The Annals of Mathematical Statistics
1970, Vol. 41, No. 3, 1034–1058

ON THE INFERENCE AND DECISION MODELS OF STATISTICS[1]

BY COLIN R. BLYTH

University of Illinois

1. Introduction and Summary. The inference and decision models considered here are those described by Neyman in [5] pages 16 and 17. For a random variable X with possible probability distributions indexed by a parameter θ, Neyman distinguishes two inference approaches to the problem of estimating θ: (i) having observed $X = x$, find the most probable values $z(x)$ of θ (This requires a priori probabilities, which often must be chosen rather arbitrarily, and whose existence may be questioned.); and (ii) having observed $X = x$, find the values $z(x)$ that are most reasonable or in which we have the greatest confidence (This requires the rather arbitrary choice of a real-valued function L, with $L(x, \theta)$ measuring the degree of confidence we have in the parameter value θ given that $X = x$ has been observed.). Of these two approaches, (i) is a special case of (ii) since in particular $L(x, \theta)$ can be taken to be the a posteriori probability of θ given $X = x$, for a specified a priori distribution of the parameter.

Neyman remarks that in his opinion "the inferential theory solves no problem" and proceeds to describe a real world situation for which the decision model is a very good one. For a random variable X with possible probability distributions indexed by a parameter θ the decision approach to estimating θ is to associate, with the use of each possible estimator $z(X)$, a random loss $W(z, X, \theta)$ whose possible distributions are indexed by θ; and to determine z so that this loss will, in some specified sense, be as small as possible on the whole over all θ values.

The purpose of the present paper is to examine Neyman's inference model in detail, to describe real situations for which it appears to be a good model, and to compare these with situations for which the decision model is more appropriate. Mathematical models are described, but there is almost no mathematics in the sense of deriving details for the models: concern is mostly with the applied mathematics question of what mathematical model to use for a real situation.

Section 2 is a detailed description of Neyman's general inference model. Here $L(x, \theta)$ is interpreted as a measure of agreement between P_θ probabilities and observed proportions; terms such as "most probable." "most reasonable," "greatest confidence" are avoided as having connotations that are difficult to support. The general inference estimator is just Wolfowitz's minimum distance estimator [12]. Wolfowitz evaluates such procedures purely from the decision viewpoint.

Section 3 is a detailed description of the general decision model. This is given in a form closely paralleling the description of Section 2, in order that the two models can be compared easily.

Received September 20, 1966; revised September 6, 1968.

[1] Work supported by National Science Foundation Grant U.S. N.S.F. GP 3814.

In Section 4, the inference and decision models are compared from several viewpoints. The inference problem seeks an estimator z such that the $P_{z(x)}$ probabilities will be (for all x) close to the proportions observed in x; the decision problem of estimating θ seeks an estimator z such that $z(X)$ will be (for all θ) close to θ. The decision problem requires the idea of distance or error, measured by loss, in the parameter space; this idea is completely or partially absent in the inference model, where θ merely indexes the possible probability models. It is this presence or absence of a loss function that distinguishes between the decision and inference models as defined here: in the decision problem the use of an estimator z results in definite losses and we are to determine a z for which they are small; in the inference model the idea of definite losses does not appear. The decision model is a much more specific model for a much more specific real problem. In both problems we want to choose a probability model for the real situation: in the decision problem we know what the model is to be used for; in the inference problem we do not.

It can be reasonably argued, following Neyman [5] page 17, that the inference model is sometimes used in the mistaken belief that making $P_{z(x)}$ close to the observed proportions will make $P_{z(X)}$ close to P_θ, and that if so it should be replaced by the decision model with loss $W[P_{z(X)}, P_\theta]$. In the inference problem as described here, we do not have this definite aim of making $P_{z(X)}$ close to P_θ: we are uncertain as to whether we want $P_{z(X)}$ close to P_θ or want $z(X)$ close to θ or want $[z(X)]^2$ close to θ^2 or want something else; vaguely, we have all decision-type aims, but we have no *definite* one.

