LARGE-SAMPLE THEORY: PARAMETRIC CASE[1]

By Herman Chernoff

Stanford University

1. Introduction. Large-sample theory is a branch of statistics which seems to have developed because the existence of certain theorems in the theory of probability made it relatively easy to obtain good approximate results if the sample size is large. These theorems, like the law of large numbers and the central limit theorem, are extremely elegant, and frequently their elegance is captured by these "easily" obtained results. This elegance has undoubtedly stimulated a great many people to do work in statistics.

However, since one is seldom faced with an infinite sample, it is relevant to ask whether asymptotic results are useful, and if so, where. In particular, one is often asked whether a given sample size is large enough to justify the use of asymptotic results. Frequently this question is embarrassing, and no answer is available simply because the answer would involve the solution of the more difficult finite-sample-size problem and the use of nonexistent related tables. In some cases, where this question has been treated, it has been shown that these asymptotic results are very good approximations. One example is the study wherein it was shown that the chi-square goodness-of-fit statistic has approximately the chi-square distribution for rather small sample sizes [1].

Even though results of this type are not available for a particular problem, the study of the large-sample case could be justified on other grounds. Asymptotic solutions of a problem frequently give insight into what constitutes a reasonable procedure for the finite-sample-size case. Everyone who has had the experience of seeing how obvious the solution to a certain problem is after spending hours deriving it can appreciate how suggestive an asymptotic result can be for the finite-sample-size problem. For somewhat similar reasons, the method of maximum likelihood estimation, which has various good large-sample properties, has become extremely popular, even for small samples. In fact, a glance at the literature gives the impression that the property of being a maximum likelihood estimate has almost been adopted as the criterion of optimality.

In this paper we deal with the parametric case. Ordinarily this is assumed to mean that our observations come from a population whose distribution is specified by the value of a parameter θ, which may be a k-dimensional vector. A specific problem would be that of testing whether two normal populations with the same variance have the same mean. It seems that once more we must face the fact that our problems may not reflect reality completely. There is a considerable class of problems for which the parametric formulation is more than a

[1] Presented as a special invited address at the Annual Meeting of the IMS in Berkeley, California, December 27, 1954. This work was prepared with the partial support of the Office of Naval Research.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Mathematical Statistics.* www.jstor.org

convenient and very rough approximation. On the other hand, there is a considerable class for which this is not so. Even in these cases the same sort of reasoning which was advanced to advocate the study of large samples results is *apropos* to justify the study of parametric theory, and even of its application to problems where the parametric formulation seems quite rough.

Another point of some interest is that the normal distribution, on occasion, plays the role of a worst distribution. In such cases one may obtain quasi-maximum likelihood estimates, i.e., estimates derived by the use of maximum likelihood on the not necessarily correct assumption that certain random variables are normally distributed. These estimates may be inefficient compared with the true maximum likelihood estimates. Still, these quasi-maximum likelihood estimates have the same or as good asymptotic distributions as they would have were the assumptions of normality correct. They also have the advantage that their computation does not involve the knowledge of the true distribution of these variables. Some complex examples are treated in [2]. A trivial example which illustrates this point is the following. On the basis of a sample of n independent observations, estimate the mean of the population when it is assumed to be normal and it really is rectangular. Here, $\bar{X}$ is the quasi-maximum likelihood estimate and the true maximum likelihood estimate is $\hat{\mu}$, the average of the smallest and largest observations. The asymptotic distribution of $\sqrt{n}(\bar{X} - \mu)$ is normal with mean 0 and variance σ^2 (the variance of the population), whether the population is normal or rectangular. However, if it is rectangular, $\hat{\mu}$ will be considerably more efficient.

This paper will be divided into two main parts. In the first I shall summarize several techniques and results which are useful tools in the study of large-sample theory and which, I feel, have been unfortunately neglected in the literature. In the second part I shall consider some results in inference in the large-sample parametric case. There, much of the space will be devoted to material which has been of special interest to me. In this way I hope to communicate some of my outlook rather than merely to present a long list of accomplishments.

Part I

2. Stochastic limit and order relationships. The title of this section is taken from that of a paper of Mann and Wald [3]. Their stated purpose was to provide readers with certain general results which would eliminate the necessity on the part of future authors of laboriously proving special cases, not to mention confusing the readers. This aim seems to have been largely frustrated mainly by the fact that the paper was practically forgotten. I wish to discuss some of these general results and notations and some useful generalizations of these.

In standard notation one writes $a_n = O(r_n)$ if $\{a_n\}$ is a sequence of real numbers and $\{r_n\}$ is a sequence of positive numbers such that a_n / r_n is bounded. If $a_n / r_n \to 0$ as $n \to \infty$, one writes $a_n = o(r_n)$. This notation is frequently convenient and suggestive. For example, if $a_n \to 0$ and b_n is bounded, it follows that $a_n b_n \to 0$. This may be simply written as follows: $o(1)\ O(1) = o(1)$.

An analogous notation may be defined for sequences of chance variables $\{x_n\}$. We may write $x_n = O_p(r_n)$ (x_n / r_n is bounded in probability) if for each $\epsilon > 0$ there is an M_ϵ and an N_ϵ such that

$$\Pr\{|x_n| \geqq M_\epsilon r_n\} \leqq \epsilon \qquad \text{for } n \geqq N_\epsilon .$$

Finally, we may write $x_n = o_p(r_n)$ ($|x_n| / r_n$ approaches zero in probability) if

$$\Pr\{|x_n| \geqq \epsilon r_n\} \to 0 \qquad \text{as } n \to \infty \qquad \text{for each } \epsilon > 0.$$

It might be well to note here that these concepts are easily extendable to the case where x_n is not necessarily a real chance variable, but where x_n may take on values in an arbitrary space on which an "absolute value" is defined.

One of the results obtained by Mann and Wald is part of their Corollary 1, which states essentially that the algebra of o and O extends to o_p and O_p. A paraphrase of this result, which I have found very useful, is due to John Pratt and is stated as follows: Suppose that $\{x_n\}$ is a sequence of chance variables defined on an arbitrary space. Let $\{g_n(x_n)\}$ and $\{f_n^{(j)}(x_n)\}$, $j = 1, 2, \cdots, k$, be $k + 1$ sequences of measurable functions, and let $\{r_n\}$ and $\{r_n^{(j)}\}$ be $k + 1$ sequences of positive numbers.

THEOREM 1. *Suppose that*

$$(1) \qquad \begin{aligned} f_n^{(j)}(x_n) &= O_p(r_n^{(j)}), && j = 1, 2, \cdots, k_1, \\ f_n^{(j)}(x_n) &= o_p(r_n^{(j)}), && j = k_1 + 1, \quad k_1 + 2, \cdots, k, \end{aligned}$$

and that

(2) *for any (nonrandom) sequence* $\{a_n\}$ *for which*

$$f_n^{(j)}(a_n) = O(r_n^{(j)}), \qquad j = 1, 2, \cdots, k_1,$$

and

$$f_n^{(j)}(a_n) = o(r_n^{(j)}), \quad j = k_1 + 1, \quad k_1 + 2, \cdots, k,$$

hold, it follows that $g_n(a_n) = O(r_n)$.

Then, it follows that $g_n(x_n) = O_p(r_n)$. *Furthermore, if the last line of* (2) *is replaced by* $g_n(a_n) = o(r_n)$, *the conclusion is* $g_n(x_n) = o_p(r_n)$.

The following are some examples which may serve to illustrate the use of this result.

