, in estimating the center of a one-parameter Cauchy distribution, he points out that the first sample moment, the sample mean, which is the method of moments estimator is not consistent but the median is. He also cautions against the statistical rejection of outliers unless there are other substantive reasons. Rather than outright rejection, he proposes that it seriously be considered that the error distribution is not normal. Fisher effectively argues that the specification of the underlying probability law will generally require the full set of observations. A sufficient reduction is only meaningful once the probability law has been adequately established.

3. Maximum Likelihood

Fisher begins this part of his discourse acknowledging, first, that properties such as sufficiency, efficiency, and consistency per se were inadequate in directly obtaining an estimator. In solving any particular problem, we would

* In the author's note, Fisher (1950), there is a handwritten correction to the definition of intrinsic accuracy replacing sufficiency by efficiency, possibly based on his later recognition that maximum likelihood estimators were not always sufficient.

require a method that would lead automatically to the statistic which satisfied these criteria. He proposes such a method to be that of maximum likelihood, while admitting dissatisfaction with regard to the mathematical rigor of any proof that he can devise toward that result. Publication would have been withheld until a rigorous proof was found, but the number and variety of new results emanating from this method pressed him to publish. With some uncharacteristic humility, he says, "I am not insensible of the advantage which accrues to Applied Mathematics from the cooperation of the Pure Mathematician and this cooperation is not infrequently called forth by the very imperfections of writers on Applied Mathematics." This totally disarming statement would preclude any harsh commentary on the evident lack of rigor in many of his "proofs" here. Such evident modesty and good feelings toward mathematicians would never again flow from his pen.

Fisher (1912) had earlier argued for a form of maximum likelihood estimation. He had taken a Bayesian approach because the maximizing procedure resembled the calculation of the mode of a posterior probability. In the present paper, he is very concerned to differentiate it from the Bayesian approach. He also argues against the "customary" Bayesian use of flat priors on the grounds that different results are obtained when different scales for the parameters are considered.

To illustrate Fisher's argument, suppose x denotes the number of successes out of n independent trials with probability of success; then the likelihood function is

$$L(p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \qquad (0 < p < 1),$$

which is maximized when p is chosen to be x/n. Now, if a uniform distribution on (0, 1) is taken to be the prior distribution of p, then Bayesian analysis would yield

$$\pi(p) \propto p^x (1-p)^{n-x}$$

as the posterior density of p. But if we parameterize this Bernoulli process in a different way, say, in terms of θ with $\sin\theta = 2p - 1$, then the likelihood function of θ is

$$L(\theta) = \frac{n!}{x!(n-x)!} \frac{(1+\sin\theta)^x}{2^x} \frac{(1-\sin\theta)^{n-x}}{2^{n-x}} \qquad \left(-\frac{\pi}{2} < \theta < \frac{\pi}{2}\right),$$

which, when maximized with respect to θ, gives $\sin\hat{\theta} = (2x - n)/n = 2\hat{p} - 1$. Thus, the maximum likelihood estimate is invariant under a 1-1 transformation. For the Bayes approach, he questions the assignment of a prior assigned to θ. The uniformity of θ on $(-\pi/2, \pi/2)$ leads to the posterior density of p as

$$\pi(p) \propto p^{x-1/2} (1-p)^{n-x-1/2},$$

which is different from the previous result above. Due to this inconsistency and other reasons, Fisher derides the arbitrariness of the Bayes prior and

chooses not to adopt the Bayesian approach. He apparently was strongly influenced in this regard by the prior criticisms of inverse probability by Boole (1854) and Venn (1866). Venn's criticism led to the elimination of the material on inverse probability from the very prominent textbook on algebra at this time by Chrystal (1886).

Fisher's argument regarding invariance was convincing to many and undoubtedly was a setback for the Bayesian approach until Jeffreys (1946) proposed a transformation-invariant procedure for calculating a prior density. There is a bit of irony here in that Fisher's expected information quantity, used in this paper but precisely defined later by Fisher (1925), was the effectuating ingredient for Jeffreys. If $\tau = g(\theta)$ satisfies certain conditions, then the expected amount of information is

$$I(\theta) = E\left[\frac{\partial \log L(\theta)}{\partial \theta}\right]^2 = \left[\frac{d\tau}{d\theta}\right]^2 I(\tau),$$

and therefore

$$I^{1/2}(\theta)\, d\theta = I^{1/2}(\tau)\, d\tau.$$

Hence, using the positive square root of the expected amount of information as a prior on θ will transform invariantly on a prior for τ and vice-versa.

Fisher also suggests using $L(\theta)$ as a relative measure of the plausibility of various values of θ and introduces the term "likelihood" to distinguish the concept from probability, confessing that earlier he had confused the two. He says that "likelihood is not here used loosely as a synonym for probability, but simply to express the relative frequencies with which such values of the hypothetical quantity θ would in fact yield the observed sample." The likelihood is then an alternative measure of the degree of rational belief when inverse probability was not applicable, which he believed was true most of the time. In more recent years this has led to a school of inference using the likelihood in various ways [Barnard et al. (1963), Edwards (1973)].