Both the inference and decision models require the making of somewhat analogous and rather arbitrary choices at two levels. (Here, and throughout this paper "arbitrary" is used in its primary dictionary meaning of "depending on will or discretion; discretionary; can be freely chosen," with none of the secondary dictionary meanings of "unreasoned, despotic." Possible synonyms such as "subjective," "individualistic," "personalistic" are avoided because of their technical meanings.) Neyman's view that the user should be fully entitled to any choices he cares to make and that no attempt should be made to impose particular choices on all, is followed here.

In Section 5 the most commonly used inference methods (Likelihood, Least Squares and Moments, Chi-square, Kolmogorov–Smirnov) are examined as special cases of the general inference method.

Randomization. The language used throughout will be that of non-randomized procedures, but is to be understood as including randomization. Actually, there is no such thing as a randomized procedure: randomization has to be based (in a non-random way) on the outcome of an additional random experiment; and this randomization device must be described explicitly, for an exact account of randomization. Therefore, all we need to do is take our X as including all available randomization devices. For example, instead of taking X to be Normal $(\mu, 1)$ and considering randomized procedures, we can take $X = Y, Z$ where Y and Z are independent, Y is Normal $(\mu, 1)$, Z is Rectangular $(0, 1)$, and consider non-randomized procedures. Moreover it is necessary to do some such thing because

the first formulation fails to specify what randomization devices are available to us.

The randomized procedures based on Y are commonly taken to consist of all the non-randomized procedures based on Y, Z where Z is Rectangular (0, 1) and independent of Y. It would be no more general to use a real, vector or sequence-valued randomization device with arbitrarily specified conditional distribution given $X = x$, because it is easy to construct a function t such that $t(Z)$ duplicates such a device. This can be done in two steps: first, write Z in decimal form $Z = \cdot U_1 U_2 U_3 \cdots$ and make the usual one-to-one mapping of the unit interval on to sequences of numbers from the unit interval

$$Z_1 = \cdot U_1 U_3 U_6 \cdots$$
$$Z_2 = \cdot U_2 U_5 \cdots$$
$$Z_3 = \cdot U_4 \cdots$$
$$\cdots$$

Here $Z_1, Z_2, \cdots$ is a function of Z, and the Z_i's are independent, each Rectangular (0, 1); second, for the ith component of $t(Z)$ take $F^*(Z_i)$ where F^* is essentially the inverse of the required cumulative probability function for the ith component of $t(Z)$, given the earlier components and given $Y = y$.

The objection is sometimes made that the use of such an extraneous Z ought to be ruled out as intuitively unreasonable. This objection can be overcome, in many problems, by noticing that no such Z is needed. The argument used in the following example shows that Z is unnecessary if there is a sufficient statistic $S = s(Y)$ such that the conditional distribution of Y given $S = s_0$ has no points of positive probability.

EXAMPLE. Let $Y_1, \cdots, Y_n, Z$ be independent, where each Y_i is Normal $(\mu, 1)$ and Z is Rectangular (0, 1), write Φ for the Normal (0, 1) cumulative probability function, and consider the following classes of possible distributions of non-randomized procedures:

A_1: All those for procedures based on $Y_1, \cdots, Y_n$

A_2: All those for procedures based on $Y_1, \cdots, Y_n, Z$

A_3: All those for procedures based on $\bar{Y}, Z$

A_4: All those for procedures based on $\bar{Y}, \Phi(2^{-\frac{1}{2}}(Y_1 - Y_2))$.

Then $A_1 \subset A_2$ is obvious; $A_2 \subset A_3$ by the usual sufficiency argument, [4] page 18; $A_3 = A_4$ because the same joint distribution is involved in both; and $A_4 \subset A_1$ is obvious. Therefore $A_1 = A_2 = A_3 = A_4$ and so any one of them can be used for the randomized procedures based on $Y_1, \cdots, Y_n$. Ordinarily we would work with A_3 as easiest; if anyone objects to the Z we could change to the equivalent A_4 or A_1.