EXAMPLE 1. If $y_n \xrightarrow{p} y$, i.e., if y_n approaches y in probability, or $y_n - y = o_p(1)$, and if $z_n \xrightarrow{p} z$, then $y_n z_n \xrightarrow{p} yz$. This result follows because we are given, on the one hand, that $y_n - y = o_1(1)$ and $z_n - z = o_p(1)$. On the other hand, it is easy to prove (and is well known) that $b_n - b = o(1)$ and $c_n - c = o(1)$ (i.e., that $b_n \to b$ and $c_n \to c$) imply that $b_n c_n - bc = o(1)$. Consequently,

$$y_n z_n - yz = o_p(1).$$

Several remarks may be made about this example. It may seem to involve a tremendous amount of machinery for a very simple result. In fact, a direct proof

may seem to be no more difficult than the "on-the-other-hand" part. Actually, my own experience in class has shown that the direct proof is usually instructive because students find it so difficult. The tremendous machinery is not so tremendous if this approach is used frequently, for then it becomes standard. Finally, this example illustrates how this approach clearly separates the nonstochastic asymptotic elements of a problem from the stochastic elements.

One point which may have not been thoroughly clarified in the above exposition is the specification of x_n, $f_n^{(j)}$, g_n, and a_n in this example. To be perfectly specific, we may let

$$x_n = (y_n, z_n, y, z); \qquad f_n^{(1)}(x_n) = y_n - y;$$
$$g(x_n) = y_n z_n - yz; \qquad f_n^{(2)}(x_n) = z_n - z;$$
$$a_n = (b_n, c_n, b, c).$$

EXAMPLE 2. If $x_n = o_p(1)$, it follows that $\sin x_n/\sqrt{x_n} = o_p(1)$. All that needs to be shown is that $\sin a_n/\sqrt{a_n} \to 0$ if $a_n \to 0$.

EXAMPLE 3. The following is the simplest of several results which concern Taylor Series Expansions.

COROLLARY 1. *If*

$$x_n = a + o_p(r_n), \tag{1}$$

where $r_n \to 0$, *and*
(2) $f(x)$ *has* s *continuous derivatives at* $x = a$, *then*

$$f(x_n) = f(a) + (x_n - a)f'(a) + \cdots + \frac{(x_n - a)^s f^{(s)}(a)}{s!} + o_p(r_n^s).$$

The following is a considerably more sophisticated example. Here the separation of stochastic and nonstochastic elements is a blessing, for the problem is not completely trivial under the best of circumstances.

EXAMPLE 4. Suppose that $x_1, x_2, \cdots, x_n$ are n independent observations on a chance variable with density

$$f(x \mid \alpha, \beta, \gamma) = \beta \qquad \text{for } 0 \leqq x \leqq \alpha,$$
$$f(x \mid \alpha, \beta, \gamma) = \gamma \qquad \text{for } \alpha < x \leqq 1,$$

where $\alpha\beta + (1 - \alpha)\gamma = 1$, $0 < \alpha < 1$, $\beta > 0$, $\gamma > 0$, and $\beta \neq \gamma$. It is not difficult to show that the maximum likelihood estimate $\hat{\alpha}_n$ of α maximizes

$$\left[\frac{F_n(\alpha)}{\alpha}\right]^{F_n(\alpha)} \left[\frac{1 - F_n(\alpha)}{1 - \alpha}\right]^{1 - F_n(\alpha)},$$

where F_n is the sample c.d.f.; i.e., $F_n(x)$ is $1/n$ times the number of observations less than or equal to x. (Note that the function F_n is itself random.) We may write $\hat{\alpha}_n = \Phi(F_n)$. A proof of the consistency of $\hat{\alpha}_n$ (i.e., that $\hat{\alpha}_n \xrightarrow{p} \alpha_0$ if α_0 is the true value of the parameter) is partially complicated by the possibility that

$\hat{\alpha}_n$ may get close to zero or one. We shall merely outline the proof that $\hat{\alpha}_n$ is bounded away from zero in probability, i.e., $1/\hat{\alpha}_n = O_p(1)$.

First, it is known that

$$\sup_{0 \leqq x \leqq 1} | F_n(x) - F_0(x) | = o_p(1), \tag{1}$$

where $F_0(x)$ is the true c.d.f., and it can be shown that

$$\sup_{0<x<1} \left| \frac{F_n(x)}{x} \right| = O_p(1). \tag{2}$$

Secondly, it can be shown that if $\{G_n\}$ is a sequence of nonrandom c.d.f.'s such that

$$\sup_{0 \leqq x \leqq 1} | G_n(x) - F_0(x) | = o(1), \tag{1'}$$

$$\sup_{0<x<1} \left| \frac{G_n(x)}{x} \right| = O(1), \tag{2'}$$

then $1/\Phi(G_n) = O(1)$. It follows that $1/\hat{\alpha}_n = O_p(1)$.

One may observe that the role of x_n in Theorem 1 is played here by the sample c.d.f., F_n.

Another important consideration in the Mann-Wald paper involves a generalization of a well-known result which states that if x_n has a limiting distribution, then for a continuous function g, $g(x_n)$ has the corresponding limiting distribution. Hence, if x_n is asymptotically normally distributed with mean 0 and variance 1, x_n^2 has an asymptotic chi-square distribution with one degree of freedom This result was generalized to allow for the possibility that g has points of discontinuity. Unfortunately, through an oversight, a slightly weaker result than could have been obtained was presented. The stronger version will be stated after we introduce some appropriate notation.

We write $\mathcal{L}(x_n) \to \mathcal{L}(x)$ (read: the distribution law of x_n converges to the distribution law of x) or $\lim_{n\to\infty} \mathcal{L}(x_n) = \mathcal{L}(x)$ if $F_n(a) \to F(a)$ at every point a of continuity of F, where F_n and F are the c.d.f.'s of x_n and x, respectively. Here, $\mathcal{L}(x_n)$ and $\mathcal{L}(x)$ represent the probability measures associated with x_n and x. Let $D(g)$ be the set of discontinuities of the function g.

THEOREM 2. *If*

$$\mathcal{L}(x_n) \to \mathcal{L}(x) \tag{1}$$

and

$$\mathcal{L}(x; D(g)) \equiv P\{x \in D(g)\} = 0, \tag{2}$$

then

$$\mathcal{L}[g(x_n)] \to \mathcal{L}[g(x)].$$

EXAMPLE 1. If $\mathcal{L}(x_n, y_n) \to \mathcal{L}(x, y)$, where x and y are independently and normally distributed with mean 0 and variance 1, then $\mathcal{L}(x_n/y_n) \to \mathcal{L}(x/y)$, which is a Cauchy distribution.

This theorem was extended by Rubin [4] to the case where x_n and x take on values in a topological space X. Here, the notion of convergence in distribution law must be extended. Rubin uses the following definition:[2]

$$\mathcal{L}_n \to \mathcal{L} \quad \text{if for every closed set} \quad S, \qquad \mathcal{L}(S) \geqq \limsup_{n \to \infty} \mathcal{L}_n(S)$$

or, equivalently,

$$\mathcal{L}_n \to \mathcal{L} \quad \text{if for every open set} \quad S, \qquad \mathcal{L}(S) \leqq \liminf_{n \to \infty} \mathcal{L}_n(S).$$

For many spaces, in particular for metric spaces, this definition coincides with the following one used by other authors [5], [6]: $\mathcal{L}_n \to \mathcal{L}$ if for every bounded continuous function h,

$$\int h(x)\, d\mathcal{L}_n(x) \to \int h(x)\, d\mathcal{L}(x).$$

Both of these are extensions of the definition for Euclidean spaces. With Rubin's definition, it follows rather easily that Theorem 2 applies whenever g is a measurable transformation from one topological space into another.