Assuming that the distribution of the maximum likelihood estimator tends to normality, Fisher demonstrates that the variance is the reciprocal of the Fisher information. That is, if T is an "optimal" statistic satisfying

$$\frac{\partial \log L(\theta)}{\partial \theta} = 0$$

at $\theta = T$, then the variance† of the large-sample normal distribution of T is the inverse of

$$-\frac{\partial^2 \log L(\theta)}{\partial \theta^2}.$$

The general consistency of the maximum likelihood estimator is not used but essentially assumed, and the "proof" relies heavily on the initial assumption.

† In the ultimate formula on page 31 of the following abridged version, x should be replaced by n.

Fisher consistency of the maximum likelihood estimator is also assumed without proof.

Fisher derives the explicit form of the limiting normal distribution for the maximum likelihood estimator, after having "demonstrated" that a sufficient estimate has the smallest-variance normal distribution in large samples. Now the theory would be complete if the maximum likelihood estimator were found to be sufficient, since then the reciprocal of Fisher-expected information would be a lower bound against which efficiency could be measured. Fisher claims to show that it is sufficient, although his wording is ambiguous in some passages. Fisher's "demonstration" begins with the joint distribution of the maximum likelihood estimator $\hat{\theta}$ and an arbitrary statistic T, whose joint density f satisfies

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

at $\theta = \hat{\theta}$. The factorization

$$f(\hat{\theta}, t; \theta) = g(\hat{\theta}; \theta)h(\hat{\theta}, T)$$

is then deduced, and the sufficiency of $\hat{\theta}$ follows.

Using the inverse of the expected information,

$$\frac{1}{-E\left[\dfrac{\partial^2 \log L(\theta)}{\partial \theta^2}\right]}$$

as a lower bound on variance, Fisher illustrates the calculations on the Pearson-type III error curve with density function

$$f(x|m, a, p) \propto \left[\frac{x-m}{a}\right]^p \exp\left\{\frac{-(x-m)}{a}\right\},$$

where only m is unknown and to be estimated. The method of moments estimator is $m = X - a(p+1)$ with variance $a^2(p+1)/n$, whereas the asymptotic variance of the maximum likelihood estimator is

$$\frac{a^2(p-1)}{n} = -\frac{1}{E\left[\dfrac{\partial^2 \log L(m)}{\partial m^2}\right]}.$$

Thus, the method of moments is not efficient for any n and approaches zero efficiency as $p \to 1$. In addition, he points out that for estimating the location parameter of a Cauchy distribution, the method of moments is useless. No moment of the Cauchy exists and the distribution of the sample mean is independent of the sample size! Savage (1976) later provides a curious, if not pathological, example in which a Fisher-consistent estimator is derived that is sufficient and hence loses no information in the Fisher sense, while the maximum likelihood estimator θ is not sufficient and hence does lose infor-

mation. But $\hat{\theta}$ has a smaller mean squared error for a sufficiently large sample size.

The method of maximum likelihood appears to have been anticipated by Edgeworth (1908–9) according to Pratt (1976). Although there is less than universal consensus for this view, there is ample evidence that Edgeworth derived the method in the translation case directly and also using inverse probability. It appears he also conjectured the asymptotic efficiency of the method without giving it a name.

4. Other Topics

The remainder of the paper contains mainly applications of maximum likelihood techniques and various relative efficiency calculations. There is a long discussion of the Pearson system of frequency curves. This section serves mainly to display Fisher's analytic virtuosity in handling the Pearson system, also displaying graphs that serve to characterize the system in a more useful form than previously. This enables him to calculate for the various Pearson frequency curves,‡ regions for varying percent efficiencies of the method of moments estimators of location and scale. He also determines the conditions that make them fully efficient. In the latter case, he shows that if the log of a density is quartic, under certain conditions it will be approximately normal and fit the Pearson system. In dealing with the Pearson-type III curve, he now demonstrates that the asymptotic variance of the maximum likelihood estimators of scale a and shape p is smaller than that of their method of moments counterparts. However, he fails to remark or perhaps notice the anomaly of the nonregular case. Here the asymptotic variance of the maximum likelihood estimator of a is larger when m and p are given than when only p is given. Similarly, the maximum likelihood estimator of p has smaller asymptotic variance when a and m are unknown than when a and m are known.

Interest in the Pearsonian system has declined considerably over the years, being supplanted by so-called nonparametric and robust procedures, and revival appears unlikely unless Bayesians find use for them. The final part of the paper looks at discrete distributions, where the method of minimum chi-square is related to maximum likelihood, and the problem of Sheppard's correction for grouped normal data is addressed in detail. This and the material on the Pearson system actually make up the bulk of the paper. No doubt of considerable interest 70 years ago, it is of far less interest than the preceding work on the foundations. Fisher implies as much in his author's note. However, there is a final example that deals with the distribution of observations in a dilution series that is worthy of careful examination.

‡ The density on the bottom of page 342 as well as the one on page 343 of the original paper contain misprints. The section involving this material has been omitted in the abridged version of Fisher's paper which follows.

After earlier displaying the potential lack of efficiency inherent in an uncritical application of the method of moments, Fisher in an ingenious *volte-face* produces an estimation procedure for a dilution series example, which, though inefficient, is preferable to a fully efficient one essentially for economic and practical reasons. To be sure, in later years Fisher fulminated against the wholesale introduction of utility or decision theory into scientific work, but rarely again were such principles so elegantly and unobtrusively applied to such a significant practical problem. The analysis here represents a peerless blend of theory and application.