2. General statistical inference problem. Given independent random variables $X_1, \cdots, X_n$ each with the same probability distribution which is known to belong to

a given class P_θ, $\theta \in \Omega$ of probability measures on $\mathcal{X}_1$, and having observed $X_1, \cdots, X_n = x_1, \cdots, x_n$ (abbreviated to $X = x$), the general statistical inference problem is to choose a subset $z(x)$ of Ω giving those probability models from the available class that are in best agreement with the empirical probability measure P_x. This empirical probability measure P_x is the discrete probability measure that places probability m/n on each point of $\mathcal{X}_1$ on which m of the values $x_1, \cdots, x_n$ fall.

The reason for using the notation $X_1, \cdots, X_n$ independent and equi-distributed, when the same notation with $n = 1$ is equally general (we can take X to be a random variable with possible distributions the product of the P_θ measures, on $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_1$ and on which we have a single observation x), is that for $n = 1$ the empirical probability measure P_x degenerates to placing probability 1 on the single point x, and discussion of agreement between P_θ and P_x is more natural for non-degenerate P_x. In addition, the fact that $P_x \to P_\theta$ as $n \to \infty$ (in the sense of the Borel–Cantelli Lemma) may be of interest.

Restrictions on z. The function z or the random variable $z(X)$ it generates is called an estimator of θ. In any particular problem, attention is restricted to a very small class of functions z. This class depends on the sort of inference wanted and may be further restricted by intuitive ideas of reasonableness. In point estimation of θ, for example, z is restricted to functions whose values are single points of Ω, and may be further restricted to functions having some intuitively reasonable property such as equivariance under a group of transformations on $\mathcal{X}$ and an induced group of transformations on Ω. In point estimation of $g(\theta)$, z is restricted to functions whose possible values are sets of the form $\{\text{All } \theta \text{ such that } g(\theta) \text{ has the value } g_1\}$. In confidence set estimation z is restricted to functions such that, for all $\theta \in \Omega$, $P_\theta\{\theta \in z(X)\}$ is at least as great as a prescribed constant; and z may be further restricted, for example, when θ is real valued, to functions whose possible values are intervals, or intervals of the form $(-\infty, a)$. In hypothesis testing z is restricted to functions having only two possible values: a prescribed subset H of Ω, and the complement $H' = \Omega - H$ of that subset; and there is the further restriction $P_\theta\{z(X) = H'\} \leqq \alpha$ whenever $\theta \in H$.

The restrictions on z can involve only such structure on Ω as is relevant, i.e. which corresponds to structure in the distributions and in the real situations for which these are the possible models. The more such structure there is, the more restrictions can be considered. There may be no relevant structure at all, the elements of Ω being merely arbitrary labels for the possible distributions, and distance or closeness in Ω implying no such thing for the corresponding distributions or real situations; here restrictions on z must be invariant under all one-to-one mappings of Ω onto itself; equivariance is an example of such a restriction. Closeness in Ω may be relevant; invariance under continuous mappings is then enough: the requirement that $z(x)$ be a connected set is an example. There may be a relevant ordering in Ω; invariance under monotone mappings is then enough: the requirement, for real-valued θ, that $z(x)$ be an interval of the form $(-\infty, a)$ is an example. In inference, there is indefinite structure on Ω such as closeness or order, rather than specific structure such as distance. A condition such as unbiasedness $Eg[z(X)] \equiv g(\theta)$,

which requires relevance of the linearity in the space of $g(\theta)$, is not really part of the inference approach.