Rubin [7] has applied this result to find the limiting distribution of quasi-maximum likelihood estimates of the parameters of certain sets of simultaneous linear stochastic difference equations. Donsker [8] derived a related result while engaged in the justification of a heuristic derivation of the asymptotic distribution of Kolmogorov-Smirnov statistic given by Doob [9]. It is interesting to note that in terms of our Theorem 2, Doob's paper dealt mainly with finding the distribution of $g(x)$, after indicating that it seemed reasonable to expect that in some sense $\mathcal{L}(x_n) \to \mathcal{L}(x)$. There the role of x_n was played by the sample c.d.f. in the Kolmogorov-Smirnov problem.

The above exposition is far from complete. For example, the following result for Euclidean spaces is rather useful. Note that it can be reworded so as to be extended to metric spaces.

THEOREM 3. *If* $\mathcal{L}(x_n) \to \mathcal{L}(x)$, *then* $\mathcal{L}(x_n + o_p(1)) \to \mathcal{L}(x)$.

Furthermore, it seems to me that there still remains some work to be done with a view to making the application of Theorems 1 and 2 more cut and dried. Finally, it should be remarked that direct derivations which do not separate the stochastic and asymptotic elements of the problem are sometimes simpler and neater than the techniques suggested by the above results.

3. The Cramér extension of the central limit theorem. In 1938, Cramér [10] obtained an elegant extension of the central limit theorem which, for some reason, seemed to have been overlooked by statisticians. This seems to have been unfortunate, since it appears to be more relevant than the central limit theorem in many statistical applications.

The central limit theorem is loosely described as follows. The average $\bar{X}_n$

[2] The Borel field associated with the distributions is assumed to be that generated by the closed sets.

of n observations on a chance variable X is approximately normally distributed. More precisely,

$$\Pr\left\{\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leqq a\right\} \to \int_{-\infty}^{a} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt \qquad \text{as } n \to \infty,$$

if $\bar{X}_n$ is the average of n independent observations on a chance variable with mean μ and variance σ^2. Suppose, now, that a is not fixed but is replaced by a_n, where $a_n \to -\infty$ as $n \to \infty$. Then both sides of the above expression would approach zero. In this sense, the above equation could be considered to be still valid. Even so, it is of importance, as we shall see in Section 6, to determine how fast each side approaches zero and whether the two sides are asymptotically equivalent, i.e., whether the ratio of the two terms approaches one.

In fact, Cramér has essentially shown that as long as a_n does not approach $-\infty$ too rapidly, the two sides are roughly equivalent. However, this result fails to hold when a_n is of the order of magnitude of $\sqrt{n}$. Note that if $a_n = -b\sqrt{n}$, $b > 0$, we are essentially interested in $\Pr\{\bar{X} \leqq c\}$, where $c < \mu$. This case is an especially important one. Here, it is shown that, roughly speaking, $\Pr\{\bar{X} \leqq c\} \approx m^n$, where $m = \inf_t E\{e^{t(X-c)}\}$.

A result of Esseen [11] permits us to eliminate one of the conditions which Cramér had to apply, and which led him to obtain weaker results for the case where the chance variable is discrete. We shall state a version of Cramér's result.

THEOREM 1. *If* $E(e^{tX}) < \infty$ *in some neighborhood of* $t = 0$, *and if* $a_n < -1$, *and* $a_n = o(\sqrt{n})$, *then*

$$\Pr\left\{\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \leqq a_n\right\} = \left[\int_{-\infty}^{a_n} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt\right] \cdot \exp\left[\frac{a_n^3}{\sqrt{n}}\,\lambda\left(\frac{a_n}{\sqrt{n}}\right)\right] \cdot \left[1 + o\left(\frac{a_n}{\sqrt{n}}\right)\right],$$

where $\lambda(t)$ is an analytic function of t whose coefficients depend on the moments of X.

A similar result is obtainable for the case where a_n is positive. Note that

$$(a_n^3/\sqrt{n})\lambda(a_n/\sqrt{n})$$

may become large if a_n is larger in magnitude than $n^{1/6}$. However, this term contributes a relatively unimportant amount compared with the normal approximation term which is asymptotically equivalent to $(\sqrt{2\pi}a_n)^{-1} \exp(-a_n^2/2)$.

THEOREM 2. *If* $E(e^{tX}) < \infty$ *for* t *in some neighborhood of* 0 *and* $c \leqq E(X)$, *then*

$$P\{\bar{X}_n \leqq c\} = \frac{1}{\sqrt{n}}\,m^n\left[b_0 + \frac{b_1}{n} + \cdots + \frac{b_{k-1}}{n^{k-1}} + O\left(\frac{1}{n^k}\right)\right],$$

where $b_0 > 0$ *and* $m = \inf_t E(e^{t(X-c)})$; the quantities b_i depend on c, and k is an arbitrary positive integer.

Cramér's results were generalized by Feller [12] for the case where the observations do not necessarily have the same distribution.

Part II

4. Estimation. The development of the large-sample theory of estimation was given great impetus with the publication by Fisher [13], [14] of his works on estimation, where he proposed the method of maximum likelihood and suggested, among others, the concepts of consistency, efficiency, and sufficiency. The importance of the notions Fisher developed was soon recognized and the method of maximum likelihood became very popular among statisticians. However, these notions and the properties of the method of maximum likelihood were somewhat more complicated than Fisher or his immediate followers realized. Consequently, many proofs dealing with the properties of these estimates were found to be in error. Considerable light was thrown on these complications when J. L. Hodges, Jr., produced an example of superefficiency. This concept was later treated by Le Cam [15], who also presented an excellent historical survey of the field of maximum likelihood estimation. We shall discuss these notions very briefly, referring the reader to Le Cam's paper for a more detailed discussion.

Let X be a chance variable whose distribution is determined by the value of a parameter θ which is assumed to be in a prescribed set Ω. For the purpose of large-sample theory, Fisher defines an estimate T as a sequence of functions $\{T_n = T_n(X_1, X_2, \cdots, X_n)\}$, where $T_n(X_1, \cdots, X_n)$ represents the "estimated" point of Ω when a sample $X_1, \cdots, X_n$ of n independent observations on X are observed.[3]

Definition 1. T is consistent if $T_n(X_1, \cdots, X_n) \to \theta$ in probability as $n \to \infty$.

Suppose that the distribution of X is characterized by the density $f(x, \theta)$. Then, an estimate T^* is a maximum likelihood estimate of θ if $\prod_{i=1}^n f(X_i, \theta)$ assumes its maximum value at $\theta = T_n^*(X_1, X_2, \cdots, X_n)$. (In most applications, the class of distributions may be represented by densities with respect to some σ-finite measure.) It may turn out that the maximum likelihood estimate does not exist. For example, there will be no such estimate for the mean μ of a normal distribution if it is assumed that μ is in the *open* interval $(-1, 1)$ and that the sample mean is greater than one.

When Fisher introduced the notion of asymptotic efficiency, he did this for the case where θ was assumed to be on the real line. Then T was said to be asymptotically efficient if its asymptotic distribution (when properly normalized) was normal with no larger variance than that obtained for any other consistent asymptotically normally distributed statistic. (The variance of the asymptotic distribution will be called the asymptotic variance and is, in general, no larger than the limit of the variance of the normalized estimate.) Apparently, the restriction to asymptotically normally distributed statistics was felt necessary, because Fisher had no way of comparing two dissimilar limiting distributions.

[3] The extension of this notion to the case where the observations need not be independent nor identically distributed is rather evident and we shall not formally treat of that case here.