An important monitoring procedure, of ongoing interest and wide applicability, used in this instance for estimating the density of protozoa in soils, was brought to Fisher's attention. A series of dilutions of a soil sample solution were made such that each is reduced by a factor a. At each dilution, a uniform amount of the solution is deposited on s different plates containing a nutrient. After a proper incubation period, the number of protozoa on each plate is to be counted. A reasonable model for such a situation is that the chance of z protozoa on a plate is Poisson-distributed with expected value θ/a^x, where θ is the density or number of protozoa per unit volume of the original solution, and x the dilution level. A large number of such series were made daily for a variety of organisms. It proved either physically impossible or economically prohibitive to count the number of such organisms on every plate for many such series in order to estimate θ. First, Fisher suggests that only those plates containing no organisms be counted; the chance of such an occurrence at level x is $p_x = \exp(-\theta/a^x)$. By this device, an experimentally feasible situation is attained that produces a joint likelihood for Y_x, the number of sterile plates at level x, as

$$L = \prod_{x=0}^{k} \binom{s}{y_x} p_x^{y_x}(1 - p_x)^{s-y_x}$$

for dilution levels $x = 0, 1, \ldots, k$. He then calculates the contribution of a plate at level x to the information about $\log \theta$ to be

$$w_x = p_x(1 - p_x)^{-1}(\log p_x)^2.$$

This is informative as to the number of dilution levels necessary in such experiments. Further, the total expected information is approximately given as

$$s \sum_x w_x \approx \frac{s\pi^2}{(6 \log a)}.$$

The maximum likelihood solution to the problem, however, required a heavy investment of time and effort given the computational facilities of 1922. (Of course, it can easily be done today.)

At this point, Fisher employs a second wrinkle that makes the problem tractable. He suggests that the expected total number of sterile plates be equated to the observed total number in order to obtain an estimate of θ. This "rough" procedure has expected information with respect to $\log \theta$ of approxi-

mate value

$$\frac{s}{\log 2 \log a}.$$

This results in a very quick and easy procedure possessing an efficiency, independent of the dilution factor, of about 88%.

This may very well be one of the earliest statistical applications of a decision like approach to the analysis of data.

5. Summary

Clearly Fisher's paper was a landmark event in theoretical statistics. While it suffered from a lack of mathematical rigor, long analytic excursions into areas of lesser interest, and some confusion in parts, the novelty and number of ideas expressed here, both those developed from previous work and newly introduced, are still compelling for most statisticians. Although this paper is infrequently cited, its influence completely pervades the subsequent paper [Fisher (1925)],§ which presents a clearer exposition of his views. However, he poured into the 1922 paper, pell-mell, all his creative thinking and work on the foundations of statistics, the major exception being the fiducial argument. This work, filtered through the 1925 paper, has had a profound impact on statistical thinking unto this day. One has only to scan any serious work on the foundations to see that these ideas still have relevance in statistical theory, although citation is almost always to the 1925 paper.

References

Barnard, G., Jenkins, G.M. and Winsten, C.B. (1963). Likelihood inference and time series, *Jo. Roy. Statist. Soc., Ser. A*, **125**, 321–372.

Boole, G. (1854). *The Laws of Thought*. Dover, New York.

Chrystal, G. (1886). *Algebra*. Adam and Charles Black, London.

Edgeworth, F.Y. (1908–9). On the probable errors of frequency-constants, *Jo. Roy. Statist. Soc.*, **71** 381–397, 499–512, 651–678. Addendum ibid. **72**, 81–90.

Edwards, A.W.F. (1972). *Likelihood*. Cambridge University Press, New York.

Fisher, R.A. (1912). On an absolute criterion for fitting frequency curves, *Messenger of Math.*, **41**, 155–160.

Fisher, R.A. (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 Roy. Astro. Soc.*, **80**, 758–770.

§ This paper is cited a number of times in Fisher (1956), his final work on the foundations, while the seminal 1922 paper is not mentioned. In fact, the dozen or so times that Fisher subsequently cites the 1922 paper, he misdates it about half the time as 1921. This Fisherian slip, making him a year younger at its publication, accords with the author's note attributing certain deficiencies in the paper to youth.

Fisher, R.A. (1925). Theory of statistical estimation, *Proc. Cambridge Philos. Soc.*, **22**, 700–725.
Fisher, R. A. (1950). *Contributions to Mathematical Statistics.* Wiley, New York.
Fisher, R.A. (1956). *Statistical Methods and Scientific Inference.* Oliver and Boyd, Edinburgh.
Geisser, S. (1975). The predictive sample reuse method with applications, *Jo. Amer. Statist. Assoc.*, **70**, 320–328.
Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems, *Proc. Roy. Soc. London, Ser. A*, **186**, 453–454.
Pratt, J.W. (1976). F.V. Edgeworth and R.A. Fisher on the efficiency of maximum likelihood estimation, *Ann. Statist.*, 501–514.
Savage, L.J. (1976). On Rereading R.A. Fisher (with discussion), *Ann. Statist.*, **4**, 441–500.
Stigler, S.M. (1976). Discussion of Savage (1976), *Ann. Statist.*, **4**, 498–500.
Venn, J. (1866). *The Logic of Chance.* Macmillan, London.