In point estimation, the sense of best agreement between P_θ and P_x can be made specific by specifying $d(P_\theta, P_x)$, where d is a nonnegative-valued function, as measuring the discrepancy between P_θ and P_x. Best agreement between P_θ and P_x is then taken to mean that $d(P_\theta, P_x)$ is as small as possible: if $d(P_\theta, P_x)$ has, for every x, a minimum over $\theta \in \Omega$, then $z(x)$ is taken to be any θ achieving this minimum, and this determines a unique best inference estimator z (or a class of estimators, all equivalently best inference).

If restrictions on z rule out the above estimator (as may happen in point estimation and does happen in other problems) the sense of best agreement is not made specific by specifying a discrepancy measure d; because when $X = x$ is observed the best available value for $z(x)$ is now the set $\{\theta: d(P_\theta, P_x) \leqq c(x)\}$ where $c(x)$, subject to the restrictions on z, is taken to be as small as possible as a whole over all x (in a sense that must still be specified). More generally, if the values of $z(x)$ are restricted to a class $\mathscr{A}$ of subsets of Ω, then $d^*(A, P_x) = \inf_{\theta \in A} d(P_\theta, P_x)$ is taken as a measure of discrepancy between the set of probability measures P_θ, $\theta \in A$ and the empirical probability measure P_x. When $X = x$ is observed the most reasonable value for $z(x)$ is now any $A \in \mathscr{A}$ such that $d^*(A, P_x) = c(x)$ is as small as possible as a whole over all x (in some specified sense), subject to any additional restrictions on z.

In the preceding discussion we could equivalently have specified $a(P_\theta, P_x)$, where a is a nonnegative-valued function, as measuring the agreement between P_θ and P_x, and interchanged inf and sup throughout. We can always change from an agreement measure a to a discrepancy measure d which is a strictly decreasing function of a, so there is no need for a separate general discussion using agreement measures. In discussing particular inference procedures we will use either an agreement or a discrepancy measure, whichever is usual for the procedure under discussion.

One convenient way of constructing a discrepancy measure d is as follows. Having observed $X = x$, choose a class $\mathscr{T}$, depending on x, of real-valued functions t on $\mathscr{X}_1$, and take for d some measure of discrepancy between the expectations of $t(X_1)$ under P_θ and under P_x over the class $\mathscr{T}$, i.e. between $E_\theta\, t(X_1)$ and $\sum_{i=1}^n t(x_i)/n$ over the class $\mathscr{T}$. For a set $S \subset \mathscr{X}_1$, asking for agreement between $P_\theta(S)$ and the proportion of x_i's in S is to take for t the characteristic function of the set S; such functions are often used. This method of constructing a discrepancy measure is fully general, since $\mathscr{T}$ can be constructed so that $E_\theta\, t(X_1)$, $\theta \in \Omega$ duplicates P_θ, $\theta \in \Omega$; for example, when X_1 is real valued, take for $\mathscr{T}$ the indicator functions of the sets $(-\infty, t]$, for all t, $-\infty < t < \infty$.

The general inference method described above is due to Wolfowitz—his minimum distance method [12]. It is the general inference method in that it includes as special cases all the usual inference methods such as estimation by the methods of maximum likelihood, minimum chi-square, moments, least squares; fiducial confidence intervals; and likelihood ratio, chi-square, and Kolmogorov–Smirnov tests. For

point estimation of θ with no additional restrictions Wolfowitz shows, subject to regularity conditions, that an inference method applied successively for $n = 1, 2, \cdots$ gives a consistent sequence of estimators. His proof begins with the fact that as $n \to \infty$, P_X gets close to P_θ, where θ is the true parameter value, in the sense of the Borel–Cantelli lemma. The first regularity condition is that this should imply $d(P_\theta, P_X) \to 0$ in probability. This in turn implies $d(P_{z(X)}, P_X) \to 0$ in probability, because $z(x)$ is the θ' value minimizing $d(P_{\theta'}, P_x)$ over $\theta' \in \Omega$. A second regularity condition is that $P_{z(x)}$ and P_θ being both close to P_x should imply that they are close to each other—Wolfowitz ensures this by having d satisfy the distance axioms. Completing the proof is the identification and continuity regularity condition that $P_{z(x)}$ close to P_θ should imply $z(x)$ close to θ. The first regularity condition is a critical one: for some inference methods (e.g. Kolmogorov–Smirnov) it is satisfied, but for most inference methods (e.g. maximum likelihood) it is not satisfied, and consistency for those methods does not follow from the Wolfowitz theorem.