Fisher and various followers claimed that under suitable mild restrictions the maximum likelihood estimates were consistent and efficient. That the attempts to establish efficiency with the above definition would encounter grave difficulties seems clear when an example of superefficiency is given. Le Cam's example is that of observations from a normal population with unknown mean μ and variance 1. Let T_n represent the maximum likelihood estimate which is the mean of n observations and let T_n' be defined as follows:

$$T_n' = T_n \qquad \text{if} \quad |T_n| \geqq \frac{1}{n^{1/4}},$$

$$T_n' = \alpha T_n \qquad \text{if} \quad |T_n| < \frac{1}{n^{1/4}},$$

where α is an arbitrary constant. Then it is clear that[4]

$$\mathfrak{L}\{\sqrt{n}(T_n - \mu)\} \to N(0, 1),$$

while

$$\mathfrak{L}\{\sqrt{n}(T_n' - \mu)\} \to N(0, 1) \qquad \text{if } \mu \neq 0,$$

but

$$\mathfrak{L}\{\sqrt{n}(T_n' - \mu)\} \to N(0, \alpha^2) \qquad \text{if } \mu = 0.$$

Hence, if $0 < \alpha^2 < 1$, T_n' is asymptotically normally distributed with asymptotic variance which is never larger, and sometimes smaller, than that of T_n. Let us call the set of θ, on which a statistic T_n' is more "efficient" than the maximum likelihood estimate T_n, the set of superefficiency. Le Cam has shown under certain conditions that a set of superefficiency must have Lebesgue measure zero. In this sense the maximum likelihood estimate is *efficient.*

In his paper Le Cam makes use of Wald's decision-theory formulation [16] of the estimation problem. (Similar techniques were independently applied by Wolfowitz [17].) His definition of efficiency and superefficiency involves the loss function $L_n(t, \theta)$, which is introduced to represent the loss to the statistician when he observes a sample of size n and estimates t, while θ is the true value of the parameter. Le Cam derives and uses the properties of Bayes' estimates in his attack. I wish to indicate an alternative approach which yields somewhat weaker results but which will be useful to us later. We may assume that L is measured in terms of negative utility [18], so that it makes sense to attempt to select T so as to minimize the "risk" or expected loss $E\{L_n(T_n, \theta)\}$. Then, corresponding to an estimate T, we have a sequence of risk functions

$$R_n(T_n, \theta) = E\{L_n(T_n(X_1, \cdots, X_n), \theta)\}.$$

This formulation permits us to compare estimates which are (1) not necessarily confined to the real numbers and (2) do not necessarily have similar distribu-

[4] $N(0, 1)$ represents the normal distribution with mean 0 and variance 1.

tions. However, a difficulty appears. First of all, it is usually quite difficult to evaluate the loss function that the statistician really faces. On the other hand, in many cases, it is reasonable to assume that $L_n(t, \theta)$ is a minimum at $t = \theta$ and is well behaved near $t = \theta$. Hence, it is often reasonable to assume (in the one-dimensional case) that for t close to θ, $L_n(t, \theta)$ is approximately

$$c_{0n}(\theta) + c_{2n}(\theta)(t - \theta)^2,$$

where $c_{2n}(\theta) > 0$. Intuitively, this would seem to furnish a good excuse for selecting estimates which minimize the second moment about θ. However, some misgivings may arise when we note that $\lim_{n\to\infty} n\{E(T_n - \theta)^2\}$ and the variance of the limiting distribution of $\sqrt{n}(T_n - \theta)$ need not coincide. In extreme cases, it is possible for an estimate to have a very good asymptotic distribution but have infinite variance for each sample size. This estimate would not show up well if we used $E\{(T_n - \theta)^2\}$ as a criterion. In fact, a utility function which satisfies the von Neumann-Morgenstern axioms [18] must be bounded. Hence, $L_n(t, \theta)$ should be taken to be bounded, whereas the above approximation, which may be reasonable for t close to θ, is not. It is difficult to say what is an appropriate criterion without referring to the true $L_n(t, \theta)$. One might propose the asymptotic variance of $T_n - \theta$ (when suitably normalized), but objections could easily be raised against this.

Suppose that one considered estimates T such that

$$T_n - \theta = O_p(1/\sqrt{n}).$$

Let us treat the expectation of the normalized loss function

$$L_n^*(t, \theta) = n\left[\frac{L_n(t, \theta) - c_{0n}(\theta)}{c_{2n}(\theta)}\right],$$

where we assume

$$L_n^*(t, \theta) = n[(t - \theta)^2 + o(t - \theta)^2],$$

and o is assumed to hold uniformly in n as $t \to \theta$. Then,

$$\liminf_{n\to\infty} \frac{E\{L_n^*(T_n, \theta)\}}{nE\left\{\min\left[\left(T_n - \theta\right)^2, \frac{k^2}{n}\right]\right\}} \geqq 1;$$

$$\lim_{k\to\infty} \liminf_{n\to\infty} \frac{E\{L_n^*(T_n, \theta)\}}{nE\left\{\min\left[(T_n - \theta)^2, \frac{k^2}{n}\right]\right\}} \geqq 1.$$

If $\sqrt{n}(T_n - \theta)$ has a limiting distribution with second moment $\sigma^2(\theta)$, it follows that

$$\lim_{k\to\infty} \lim_{n\to\infty} nE\left\{\left[\min\,(T_n - \theta)^2, \frac{k^2}{n}\right]\right\} = \sigma^2(\theta)$$

and the asymptotic variance $\sigma^2(\theta)$ may be regarded as a lower bound for the normalized risk function.

On the other hand, if $P\{|T_n - \theta| > k\} = o(1/n)$ for each k, it is possible to show that

$$\liminf_{n\to\infty} \frac{nE\{(T_n - \theta)^2\}}{E\{L_n^*(T_n, \theta)\}} \geqq 1,$$

and then the normalized risk is sandwiched between the real variance (normalized) and the asymptotic variance. A similar discussion is given by Hodges and Lehmann [19].

I believe that without unreasonable modifications the standard derivations of the asymptotic normal distribution of the maximum likelihood estimates can be used to show that for the maximum likelihood estimates $\lim_{n\to\infty} E\{L_n^*(T_n, \theta)\}$ is *equal to the asymptotic variance*. As far as I know, no such proof exists yet in the literature.

The above discussion extends easily to the k-dimensional parameter case where the role of the asymptotic variance is played by an expression of the form $\sum_{i,j} a_{ij}(\theta)\sigma_{ij}(\theta)$ Here $A = \|a_{ij}\|$ is a nonnegative symmetric matrix whose elements correspond to the second-order partial derivatives of the loss function at θ (provided that these derivatives or their ratios converge as $n \to \infty$), and $\|\sigma_{ij}(\theta)\|$ is the asymptotic covariance matrix.

A technique that had been used in previous attempts to establish the efficiency of maximum likelihood estimates was the derivation of a lower bound for the variance of an estimate and the proof that this lower bound was "asymptotically" attained by the maximum likelihood estimates.

Results in this direction were apparently first obtained by Fréchet [20] and Darmois [21] and later given by Cramér [22] and Rao [23] and called the Cramér-Rao inequality. Savage [24] has tentatively suggested alternatively using the name "Information inequality". These results were extended in various directions by Bhattacharya [25], [26], Barankin [27], Wolfowitz [28], Seth [29], Chapman and Robbins [30], Kiefer [31], and Fraser and Guttman [32]. Because these results invoked regularity conditions on the estimates, the possibility of superefficiency was hidden. Let us consider the following form of this result which does not use regularity conditions on the estimates. (This form and a variant of it were communicated to me by Charles Stein and Herman Rubin, respectively.)

First, we consider the *nonasymptotic* case where the parameter space Ω is a subset of the real line containing the origin as an inner point. Let us define Fisher's information by

$$I(\theta) = E_\theta\left[\left(\frac{\partial \log f(X, \theta)}{\partial\theta}\right)^2\right],$$

where E_θ represents expectation with respect to the distribution determined by θ. We digress slightly to point out that $I(\theta)$ is additive. That is, if several inde-

pendent observations are combined, the corresponding information is the sum of the individual informations. In particular, when n independent observations are taken on a chance variable X, the information is multiplied by n.