On the Mathematical Foundations of Theoretical Statistics

R.A. Fisher
Fellow of Gonville and Caius College,
Chief Statistician, Rothamsted Experimental Station

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 efficiency. (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. (1.)

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 Fundamental 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 hap-

pened 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; ar 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

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

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; for 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 formulae. 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 probabil-

ity 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 formulae 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—

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

and

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

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 $n^{-1/2}$, 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 $n^{-1/2}$. Let this distribution be

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

then the distribution of θ_1 is

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

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

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

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 formulae 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

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

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

$$\mathrm{A}_0 + \mathrm{A}_1 \sin\theta + \mathrm{A}_2 \sin 2\theta \ldots + \mathrm{B}_1 \cos\theta + \mathrm{B}_2 \cos 2\theta \ldots .$$

Then

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

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

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

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

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

Inserting the values 1, 2, 3 and 4 for k, we obtain for the aperiodic terms of

the four moments of the grouped population

$$
\begin{aligned}
{}_1A_0 &= \int_{-\infty}^{\infty} xf(x)\,dx,\\
{}_2A_0 &= \int_{-\infty}^{\infty} \left(x^2 + \frac{a^2}{12}\right) f(x)\,dx,\\
{}_3A_0 &= \int_{-\infty}^{\infty} \left(x^3 + \frac{a^2 x}{4}\right) f(x)\,dx,\\
{}_4A_0 &= \int_{-\infty}^{\infty} \left(x^4 + \frac{a^2x^2}{2} + \frac{a^4}{80}\right) f(x)\,dx.
\end{aligned}
$$

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

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

But

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

therefore

$$
{}_1A_S = (-)^{s+1} \frac{a}{\pi s} \int_{-\infty}^{\infty} \cos \frac{2\pi sx}{a} f(x)\,dx;
$$

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

For a normal curve referred to the true mean

$$
{}_1A_S = (-)^{s+1} \frac{2\varepsilon}{s} e^{-(s^2\sigma^2/2\varepsilon^2)},
$$

$$
{}_1B_S = 0,
$$

in which

$$
a = 2\pi\varepsilon.
$$

The error of the mean is therefore

$$-2\varepsilon(e^{-(\sigma^2/2\varepsilon^2)}\sin\theta - \tfrac{1}{2}e^{-(4\sigma^2/2\varepsilon^2)}\sin 2\theta + \tfrac{1}{3}e^{-(9\sigma^2/2\varepsilon^2)}\sin 3\theta - \cdots).$$

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

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

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

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

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

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

putting

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

$$n = \frac{1}{960\pi^2}e^{4\pi^2} = 147 \times 10^{12}.$$

While for the fourth moment

$$B_s = (-)^{s+1}\frac{8\sigma^6 s^2}{\varepsilon^2}\left\{1 - (\pi^2 s^2 - 3)\frac{\varepsilon^4}{s^4\sigma^4} - (\pi^2 s^2 - 6)\frac{\varepsilon^6}{s^6\sigma^6}\right\} e^{-(s^2\sigma^2/2\varepsilon^2)},$$

so that, if we put,

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

$$n = \frac{3}{3200\pi^4}e^{4\pi^2} = 1.34 \times 10^{12}.$$

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 al-

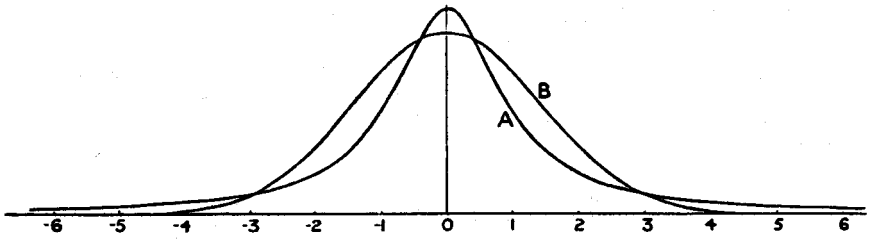

Figure 1. Symmetrical error curves of equal intrinsic accuracy:

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

$$\text{B}.....df = \frac{1}{2\sqrt{\pi}}e^{-x^2/4}$$

luded 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. 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 $\theta_1, \theta_2, \theta_3, \ldots$, the chance of an observation falling in the range dx be represented by

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

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

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

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

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

a maximum for variations of $\theta_1, \theta_2, \theta_3$, &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,

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

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 $\hat{p}$, 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.

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,

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

and differentiating with respect to θ,

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

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

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

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 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

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

* 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.

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 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

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

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

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

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

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

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

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

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

or

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

we have

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

now for optimum statistics

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

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 $n^{-1/2}$, the only terms in $\log \phi$ which are not reduced without limit, as n is increased, are

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

hence

$$\phi \propto e^{(1/2)n\bar{b}\overline{\theta-\theta_1}^2}.$$

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

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

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

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

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.*

* 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 formulae 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 corre-

For example, to find the standard deviation of

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

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

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

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

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

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.

sponding 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 formulae 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.

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

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

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,

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

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

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....

12. Discontinuous Distributions

The applications hitherto made of the optimum statistics have been problems in which the data are ungrouped, or at least in which the grouping intervals are so small as not to disturb the values of the derived statistics. By grouping, these continuous distributions are reduced to discontinuous distributions, and in an exact discussion must be treated as such.