In problems where, for every $x \in \mathcal{X}$, there is a θ giving $P_\theta \equiv P_x$, this gives the inference point estimator (no additional restrictions on z) of θ, which is the same for every d because any reasonable d must be minimized by $P_\theta \equiv P_x$. An example is the binomial problem $X_1, \cdots, X_n$ independent, $P(X_i = 1) = \theta$, $P(X_i = 0) = 1 - \theta$, where $\theta = \sum x_i$ gives the above identity, and where every inference point estimator is $\sum X_i/n$.

In the inference problem an uncertain inference or guess or estimate is to be made as to the value of θ. This inference has to do with agreement between θ-probabilities and the observed proportions; it has nothing to do with any possible losses resulting from poor guesses. The ideas of specific decisions to be made, and definite losses resulting from poor decisions, are not a part of the inference model.

3. General statistical decision problem. The probability model is exactly the same as in the inference problem: we are given independent random variables $X_1, \cdots, X_n$ each with the same probability distribution which is known to belong to a given class P_θ, $\theta \in \Omega$ of probability measures on $\mathcal{X}_1$. Having observed $X_1, \cdots, X_n = x_1, \cdots, x_n$ the general statistical decision problem is to choose a decision $\delta(x_1, \cdots, x_n)$ from a given class D of possible decisions. In many decision problems each element of D can be thought of as "acting as though we believe that θ lies in some particular subset of Ω." In the following discussion $\delta(x_1, \cdots, x_n)$ will be identified with this corresponding subset $z(x_1, \cdots, x_n)$, and the language of estimation rather than decision-making used, in order to make easier comparisons between the inference and decision problems.

Associated with a parameter value θ and the use of an estimator z is a family L_i, $i \in I$ of random losses, $L_i = W_i(\theta, z, X_1, \cdots, X_n)$ being a nonnegative-valued random variable, with $W_i(\theta, z, x_1, \cdots, x_n)$ being the type i loss incurred when θ is the parameter value, the estimator z is used, and $X_1, \cdots, X_n = x_1, \cdots, x_n$ is observed. The general statistical decision problem is to find an estimator z for which the random losses L_i, $i \in I$ are as small as possible (in some specified sense).

The following are some commonly used systems L_i, $i \in I$ of loss functions. In

point estimation I contains a single element, with L_1 depending on θ and $z(X_1, \cdots, X_n)$ only. In hypothesis testing (Neyman–Pearson model) I contains two elements, with L_i having the value 1 or 0 according as a type i error is or is not committed. In confidence set estimation (Neyman model) I is a duplicate of Ω and L_i has the value 0 or 1 according as $i \in z(X_1, \cdots, X_n)$ or not when $i = \theta$; and L_i has the value 1 or 0 according as $i \in z(X_1, \cdots, X_n)$ or not when $i \neq \theta$; that is, for each possible value i of the parameter we consider a loss L_i which is 1 if either i is the true parameter value and is not covered or is not the true parameter value and is covered, and is otherwise 0. In sequential estimation, $n = 1$, X_1 is sequence-valued, and I contains two elements with L_1 depending on θ and $z(X_1)$ only, and L_2 depending on the number of elements of X_1 needed to determine $z(X_1)$. In design of experiments $n = 1$, $X_1 = Z_\alpha$, $\alpha \in A$ consists of all possible experiments under consideration, and I contains two elements with L_1 depending on θ and $z(X_1)$, and L_2 depending on the set of α's needed to determine $z(X_1)$ [this is an experiment being designed for point estimation].