Now let $T(X)$ be an estimate based on the observation X. Under mild conditions on the distribution of X (and not on T) we have

LEMMA 1. *For every* ϵ, $0 < \epsilon < 1$, *and any estimate* T,

$$\sup_{-a<\theta<a} [E_\theta\{(T(X) - \theta)^2\}I(\theta)] \geqq 1 - \epsilon$$

if

$$\frac{1}{2a}\int_{-a}^{a} \frac{d\theta}{I(\theta)} \leqq \frac{a^2\epsilon^2}{12(1-\epsilon)}.$$

Otherwise,

$$\sup_{-a<\theta<a} E_\theta\{(T(X) - \theta)^2\} \geqq \frac{a^2\epsilon^2}{4}.$$

This result can be applied to the large-sample case. To deal with estimates which may behave well asymptotically, but which may have large or even infinite variances, we introduce the truncated estimate T^*,

$$T^* = T \text{ if } |T| \leqq 2a; \qquad T^* = 2a \text{ if } T > 2a; \qquad T^* = -2a \text{ if } T < -2a.$$

Since $\min [(T - \theta)^2, 16a^2] \geqq (T^* - \theta)^2$ for $-a < \theta < a$, we can easily derive the following theorem for the case of n independent observations on X.

THEOREM 1.

$$\lim_{k\to\infty} \liminf_{n\to\infty} \sup_{-k/4\sqrt{n}<\theta<k/4\sqrt{n}} E_\theta\left\{nI(\theta) \min\left[(t-\theta)^2, \frac{k^2}{n}\right]\right\} \geqq 1$$

if $I(\theta)$ *is measurable and bounded away from* 0 *in some neighborhood of* $\theta = 0$. (This statement might be easier to read if it were weakened by replacing, under "sup," the interval $-k/4\sqrt{n} < \theta < k/4\sqrt{n}$ by $-\delta < \theta < \delta$.)

This result clearly allows for the possibility of superefficiency. It is weaker than Le Cam's results, since it does not confine superefficiency to a set of measure zero. On the other hand, this statement fits in very well with our discussion of the normalized risk functions. It states that for an arbitrary estimate *the reciprocal of the information is "essentially" asymptotically a lower bound for the asymptotic variance and hence for the normalized risk function.* This, together with the above-mentioned conjecture that for the maximum likelihood estimate, the normalized risk approaches the asymptotic variance (which coincides with the reciprocal of the information), would give the "essential" efficiency of the maximum likelihood estimate from the normalized risk-function point of view.

Theorem 1 also has the advantage that it can be easily extended to the case where the independent observations are not necessarily from the same population. If the average information per observation is given by

$$\bar{I}_n(\theta) = \frac{1}{n}[I_1(\theta) + \cdots + I_n(\theta)],$$

where $I_j(\theta)$ is the information corresponding to the jth observation or experiment, we can replace $I(\theta)$ in Theorem 1 by $\bar{I}_n(\theta)$, provided $\bar{I}_n(\theta)$ is measurable and

$$\lim_{n\to\infty} \inf_{-k/\sqrt{n}<\theta<k/\sqrt{n}} \bar{I}_n(\theta) > 0$$

for each k. The case where

$$\lim_{n\to\infty} \sup_{-k/\sqrt{n}<\theta<k/\sqrt{n}} \bar{I}_n(\theta) = 0$$

for each k gives no difficulty.

5. Optimal designs for estimating parameters. Suppose that there is available a class of experiments $\{E\}$. A design will consist of a selection of n of these experiments to be performed independently. Suppose that the outcome of each experiment depends only on a real-valued parameter θ which is to be estimated. We shall assume that the true value of θ is approximately known so that it makes sense to consider locally optimal designs. That is to say, that we shall be interested in selecting n experiments so that an estimate of θ, based on the outcomes, will be very good if θ is close to some specified value θ^0.

If n is large, it seems reasonable to select these n experiments, $E_1, E_2, \cdots, E_n$, so as to make the sum of the corresponding informations $\sum_{i=1}^n I(E_i, \theta^0)$ large. If $I(E, \theta^0)$ is maximized by an experiment E_0, it pays to repeat the experiment, E_0, n times. Then, by the Cramér-Rao type of theorem we treated, the asymptotic variance for any design is at least as large as

$$\frac{1}{\bar{I}_n} = \frac{n}{\sum_{i=1}^n I(E_i, \theta^0)} \geqq \frac{1}{I(E_0, \theta^0)},$$

which is the asymptotic variance for the maximum likelihood estimate based on n repetitions of E_0. Furthermore, if the conjecture that for maximum likelihood estimates, the asymptotic variance is equal to the normalized risk is correct, then the normalized risk is asymptotically a minimum for this design.

While the above problem is not very deep, there are certain remarks which are relevant to the extension of this problem to the multidimensional parameter case. First of all, it is quite possible that $I(E, \theta^0)$ does not attain its maximum. A trivial case is the following: Suppose that E_σ corresponds to observing a normal deviate with mean θ and variance σ^2, and suppose that all E_σ are available for $\sigma > 1$. Here, $I(E_\sigma, \theta^0) = 1/\sigma^2$ can be made arbitrarily close to 1 but cannot equal 1. It is apparent that the theoretical difficulty posed by this situation is neither significant nor important.

In general, some experiments are more costly than others, and the formulation involving the selection of a preassigned number of experiments may reasonably be changed to that of selecting an arbitrary number of experiments whose total cost is preassigned. Here, we would attempt to make $\sum I(E_i, \theta^0)$ large,

subject to the restriction $\sum c(E_i) = k$, where $c(E)$ is the cost of the experiment E. Rewriting the above as

$$\sum I(E_i, \theta^0) = \sum \left[\frac{I(E_i, \theta^0)}{c(E_i)}\right] c(E_i),$$

it is evident that we should select that E_0 which maximizes $I(E, \theta^0)/c(E)$, the information per unit cost, and repeat E_0, $k/c(E_0)$ times.

Let us now extend the problem to the following case. Suppose that it is desired to estimate a parameter θ_1, but the distribution of the outcomes of the available experiments depends not only on θ_1, but also on $\theta_2, \cdots, \theta_k$. A special case of this would be that of estimating the slope β of the regression line of Y on x, where $Y = \alpha + \beta x + u$, $\mathcal{L}(u) = N(0, 1)$. Each level x represents an experiment E_x; then let us assume that one has available the set of E_x for which $-1 \leqq x \leqq 1$. It is well known that in this special example the optimal experiment consists of performing E_1 and E_{-1} each half of the time.

To formulate this problem properly, we first note that in the case of k parameters, the information is replaced by the information matrix

$$I(\theta) = \left\| E_\theta \left\{ \frac{\partial \log f(X, \theta)}{\partial \theta_i} \cdot \frac{\partial \log f(X, \theta)}{\partial \theta_j} \right\} \right\|, \qquad i, j = 1, 2, \cdots, k.$$

The information matrix $I(\theta)$ has the additive property; i.e., the information matrix corresponding to the outcome of several independent experiments E_i is equal to the sum of the corresponding information matrices $\sum I(E_i, \theta)$. Another property of interest is the following: Consider the randomized experiment where E_i is performed with probability p_i. Then, the information matrix for the randomized experiment is given by the average $\sum p_i I(E_i, \theta)$.

Let $I_{ij}(\theta)$ represent the (i, j) term of $I(\theta)$ and let $I^{ij}(\theta)$ be the (i, j) term of $I^{-1}(\theta)$. As $1/I(\theta)$ represented the asymptotic variance in the one-dimensional case, so $I^{-1}(\theta)$ represents the asymptotic covariance matrix in the k-dimensional case. In particular, $I^{11}(\theta)$ represents the asymptotic variance of $\sqrt{n}(\hat{\theta}_1 - \theta_1)$.