If p_s be the probability of an observation falling in the cell (s), p_s being a function of the required parameters $\theta_1, \theta_2 \ldots$; and in a sample of N, if n_s are found to fall into that cell, then

$$\mathrm{S}(\log f) = \mathrm{S}(n_s \log p_s).$$

If now we write $\bar{n}_s = p_s \mathrm{N}$, we may conveniently put

$$\mathrm{L} = \mathrm{S}\left(n_s \log \frac{n_s}{\bar{n}_s}\right),$$

where L differs by a constant only from the logarithm of the likelihood, with sign reversed, and therefore the method of the optimum will consist in finding the *minimum* value of L. The equations so found are of the form

$$\frac{\partial \mathrm{L}}{\partial \theta} = -\mathrm{S}\left(\frac{n_s}{\bar{n}_s}\frac{\partial \bar{n}_s}{\partial \theta}\right) = 0. \tag{6}$$

It is of interest to compare these formulae with those obtained by making the Pearsonian χ^2 a minimum.

For

$$\chi^2 = \mathrm{S}\frac{(n_s - \bar{n}_s)^2}{\bar{n}_s},$$

and therefore

$$1 + \chi^2 = \mathrm{S}\left(\frac{n_s^2}{\bar{n}_s}\right),$$

so that on differentiating by $d\theta$, the condition that χ^2 should be a minimum for variations of θ is

$$-\mathrm{S}\left(\frac{n_s^2}{\bar{n}_s^2}\frac{\partial \bar{n}_s}{\partial \theta}\right) = 0. \tag{7}$$

Equation (7) has actually been used (12) to "improve" the values obtained by the method of moments, even in cases of normal distribution, and the Poisson series, where the method of moments gives a strictly sufficient solution. The discrepancy between these two methods arises from the fact that χ^2 is itself an approximation, applicable only when $\bar{n}_s$ and n_s are large, and the difference between them of a lower order of magnitude. In such cases

$$L = S\left(n_s \log \frac{n_s}{\bar{n}_s}\right) = S\left(\overline{m+x} \log \frac{m+x}{m}\right) = S\left\{x + \frac{x^2}{2m} - \frac{x^3}{6m^2} \cdots\right\},$$

and since

$$S(x) = 0,$$

we have, when x is in all cases small compared to m,

$$L = \frac{1}{2} S\left(\frac{x^2}{m}\right) = \frac{1}{2}\chi^2$$

as a first approximation. In those cases, therefore, when χ^2 is a valid measure of the departure of the sample from expectation, it is equal to 2L; in other cases the approximation fails and L itself must be used.

The failure of equation (7) in the general problem of finding the best values for the parameters may also be seen by considering cases of fine grouping, in which the majority of observations are separated into units. For the formula in equation (6) is equivalent to

$$S\left(\frac{1}{\bar{n}_s} \frac{\partial \bar{n}_s}{\partial \theta}\right)$$

where the summation is taken over all the observations, while the formula of equation (7), since it involves n_s^2, changes its value discontinuously, when one observation is gradually increased, at the point where it happens to coincide with a second observation.

Logically it would seem to be a necessity that that population which is chosen in fitting a hypothetical population to data should also appear the best when tested for its goodness of fit. The method of the optimum secures this agreement, and at the same time provides an extension of the process of testing goodness of fit, to those cases for which the χ^2 test is invalid.

The practical value of χ^2 lies in the fact that when the conditions are satisfied in order that it shall closely approximate to 2L, it is possible to give a general formula for its distribution, so that it is possible to calculate the probability, P, that in a random sample from the population considered, a worse fit should be obtained; in such cases χ^2 is distributed in a curve of the Pearsonian Type III.,

$$df \propto \left(\frac{\chi^2}{2}\right)^{(n'-3)/2} e^{-\chi^2/2}\, d\left(\frac{\chi^2}{2}\right)$$

or

$$df \propto L^{(n'-3)/2} e^{-L}\, dL,$$

where n' is one more than the number of degrees of freedom in which the sample may differ from expectation (17).

In other cases we are at present faced with the difficulty that the distribution L requires a special investigation. This distribution will in general be

discontinuous (as is that of χ^2), but it is not impossible that mathematical research will reveal the existence of effective graduations for the most important groups of cases to which χ^2 cannot be applied.

We shall conclude with a few illustrations of important types of discontinuous distribution.

1. The Poisson Series

$$e^{-m}\left(1, m, \frac{m^2}{2!}, \ldots, \frac{m^x}{x!}, \ldots\right)$$

involves only the single parameter, and is of great importance in modern statistics. For the optimum value of m,

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

whence

$$\mathrm{S}\left(\frac{x}{\hat{m}} - 1\right) = 0,$$

or

$$\hat{m} = \bar{x}.$$

The most likely value of m is therefore found by taking the first moment of the series.

Differentiating a second time,

$$-\frac{1}{\sigma^2_{\hat{m}}} = \mathrm{S}\left(-\frac{x}{m^2}\right) = -\frac{n}{m},$$

so that

$$\sigma^2_{\hat{m}} = \frac{m}{n},$$

as is well known.