The meaning of smallness of a nonnegative-valued random loss (in the above notation, when I consists of a single element we are concerned with one such loss only) is specified in the following way. *First*, for a random loss L with known distribution we specify r_L, called the risk of L, as measuring how bad we consider this loss to be, where r is a nonnegative-valued function of the distribution of L; i.e. of two such losses we consider the one with smaller r_L to be the smaller. This associates with L a risk function r_L, where $r_L(\theta)$ is the risk specified for the distribution that L has when θ is the parameter value. Should there be an estimator z which minimizes $r_L(\theta)$ for all θ, the meaning of smallness is now completely specified and such a uniformly minimum risk estimator is used as best. Uniformly minimum risk estimators often exist when the class of available z's is a fairly restrictive one: for example the class of unbiased point estimators of θ often contains a uniformly minimum risk member. More usually, different estimators z minimize $r_L(\theta)$ for different θ values, so that the meaning of smallness is not yet specified, and a second specification is needed: *Second*, we specify $q(r_L)$ as measuring how large we consider $r_L(\theta)$ to be as a whole over all $\theta \in \Omega$, where q is a nonnegative-valued function of the function r_L. Then a best estimator z is taken to be one for which the corresponding $q(r_L)$ is as small as possible.

The specifications of r and q usually worked with are those used by Wald [9], [11], namely $r_L = EL$ for risk and $q(r_L) = \sup_{\theta \in \Omega} r_L(\theta)$ for overall size—such a best estimator is called minimax. Other specifications of risk could be used, such as $r_L =$ median of L, or $r_L = 95$ percentile of L, or $r_L = P(L \geqq c)$ where c is a disaster level. Other specifications of overall size are used, such as $q(r_L) = \int_\Omega r_L(\theta)\, d\lambda(\theta)$ [such a best estimator is called Bayes (λ)], or $q(r_L) = \sup_{\theta \in \Omega} \{r_L(\theta) - \inf_z r_L(\theta)\}$ [such a best estimator is said to minimize the maximum regret].

When there is more than one nonnegative-valued random loss, i.e. when I contains more than one element, the losses L_i, L_j for $i \neq j$ are understood not to be fully comparable or interchangeable. That is, while $c_1 L_1 + c_2 L_2$ can be worked with formally, it is not possible to replace the L_1, L_2 losses by the single loss $c_1 L_1 + c_2 L_2$.

Sometimes such a replacement is possible in principle, but we do not want to commit ourselves on c_1, c_2 values. This idea of several kinds of losses that are not fully comparable or interchangeable was introduced by Neyman and Pearson [7] who considered the hypothesis testing problem of minimizing the probability of a type II error subject to a bound on the probability of a type I error, with the relative seriousness of the two kinds of error difficult to specify. It is used by Wald [10] in the sequential test problem of minimizing expected sample size subject to bounds on the probabilities of type I and II errors, where increases in these probabilities cannot be traded for smaller sample size; and in sequential estimation subject to a budget on expected sampling cost, where we are not free to exchange greater sampling cost for greater accuracy.

When there is more than one L_i, a risk and overall size of risk are specified separately for each L_i, $i \in I$ and the problem considered is that of finding a z for which some named L_i is best, subject to given bounds on the risks of the others. This includes the problem of finding a z for which several L_i's are best subject to bounds on the risks of the others, should there be a z that does this uniformly in the named i's; failing this, a further specification of smallness as a whole over the named i's would be needed.

In the decision problem, the choice of an estimator z results in definite losses, and the aim is to determine z so that these losses will be small; in the inference problem, no definite losses are involved, and the aim is to determine a z giving probability models from the given class that are in best agreement with observed proportions.

4. Comparison of the inference and decision problems. In this section the statistical inference and decision problems are compared from several viewpoints.