It now becomes very natural to formulate our problem as being that of selecting n experiments to minimize

$$\left[\sum_{i=1}^{n} I(E_i, \theta^0)\right]^{11}.$$

We may equivalently minimize the (1, 1) element of the inverse of the average information per observation, i.e., we minimize

$$\bar{I}_n^{11}(\theta) = \left[\frac{1}{n}\sum_{i=1}^{n} I(E_i, \theta^0)\right]^{11}.$$

Now the expression on the right-hand side corresponds to the randomized experiment where each E_i is performed with probability $1/n$. By taking n large enough, we can approximate each randomized experiment arbitrarily closely. Hence, we might reformulate our problem as that of selecting that randomized experiment for which $I(E, \theta^0)^{11}$ is minimized.

Each information matrix is nonnegative definite symmetric and may be identified with the point in $k(k+1)/2$-dimensional space whose coordinates are the elements on and below the main diagonal of the matrix. The class of matrices corresponding to the randomized experiments is the convex set generated by the matrices of the pure experiments. Hence, our problem reduces to that of minimizing a function on a convex set.

I^{11} is a continuous function of I on the set of positive definite symmetric matrices However, I^{11} is not defined for singular matrices. If the distribution of the outcome of an experiment E depended on less than k independent parameters, the information matrix would be singular. Nevertheless, in this case, it can be shown that it would be meaningful to redefine I^{11} by $\lim_{\lambda \to 0+} (I + \lambda A)^{11}$, where A is an arbitrary positive definite symmetric matrix. We then have [33].

THEOREM 1. *If the set R of randomized information matrices $I(\theta^0)$ is closed and bounded, the function $I^{11}(\theta^0)$ attains its maximum on R at a matrix which is a convex combination of $r \leqq k$ of the information matrices corresponding to the nonrandomized experiments.*

This theorem states that there is a locally optimal design for large n which involves at most k of the original experiments. This result considerably reduces the computational problem involved in computing the optimal design. It constitutes a generalization of a similar result by Elfving [34], which applies to linear regression problems with normal deviates. In connection with his result, Elfving indicated an elegant geometrical technique of finding the optimal solution. His technique applies to our more general problem if all the information matrices resemble those of the regression case; i.e., if the typical information matrix for each experiment can be expressed as $\| x_i x_j \|$. In fact, this case occurs quite frequently in applications which are not normal linear regression.

Finally, this result extends to the case where one is interested in estimating s out of the k parameters involved in the experiments. Then the optimal design involves no more than $k + (k-1) + \cdots + (k-s+1)$ experiments. This last result is of limited computational applicability if k and s are not small numbers.

6. Testing simple hypotheses. The easiest problem in statistical inference is that of testing a simple hypothesis against a simple alternative. Suppose that the hypothesis H_0 specifies that n independently distributed observations, $X_1, X_2, \cdots, X_n$, have density $f_0(x)$, whereas the alternative H_1 specifies the density $f_1(x)$. It is well known that the class of best tests are the likelihood ratio tests characterized by critical regions which contain all points where the ratio

$$\prod_{i=1}^{n} f_1(X_i) / \prod_{i=1}^{n} f_0(X_i)$$

exceeds some constant c and a subset of those points for which the ratio is equal to c. It is peculiar that in this example, where the small-sample theory is so well understood, the large-sample theory yields resulst of interest.

First, let us note that the above test can be considered to be one that is based on $\bar{Y}_n = 1/n \sum_{i=1}^{n} Y_i$, where $Y_i = \log f_1(X_i)/f_0(X_i)$. But for tests based on

averages of observations, Cramér's results, which were expressed in Section 3, are applicable. These results also apply to tests which are not necessarily likelihood ratio tests. In what follows, we shall assume that Y_i is not necessarily of the above form, but that the test consists of rejecting H_0 if $\bar{Y}_n > a_n$ and that $\mu_0 = E(Y \mid H_0) < E(Y \mid H_1) = \mu_1$.

The probabilities of the two types of error are given by

$$\alpha_n = P\{\bar{Y}_n > a_n \mid H_0\} \qquad \text{and} \qquad \beta_n = P\{\bar{Y}_n \leqq a_n \mid H_0\}.$$

There are several principles which may be invoked for selecting a_n. One of these is that of minimizing $\alpha_n + \lambda\beta_n$ for some $\lambda > 0$. This principle would be especially meaningful if there were an a priori probability ξ, $0 < \xi < 1$, attached to H_0. Then, if l_{ij} represents the loss due to accepting H_i when H_j is correct, the risk would be given by

$$R = \xi l_{00}(1 - \alpha_n) + \xi l_{10}\alpha_n + (1 - \xi)l_{01}\beta_n + (1 - \xi)l_{11}(1 - \beta_n)$$

$$= \xi l_{00} + (1 - \xi)l_{11} + \xi(l_{10} - l_{00})\left[\alpha_n + \frac{(1 - \xi)(l_{01} - l_{11})}{\xi(l_{10} - l_{00})}\beta_n\right].$$

But for reasonable loss functions, $l_{01} - l_{11}$ and $l_{10} - l_{00}$ are positive. Hence, minimizing R is equivalent to minimizing $\alpha_n + \lambda\beta_n$, where

$$\lambda = \frac{(1 - \xi)(l_{01} - l_{11})}{\xi(l_{10} - l_{00})} > 0.$$

Another situation in which it would be appropriate to use this criterion would be one where it is desired to minimize some function $F(\alpha_n, \beta_n)$, where neither $\partial F(0, 0)/\partial\alpha$ nor $\partial F(0, 0)/\partial\beta$ vanish. Essentially, this boils down to requiring that as $n \to \infty$, α_n and β_n converge to zero at the same rate.

Let

$$m_i(a) = \inf_t E\{e^{t(X-a)} \mid H_i\}, \qquad i = 0, 1,$$

$$\rho(a) = \max [m_0(a), m_1(a)], \qquad \rho = \inf_{\mu_0 \leqq a \leqq \mu_1} \rho(a).$$

A consequence of Cramér's result (see [35]) is

THEOREM 1.

$$\lim_{n\to\infty} [\inf_{a_n} (\beta_n + \lambda\alpha_n)]^{1/n} = \rho \text{ (independent of } \lambda).$$

This theorem permits us to compare the relative efficiency of two tests. For the above test, $\beta_n + \lambda\alpha_n$ behaves roughly like ρ^n. Suppose that a similar test is based on the average of another statistic Z. If ρ^* is the corresponding value of ρ for this new test, then

$$e = \frac{\log \rho^*}{\log \rho}$$

is a reasonable measure of the relative efficiency of the test based on Z to the test based on Y. The reason for this is that if n_1 and n_2 are large sample sizes for which the $\alpha_{n_i} + \lambda\beta_{n_i}$ of the two tests are approximately equal, then n_1/n_2 is close to e. In other words, the first test requires en_2 observations to do as well as the second. This measure of efficiency permits us not only to compare various tests based on a given experiment, but also permits us to compare tests based on different experiments.