2. Grouped Normal Data

In the case of the normal curve of distribution it is evident that the second moment is a sufficient statistic for estimating the standard deviation; in investigating a sufficient solution for grouped normal data, we are therefore in reality finding the optimum correction for grouping; the Sheppard correction having been proved only to satisfy the criterion of consistency.

For grouped normal data we have

$$p_s = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_s}^{x_{s+1}} e^{-(\overline{x-m}^2)/2\sigma^2}\, dx,$$

and the optimum values of m and σ are obtained from the equations,

$$\frac{\partial \mathrm{L}}{\partial m} = \mathrm{S}\left(\frac{n_s}{p_s}\frac{\partial p_s}{\partial m}\right) = 0,$$

$$\frac{\partial \mathrm{L}}{\partial \sigma} = \mathrm{S}\left(\frac{n_s}{p_s}\frac{\partial p_s}{\partial \sigma}\right) = 0;$$

or, if we write,

$$z = \frac{1}{\sigma\sqrt{2\pi}} e^{-(\overline{x-m}^2)/2\sigma^2},$$

we have the two conditions,

$$\mathrm{S}\left(\frac{n_s}{p_s}\overline{z_s - z_{s+1}}\right) = 0$$

and

$$\mathrm{S}\left\{\frac{n_s}{p_s}\left(\frac{x_s}{\sigma}z_s - \frac{x_{s+1}}{\sigma}z_{s+1}\right)\right\} = 0.$$

As a simple example we shall take the case chosen by K. Smith in her investigation of the variation of χ^2 in the neighbourhood of the moment solution (12).

Three hundred errors in right ascension are grouped in nine classes, positive and negative errors being thrown together as shown in the following table:

$0'' \cdot 1$ arc	0–1	1–2	2–3	3–4	4–5	5–6	6–7	7–8	8–9
Frequency . .	114	84	53	24	14	6	3	1	1

The second moment, without correction, yields the value

$$\sigma_v = 2.282542.$$

Using Sheppard's correction, we have

$$\sigma_\mu = 2.264214,$$

while the value obtained by making χ^2 a minimum is

$$\sigma_{\chi^2} = 2.355860.$$

If the latter value were accepted we should have to conclude that Sheppard's correction, even when it is small, and applied to normal data, might be alto-

gether of the wrong magnitude, and even in the wrong direction In order to obtain the optimum value of σ, we tabulate the values of $\frac{\partial L}{\partial \sigma}$ in the region under consideration; this may be done without great labour if values of σ be chosen suitable for the direct application of the table of the probability integral (13, Table II.). We then have the following values:

$\frac{1}{\sigma}$	0.43	0.44	0.45	0.46
$\frac{\partial L}{\partial \sigma}$	+15.135	+2.149	−11.098	−24.605
$\Delta^2 \frac{\partial L}{\partial \sigma}$		−0.261	−0.260	

By interpolation,

$$\frac{1}{\hat{\sigma}} = 0.441624$$

$$\hat{\sigma} = 2.26437.$$

We may therefore summarise these results as follows:—

Uncorrected estimate of σ	2.28254
Sheppard's correction	−0.01833
Correction for maximum likelihood	−0.01817
"Correction" for minimum χ^2	+0.07332

Far from shaking our faith, therefore, in the adequacy of Sheppard's correction, when small, for normal data, this example provides a striking instance of its effectiveness, while the approximate nature of the χ^2 test renders it unsuitable for improving a method which is already very accurate.

It will be useful before leaving the subject of grouped normal data to calculate the actual loss of efficiency caused by grouping, and the additional loss due to the small discrepancy between moments with Sheppard's correction and the optimum solution.

To calculate the loss of efficiency involved in the process of grouping normal data, let

$$v = \frac{1}{a} \int_{\xi-(1/2)a}^{\xi+(1/2)a} f(\xi)\, d\xi,$$

when a_σ is the group interval, then

$$v = f(\xi) + \frac{a^2}{24}f''(\xi) + \frac{a^4}{1920}f^{iv}(\xi) + \frac{a^6}{322{,}560}f^{vi}(\xi) + \cdots$$

$$= f(\xi)\left\{1 + \frac{a^2}{24}(\xi^2 - 1) + \frac{a^4}{1920}(\xi^4 - 6\xi^2 + 3)\right.$$

$$\left. + \frac{a^6}{322{,}560}(\xi^6 - 15\xi^4 + 45\xi^2 - 15) + \cdots\right\},$$

whence

$$\log v = \log f + \frac{a^2}{24}(\xi^2 - 1) - \frac{a^4}{2880}(\xi^4 + 4\xi^2 - 2)$$

$$+ \frac{a^6}{181{,}440}(\xi^6 + 6\xi^4 + 3\xi^2 - 1) - \cdots,$$

and

$$\frac{\partial^2}{\partial m^2}\log v = -\frac{1}{\sigma^2} + \frac{1}{\sigma^2}\left\{\frac{a^2}{12} - \frac{a^4}{720}(3\xi^2 + 2) + \frac{a^6}{30{,}240}(5\xi^4 + 12\xi^2 + 1) - \cdots\right\},$$

of which the mean value is

$$-\frac{1}{\sigma^2}\left\{1 - \frac{a^2}{12} + \frac{a^4}{144} - \frac{a^6}{1778} + \frac{31a^8}{25.12^4} - \frac{313a^{10}}{175.12^5}\right\}$$

neglecting the periodic terms; and consequently

$$\sigma_{\hat{m}}^2 = \frac{\sigma^2}{n}\left(1 + \frac{a^2}{12} - \frac{a^8}{86{,}400}\cdots\right).$$

Now for the mean of ungrouped data

$$\sigma_{\hat{m}}^2 = \frac{\sigma^2}{n},$$

so that the loss of efficiency due to grouping is nearly $\frac{a^2}{12}$.