(1) *The probability models are the same in both problems.*

(2) *More restrictions on the estimator are possible in the decision problem.* The primary restrictions (e.g. to point-valued z's, to confidence sets, to z's with only two possible values) are the same in both problems, as are some additional restrictions such as equivariance. But there is usually more relevant structure on Ω in the decision problem, and in addition there may be restrictions involving the losses, which are not available in the inference problem.

(3) *The decision problem is a more special mathematical model for a more specific real problem*; the inference problem is a more general mathematical model for a less specific real problem. To one inference problem correspond many decision problems, one for each choice of loss. For example, estimation of θ and estimation of $g(\theta)$ are the same problem when g has an inverse, provided any restrictions on z are invariant under the g mapping: if $z(x)$ is the best set of values for θ, then automatically $g[z(x)]$ is the best set of values for $g(\theta)$. But estimation of θ and estimation of $g(\theta)$ are different decision problems even when g has an inverse: stating both as problems of estimating θ, it would be understood that we want $z(X)$ close to θ in one case, $g[z(X)]$ close to $g(\theta)$ in the other; since these are not equivalent, different losses would ordinarily be involved.

There is very little competition between the inference and decision problems as

to which should be used as a model for a particular real problem. If we are estimating θ for a specific purpose and know at least approximately what losses result from errors, there seems to be fairly general agreement that the decision problem is the more suitable model. If on the other hand our purpose in estimating θ is merely to choose a probability model for the real situation in an attempt to systematize our examination of it, and with no idea as to what specific questions may be interesting [for example, with no idea as to whether it is important to have $z(X)$ close to θ, or to have $[z(X)]^2$ close to θ^2], and with no idea as to the existence of losses let alone their amounts, it would seem that we will have to make do with the inference problem as a model. One possibility is to use the inference model in the early stages of an investigation, for the purpose of deciding on what specific decision questions ought to be considered.

The necessity of actually specifying losses presents a decided practical difficulty in using the decision model: even though the existence of losses often seems self-evident it is usually very difficult to say exactly what these losses are, and they have to be chosen rather arbitrarily. Thus in point estimation squared error loss is almost always used, not because it is thought to be a very good approximation to actual losses (which are always bounded by the total fortune of the decision-maker), but because it results in easy mathematical problems.

Experimenters show decided preference for the inference model, because they have not usually decided in advance on some definite question to ask about the real situation, and even if they have, the inference choice relieves them of the difficult task of specifying what losses are involved.

(4) *In both problems, choices of a rather arbitrary sort must be made by the experimenter.* In the inference problem he must decide on a meaning for (i) best agreement between P_θ and P_x for given x and (ii) best agreement as a whole over all x. In the decision problem he must decide on a meaning for (i) smallness of a random loss L_i for given θ and (ii) smallness as a whole over all θ. In both cases, it would seem that the experimenter is fully entitled to whatever choices he cares to make. Choices for the inference problem will be discussed in Section 5.

The experimenter finds this situation unpleasant because he has the very difficult practical task of making choices suitable for his real problem. On the other hand, the statistician finds this situation pleasant because of the great freedom of choice—he can consider a wide variety of problems, and can make choices leading to easy and elegant mathematical problems.

(5) *The inference problem is usually much easier to solve.* (Neyman in [5] refers to it as the "easy way out".) The inference problem, once the arbitrary choices have been made, and provided restrictions on z do not cause complications, is solved by finding the θ that minimizes a known function of θ. In parametric problems, where the set Ω of possible θ's is a subset of a Euclidean space and continuous functions are involved, this problem is easily solved (at least in principle) using elementary calculus methods.

The decision problem, once the arbitrary choices have been made, is solved by finding the z that minimizes a known function of z. Here the set of possible z's

is a class of functions on the space of X to a family of subsets of Ω. Special methods can be used to solve this minimization problem for special kinds of problems, especially when there are strong restrictions on z. But no very general methods are available, and we often find the problem too hard to solve.