In particular, let us consider the likelihood ratio test for a given experiment. We designate the corresponding ρ by ρ_{LR}, which can be shown to be given by

$$\rho_{LR} = \inf_{0<t<1} \int [f_1(x)]^t [f_0(x)]^{1-t} \, d\nu(x)$$

if $f_1(x)$ and $f_0(x)$ are the densities of X, with respect to the measure ν, under H_1 and H_0, respectively. Because of the character of the above-mentioned measure of relative efficiency, it is natural to define the information in the experiment by

$$I = -\log \rho_{LR} .$$

Fisher's measure of information also had the property that if two experiments yield informations $I_1(\theta)$ and $I_2(\theta)$, where $I_1(\theta) = 2I_2(\theta)$, then one needs approximately $2n$ observations on the second experiment to get results comparable to those obtained with n observations on the first experiment for n large. It is interesting to note that while Fisher's measure of information is additive, the above is not. In fact, it has the following properties:

(1) The information derived from n independent observations on a chance variable is n times the information from one observation.

(2) The information derived from observations on several independent chance variables is less than or equal to the sum of the corresponding informations.

It occasionally happens in practice that it is important to obtain β very small, whereas a relatively large value of α, like .05 or .10, is not disastrous. In such cases, it makes sense to consider in our large-sample approach the problem where one minimizes β subject to fixed α. Let β_n^* be the value of β_n which corresponds to a fixed value of α, say α_0, $0 < \alpha_0 < 1$. We have, as another consequence of Cramér's result,

THEOREM 2.

$$\lim_{n\to\infty} \beta_n^{*1/n} = \rho^* = m_1(\mu_0) \qquad (\textit{independent of } \alpha_0),$$

where $\mu_0 = E(Y \mid H_0)$.

In particular, for the likelihood ratio test, it is easy to show that we obtain ρ_{LR}^*, which is given by

$$\rho_{LR}^* = m_1(\mu_0) = e^{\mu_0} = \exp\left[\int f_0(x) \log \frac{f_1(x)}{f_0(x)} \, d\nu(x)\right].$$

This result was first obtained by Charles Stein [36]. Here again, it makes sense to define a corresponding measure of information by

$$I^* = -\log \rho_{LR}^* = -\int \log \left[\frac{f_1(x)}{f_0(x)}\right] f_0(x)\, d\nu(x).$$

It is of interest to note that I^* represents one of the Kullback-Leibler information numbers [37]; also

$$I^{**} = \int f_1(x) \log \frac{f_1(x)}{f_0(x)}\, d\nu(x)$$

would arise naturally if β_n were kept fixed and $\alpha_n \to 0$. The Kullback-Leibler numbers do have the additive property. Incidentally, the above characterization of the Kullback-Leibler numbers implies that they do exceed $I = -\log \rho_{LR}$.

Until now, we have not discussed sequential analysis from a large-sample point of view. At a first naive glance, it may seem as though the very nature of sequential analysis is such as to rule out large-sample theory. That this is not so becomes clear when one considers that reducing the cost of sampling should increase the expected sample size. In fact, let us suppose that the cost per observation is c. Consider the Bayes procedure corresponding to a fixed a priori probability ξ that H_0 is correct. The risk function is given by

$$R_0 = l_{00} + \alpha(l_{10} - l_{00}) + cE(n \mid H_0),$$

$$R_1 = l_{11} + \beta(l_{01} - l_{11}) + cE(n \mid H_1).$$

The Bayes risk, $\xi R_0 + (1 - \xi)R_1$, is minimized by Wald's sequential probability ratio test [38]. As $c \to 0$, $E(n \mid H_0)$ and $E(n \mid H_1) \to \infty$, but

$$\xi(R_0 - l_{00}) + (1 - \xi)(R_1 - l_{11}) \to 0.$$

An elementary application of Wald's inequalities concerning the operating characteristic function gives

THEOREM 3.

$$\lim_{c \to 0} \frac{R_0 - l_{00}}{(c \log 1/c)} = \frac{1}{I^*}; \qquad \lim_{c \to 0} \frac{R_1 - l_{11}}{(c \log 1/c)} = \frac{1}{I^{**}}.$$

Note that these limits do not depend on $l_{10} - l_{00}$ nor on $l_{01} - l_{11}$. This is due to the fact that as $c \to 0$, the main part of the risk is the cost of sampling.

It is rather striking that the notions of information, which are natural for the sequential and nonsequential cases, are not identical. Upon some consideration, however, it is not surprising. In the sequential case, after many observations are taken, one is almost sure which hypothesis is correct. Then if H_0 seems correct, the remaining observations may be selected from an experiment for which the corresponding Kullback-Leibler information I^* is large. In the nonsequential case, the experiment to be performed must be decided on before any data are taken. It is natural that the corresponding information should differ from I^* and I^{**}.

It is of interest to note that as the hypotheses H_0 and H_1 get closer to one another, the three measures of information behave in the following fashion:

$$I \approx \frac{I^*}{4} \approx \frac{I^{**}}{4}.$$

7. Composite hypotheses. A classical result in the large-sample theory applied to tests of composite hypotheses is that of Wilks [39]. It states that[5]

$$\mathfrak{L}(-2 \log \lambda_n) \to \mathfrak{L}(\chi^2_{k-r})$$

if λ_n is the likelihood ratio based on n independent observations for the test that a parameter θ lies on a specified r-dimensional hyperplane of k-dimensional space and the hypothesis is true. It is striking that this result does not involve the distribution of the data except in that mild regularity conditions on the distribution are required.

Many tests of composite hypotheses are not of this simple form. For example, it may be desired to test whether θ lies in the first quadrant of the plane, or it may be desired to test whether θ lies above a hyperplane or even whether θ lies inside a sphere.

For these problems, first suggested to me by Leonid Hurwicz, a natural generalization of Wilks' result is easily obtained.

Let $f(x, \theta)$ represent the density of the data. Suppose that θ lies in k-dimensional space and let ω and τ be two disjoint subsets of this space. We are interested in testing $H_0: \theta \varepsilon \omega$ against the alternative $H_1: \theta \varepsilon \tau$. Let

$$P_\omega(X_1, X_2, \cdots, X_n) = \sup_{\theta \varepsilon \omega} \sum_{i=1}^{n} f(X_i, \theta).$$

The standard definition of the likelihood ratio is given by

$$\lambda_n = \frac{P_\omega(X_1, \cdots, X_n)}{P_{\omega \cup \tau}(X_1, X_2, \cdots, X_n)}.$$

It is somewhat more convenient to treat a more symmetric form

$$\lambda_n^* = \frac{P_\omega(X_1, X_2, \cdots, X_n)}{P_\tau(X_1, X_2, \cdots, X_n)}.$$

These are related by

$$\lambda_n = \lambda_n^* \text{ if } \lambda_n^* \leqq 1; \qquad \lambda_n = 1 \text{ if } \lambda_n^* > 1.$$

We call a set C positively homogeneous if $X \varepsilon C$ implies $aX \varepsilon C$ for all $a > 0$. We say that ω is approximated by a positively homogeneous set C_ω if

$$\inf_{x \varepsilon C_\omega} |x - y| = o(|y|) \qquad \text{for } y \varepsilon \omega,$$

and

$$\inf_{y \varepsilon \omega} |x - y| = 0(|x|) \qquad \text{for } x \varepsilon C_\omega.$$

Let $I(\theta)$ represent Fisher's information matrix.

[5] χ^2_{k-r} represents chi-square with $k - r$ degrees of freedom.

Under certain regularity conditions on $f(x, \theta)$, we have the following result [40]:

THEOREM 1. *If ω and τ are approximated by two disjoint positively homogeneous sets C_ω and C_τ and the true value of θ is at the origin, then the distribution of -2 log λ_n^* is the same as it would be for the case where $\mathcal{L}(X_i) = N(\theta, I(0)^{-1})$ and ω and τ are replaced by C_ω and C_τ.*

The advantage of this result lies in the fact that the case of normally distributed data is relatively simple to treat.