The further loss caused by using the mean of the grouped data is very small, for

$$\sigma_{v_1}^2 = \frac{v_2}{n} = \frac{\sigma^2}{n}\left(1 + \frac{a^2}{12}\right),$$

neglecting the periodic terms; the loss of efficiency by using v_1 therefore is only

$$\frac{a^8}{86{,}400}.$$

Similarly for the efficiency for scaling,

$$\frac{\partial^2}{\partial\sigma^2}\log v$$

$$= \frac{1}{\sigma^2} - \frac{3\xi^2}{\sigma^2} + \frac{1}{\sigma^2}\left\{\frac{a^2}{12}(10\xi^2 - 3) - \frac{a^4}{360}(9\xi^4 + 21\xi^2 - 5)\right.$$

$$+ \frac{a^6}{30{,}240}(26\xi^6 + 110\xi^4 + 36\xi^2 - 7)$$

$$\left. - \frac{a^8}{1{,}814{,}400}(51\xi^8 + 315\xi^6 + 351\xi^4 - 55\xi^2 + 9) + \cdots\right\},$$

of which the mean value is

$$-\frac{2}{\sigma^2}\left\{1 - \frac{a^2}{6} + \frac{a^4}{40} - \frac{a^6}{270} + \frac{83a^8}{129{,}600}\cdots\right\},$$

neglecting the periodic terms; and consequently

$$\sigma_\phi^2 = \frac{\sigma^2}{2n}\left\{1 + \frac{a^2}{6} + \frac{a^4}{360} - \frac{a^8}{10{,}800}\cdots\right\}.$$

For ungrouped data

$$\sigma_\phi^2 = \frac{\sigma^2}{2n},$$

so that the loss of efficiency in scaling due to grouping is nearly $\frac{a^2}{6}$. This may be made as low as 1 per cent by keeping a less than $\frac{1}{4}$.

The further loss of efficiency produced by using the grouped second moment with Sheppard's correction is again very small, for

$$\sigma_{\nu_2}^2 = \frac{\nu_4 - \nu_2^2}{n} = \frac{2\sigma^4}{n}\left(1 + \frac{a^2}{6} + \frac{a^4}{360}\right)$$

neglecting the periodic terms.

Whence it appears that the further loss of efficiency is only

$$\frac{a^8}{10{,}800}.$$

We may conclude, therefore, that the high agreement between the optimum value of σ and that obtained by Sheppard's correction in the above example is characteristic of grouped normal data. The method of moments with Sheppard's correction is highly efficient in treating such material, the gain in efficiency obtainable by increasing the likelihood to its maximum value is trifling, and far less than can usually be gained by using finer groups. The loss of efficiency involved in grouping may be kept below 1 per cent. by making the group interval less than one-quarter of the standard deviation.

Although for the normal curve the loss of efficiency due to moderate grouping is very small, such is not the case with curves making a finite angle with the axis, or having at an extreme a finite or infinitely great ordinate. In such cases even moderate grouping may result in throwing away the greater part of the information which the sample provides....

13. Summary

During the rapid development of practical statistics in the past few decades, the theoretical foundations of the subject have been involved in great obscurity. Adequate distinction has seldom been drawn between the sample recorded and the hypothetical population from which it is regarded as drawn. This obscurity is centred in the so-called "inverse" methods.

On the bases that the purpose of the statistical reduction of data is to obtain statistics which shall contain as much as possible, ideally the whole, of the relevant information contained in the sample, and that the function of Theoretical Statistics is to show how such adequate statistics may be calculated, and how much and of what kind is the information contained in them, an attempt is made to formulate distinctly the types of problems which arise in statistical practice.

Of these, problems of Specification are found to be dominated by considerations which may change rapidly during the progress of Statistical Science. In problems of Distribution relatively little progress has hitherto been made, these problems still affording a field for valuable enquiry by highly trained mathematicians. The principal purpose of this paper is to put forward a general solution of problems of Estimation.

Of the criteria used in problems of Estimation only the criterion of Consistency has hitherto been widely applied; in Section 5 are given examples of the adequate and inadequate application of this criterion. The criterion of Efficiency is shown to be a special but important case of the criterion of Sufficiency, which latter requires that the whole of the relevant information supplied by a sample shall be contained in the statistics calculated.

In order to make clear the nature of the general method of satisfying the criterion of Sufficiency, which is here put forward, it has been thought necessary to reconsider Bayes' problem in the light of the more recent criticisms to which the idea of "inverse probability" has been exposed. The conclusion is drawn that two radically distinct concepts, both of importance in influencing our judgment, have been confused under the single name of *probability*. It is proposed to use the term *likelihood* to designate the state of our information with respect to the parameters of hypothetical populations, and it is shown that the quantitative measure of likelihood does not obey the mathematical laws of probability.

A proof is given in Section 7 that the criterion of Sufficiency is satisfied by that set of values for the parameters of which the likelihood is a maximum, and that the same function may be used to calculate the efficiency of any other statistics, or, in other words, the percentage of the total available information which is made use of by such statistics.

This quantitative treatment of the information supplied by a sample is illustrated by an investigation of the efficiency of the method of moments in fitting the Pearsonian curves of Type III.

Section 9 treats of the location and scaling of Error Curves in general, and contains definitions and illustrations of the *intrinsic accuracy*, and of the *centre of location* of such curves.

In Section 10 the efficiency of the method of moments in fitting the general Pearsonian curves is tested and discussed. High efficiency is only found in the neighbourhood of the normal point. The two causes of failure of the method of moments in locating these curves are discussed and illustrated. The special cause is discovered for the high efficiency of the third and fourth moments in the neighbourhood of the normal point.

It is to be understood that the low efficiency of the moments of a sample in estimating the form of these curves does not at all diminish the value of the notation of moments as a means of the comparative specification of the form of such curves as have finite moment coefficients.

Section 12 illustrates the application of the method of maximum likelihood to discontinuous distributions. The Poisson series is shown to be sufficiently fitted by the mean. In the case of grouped normal data, the Sheppard correction of the crude moments is shown to have a very high efficiency, as compared to recent attempts to improve such fits by making χ^2 a minimum; the reason being that χ^2 is an expression only approximate to a true value derivable from likelihood. As a final illustration of the scope of the new process, the theory of the estimation of micro-organisms by the dilution method is investigated.

Finally it is a pleasure to thank Miss W.A. Mackenzie, for her valuable assistance in the preparation of the diagrams.

References

(1) K. Pearson (1920). "The Fundamental Problem of Practical Statistics," 'Biom.,' xiii., pp. 1–16.
(2) F.Y. Edgeworth (1921) "Molecular Statistics," 'J.R.S.S.,' lxxxiv., p. 83.
(3) G.U. Yule (1912). "On the Methods of Measuring Association between two Attributes." 'J.R.S.S.,' lxxv., p. 587.
(4) Student (1908). "The Probable Error of a Mean," 'Biom.,' vi., p. 1.
(5) R.A. Fisher (1915). "Frequency Distribution of the Values of the Correlation Coefficient in Samples from an Indefinitely Large Population," 'Biom.,' x., 507.
(6) R.A. Fisher (1921). "On the 'Probable Error' of a Coefficient of Correlation deduced from a Small Sample," 'Metron.,' i., pt. iv., p. 82.

(7) R.A. Fisher (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.,' lxxx., 758.
(8) E. Pairman and K. Pearson (1919). "On Corrections for the Moment Coefficients of Limited Range Frequency Distributions when there are finite or infinite Ordinates and any Slopes at the Terminals of the Range," 'Biom.,' xii., p. 231.
(9) R.A. Fisher (1912). "On an Absolute Criterion for Fitting Frequency Curves," 'Messenger of Mathematics,' xli., p. 155.
(10) Bayes (1763). "An Essay towards Solving a Problem in the Doctrine of Chances," 'Phil. Trans.,' liii., p. 370.
(11) K. Pearson (1919). "Tracts for Computers. No. 1: Tables of the Digamma and Trigamma Functions," By E. Pairman, Camb. Univ. Press.
(12) K. Smith (1916). "On the 'best' Values of the Constants in Frequency Distributions," 'Biom.,' xi., p. 262.
(13) K. Pearson (1914). "Tables for Statisticians and Biometricians," Camb. Univ. Press.
(14) G. Vega (1764). "Thesaurus Logarithmorum Completus," p. 643.
(15) K. Pearson (1900). "On the Criterion that a given System of Deviations from the Probable in the case of a Correlated System of Variables is such that it can be reasonably supposed to have arisen from Random Sampling," 'Phil. Mag.,' l., p. 157.
(16) K. Pearson and L.N.G. Filon (1898). "Mathematical Contributions to the Theory of Evolution. IV.—On the Probable Errors of Frequency Constants, and on the influence of Random Selection on Variation and Correlation," 'Phil. Trans.," cxci., p. 229.
(17) R.A. Fisher (1922). "The Interpretation of χ^2 from Contingency Tables, and the Calculation of P," 'J.R.S.S.,' lxxxv., pp. 87–94.
(18) K. Pearson (1915). "On the General Theory of Multiple Contingency, with special reference to Partial Contingency," 'Biom.,' xi., p. 145.
(19) K. Pearson (1903). "On the Probable Errors of Frequency Constants," 'Biom.,' ii., p. 273, Editorial.
(20) W.F. Sheppard (1898). "On the Application of the Theory of Error to Cases of Normal Distribution and Correlations," 'Phil. Trans.,' A., cxcii., p. 101.
(21) J. M. Keynes (1921). "A Treatise on Probability," Macmillan & Co., London.

Source: Fisher, R. A. 1992. On the Mathematical Foundations of Theoretical Statistics. Pages 11-44 *In*: Kotz, S., and N.L. Johnson, editors. *Breakthroughs in Statistics Volume 1. Foundations and Basic Theory.* Springer Series in Statistics, Perspectives in Statistics. Springer-Verlag: New York.