It is now easy to show that if ω is a smooth r-dimensional surface and τ is the rest of the k dimensional space and $\theta \in \omega$, then

$$\lim \mathcal{L}(-2 \log \lambda_n) = \lim \mathcal{L}(-2 \log \lambda_n^*) = \mathcal{L}(\chi^2_{k-r}).$$

It is also easy to show that if ω is the set on one side of a smooth $(k-1)$-dimensional surface, τ is the rest of k-dimensional space, and θ is on the boundary,

$$\lim_{n\to\infty} \mathcal{L}(-2 \log \lambda_n^*) = \mathcal{L}(u\chi_1^2); \qquad \lim_{n\to\infty} \mathcal{L}(-2 \log \lambda_n) = \mathcal{L}(v\chi_1^2),$$

then where u is independent of χ_1^2 and takes on the values 1 and -1 with probability $\frac{1}{2}$ and $v = \frac{1}{2}(u + 1)$.

In particular, this case applies to testing whether θ lies inside or outside a sphere, and to testing whether θ lies above or below a hyperplane.

In the problem where one is interested in whether θ is in the first quadrant or not, the following is the situation. If θ is on the positive part of either axis,

$$\lim_{n\to\infty} \mathcal{L}(-2 \log \lambda_n^*) = \mathcal{L}(u\chi_1^2).$$

If θ is at the origin, the limiting distribution depends on $I(0)$ and is not difficult to evaluate mumerically.

8. Summarizing remarks. The topic of this paper is so broad and current research in it is so vigorous that it is impossible for me to do more than mention a few of those notions in it that have been of special interest to me. I have tried to give some feeling for those aspects which attract me to the subject and, in so doing, I have neglected a considerable amount of important work done by many people including among others Neyman and Wald.

REFERENCES

[1] W. G. COCHRAN, "The χ^2 distribution for the binomial and Poisson series, with small expectations," *Ann. Eugenics*, Vol. 7 (1936), pp. 207–217.

[2] H. CHERNOFF AND H. RUBIN, "Asymptotic properties of limited-information estimates under generalized conditions," *Studies in Econometric Method*, William C. Hood and L. C. Koopmans, (eds.) Cowles Commission Monograph 14, John Wiley and Sons, New York, 1953, pp. 200–212.

[3] H. B. MANN AND A. WALD, "On stochastic limit and order relationships," *Ann. Math. Stat.*, Vol. 14 (1943), pp. 217–226.

[4] H. RUBIN, "Convergence of probability measures on completely regular spaces," unpublished.

[5] B. V. GNEDENKO AND A. N. KOLMOGOROFF, *Limit distributions for sums of independent random variables*, trans. by K. L. Chung, Addison-Wesley, Cambridge, Mass., 1954, 264 pp.
[6] M. FRECHET, "Les éléments aléatoires de nature quelconque dans un espace distancié," *Ann. Inst. H. Poincaré*, Vol. 10 (1948), pp. 215–230.
[7] H. RUBIN, "Systems of linear stochastic equations," unpublished Ph.D. dissertation, University of Chicago, 1948, 50 pp.
[8] M. D. DONSKER, "Justification and extension of Doob's heuristic approach to the Kolmogorov-Smirnov theorems," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 277–281.
[9] J. L. DOOB, "Heuristic approach to the Kolmogoroff-Smirnov theorems," *Ann. Math. Stat.*, Vol. 20, (1949), pp. 393–403.
[10] H. CRAMÉR, "Sus un nouveau théorème—limite de la théorie des probabilitás," *Actualités Sci. Ind.*, No. 736, Paris (1938).
[11] C. G. ESSEEN, "Fourier analysis of distribution functions," *Acta Math.*, Vol. 77 (1945), pp. 1–125.
[12] W. FELLER, "Generalization of a probability limit theorem of Cramér," *Trans. Amer. Math. Soc.*, Vol. 54 (1943), pp. 361–372.
[13] R. A. FISHER, "On the mathematical foundations of theoretical statistics," *Philos. Trans. Roy. Soc. London*, Series A, Vol. 222 (1922), pp. 390–368.
[14] R. A. FISHER, "Theory of statistical estimation," *Proc. Cambridge Philos. Soc.*, Vol. 22, Part 5 (1925), pp. 700–725.
[15] L. LE CAM, "On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates," *Univ. California Publ. Stat.*, Vol. 1 (1953), pp. 277–330.
[16] A. WALD, *Statistical decision functions*, John Wiley and Sons, New York, 1950, 179 pp.
[17] J. Wolfowitz, "The method of maximum likelihood and the Wald theory of decision functions," *Proc. Roy. Dutch Acad. Sci.*, Vol. 56 (1953), pp. 114–119.
[18] J. VON NEUMANN AND O. MORGENSTERN, *Theory of Games and Economic Behavior*, 2d ed., Princeton University Press, Princeton, N. J,. 1947, 641 pp.
[19] J. L. HODGES AND E. L. LEHMANN, "The Robbins mono-process in the bounded case," to be published in the *Third Berkeley Symposium Proceedings*, University of California Press.
[20] M. FRÉCHET, "Sur l'extension de certains évaluations statistiques au cas de petits échantillons", *Revue de L'Institut International de Statistique*, 11 (1943), pp. 183–205.
[21] G. DARMOIS, "Sur les limites de la dispersion de certains estimations", *Revue de l'Institut International de Statistique*, 13 (1945), pp. 9–15.
[22] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, N. J., 1946, 575 pp.
[23] C. R. RAO, "Information and accuracy obtainable in one estimation of a statistical parameter," *Bull. Calcutta Math. Soc.*, Vol. 37 (1945), pp. 81–91.
[24] L. J. SAVAGE, The Foundations of Statistics, John Wiley and Sons, Inc., New York (1954), 294 pp.
[25] A. BHATTACHARYA, "On some analogues of the amount of information and their uses in statistical estimation," *Sankhyā*, Vol. 8 (1946), pp. 1–14.
[26] A. BHATTACHARYA, "On some analogues of the amount of information and their uses in statistical estimation," *Sankhyā*, Vol. 8 (1947), pp. 201–218.
[27] E. W. BARNAKIN, "Locally best unbiased estimates," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 477–501.
[28] J. WOLFOWITZ, "Efficiency of sequential estimates," *Ann. Math. Stat.*, Vol. 18 (1947), pp. 215–230.
[29] G. R. SETH, "On the variance of estimates," *Ann. Math. Stat.*, Vol. 20 (1949), pp. 1–27.
[30] D. G. CHAPMAN AND H. ROBBINS, "Minimum variance estimation without regularity assumptions," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 581–586.

[31] J. KIEFER, "On minimum variance estimators," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 627–628.
[32] D. A. S. FRASER AND I. GUTTMAN, "Bhattacharya bounds without regularity conditions," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 629–631.
[33] H. CHERNOFF, "Locally optimal designs for estimating parameters," *Ann. Math. Stat.*, Vol. 24 (1953), pp. 586–602.
[34] G. ELFVING, "Optimum allocation in linear regression theory," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 255–262.
[35] H. CHERNOFF, "A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations," *Ann. Math. Stat.*, Vol. 23 (1952), pp. 493–507.
[36] C. STEIN, "Information and comparison of experiments, unpublished.
[37] S. KULLBACK AND R. A. LEIBLER, "On information and sufficiency," *Ann. Math. Stat.*, Vol. 22 (1951), pp. 79–86.
[38] A. WALD, *Sequential Analysis*, John Wiley and Sons, New York, 1947, pp. 60–62.
[39] S. S. WILKS, "The large sample distribution of the likelihood ratio for testing composite hypotheses," *Ann. Math. Stat.*, Vol. 9 (1938), pp. 60–62.
[40] H. CHERNOFF, "On the distribution of the likelihood ratio," *Ann. Math. Stat.*, Vol. 25 (1954), pp. 573–578.