(6) *Inference methods from the decision viewpoint.* Because of ease of solution, inference methods are of great practical interest even in problems for which the decision model seems more suitable. Commonly, the decision problem cannot be solved, while one or more inference estimators can be written down rather easily, their performances from the decision viewpoint easily computed, and the best one of them used in the hope that it may be fairly good (which may or may not be true). In point estimation with squared error loss and expected loss as risk, a Schwarz inequality bound on risk can sometimes be used to show that some particular inference estimator either is optimum or cannot be much improved upon.

In some classes of problems, estimators obtained by particular inference methods can be shown to have optimum properties from particular decision viewpoints: the Gauss–Markov theorem for least squares estimation is an example. Wolfowitz [12] raises the general question of trying to determine what inference choices will give estimators that are good from some particular decision viewpoint.

From the decision point of view most inference methods are asymptotically good as $n \to \infty$, roughly because as $n \to \infty$ (identifiability and continuity assumed), θ becomes known and there is no longer any uncertainty. Wolfowitz's consistency proof [12] is the most general result of this sort. Similar, and stronger, results are known for particular methods; for example, the consistency and asymptotic efficiency results for maximum likelihood estimation.

5. Survey of inference methods. In this section some commonly used inference methods are considered as examples of the general inference method. For the hypothesis testing examples that are included, we need to consider both the Karl Pearson and Neyman–Pearson models of testing. In both, the values of $z(x)$ are restricted to a named subset H of Ω, and its complement H'; and there is the size restriction $P_\theta\{z(X) \neq H\} \leqq \alpha$ for all $\theta \in H$.

Testing Hypotheses I (Karl Pearson model). The question asked is: does H provide a reasonable model or not? The answer given is

Reject H if $d^*(H, P_x) > c$
Accept H if $d^*(H, P_x) \leqq c$

where the size condition determines c.

Testing Hypotheses II (Neyman–Pearson model). The question asked is: which of H, H' provides the more reasonable model? The answer given is

Reject H if $d^*(H, P_x)$ is large compared to $d^*(H', P_x)$.
Accept H if $d^*(H, P_x)$ is small compared to $d^*(H', P_x)$

where the size condition determines where the line is drawn. Here the meaning of

"large compared to" has to be specified. The choice made by Neyman and Pearson [6], for a particular d, is

Reject H if $d^*(H, P_x) > c \cdot d^*(H', P_x)$
Accept H if $d^*(H, P_x) \leqq c \cdot d^*(H', P_x)$.

In model I we do not care how good a model H' may provide, because H' is only a formal alternative: $z(x) = H'$ means not that we are going to use H' as model, but only that we are not going to use H. Indeed, model I is most easily put in general inference form by using H and the null set for the possible values of $z(x)$; alternatively, we can use H and H', and artificially modify a natural d to be larger for $\theta \in H'$. In model II, on the contrary, H' is a genuine alternative, on the same footing as H except for the non-symmetry introduced by the size condition. In problems where H indexes some relatively small class of distributions and H' indexes everything else, it often happens that for every $x \in \mathcal{X}$, $\theta \in H$ there exists $\theta' \in H'$ with $|d(P_\theta, P_x) - d(P_{\theta'}, Px)|$ arbitrarily small. Then $d^*(H, P_x) = d^*(H', P_x)$ for all x, and model II cannot be used without modifying d, or modifying H, H' to provide an indifference region.

(1) *Likelihood methods.* These are characterized by the choice of $p_\theta(x_1) \cdots p_\theta(x_n)$, where p_θ is a probability density of X_1 relative to some fixed measure μ, as a measure of agreement between P_θ and P_x. The intuitive reasonableness of this measure can be examined in two ways. One way is to think of the sample size as 1, with $X_1 \cdots, X_n$ taking the place of X_1 of the general discussion. Then P_x is a measure putting probability 1 on the single point x. Having P_θ maximize the probability (density if μ is not discrete) on x is one obvious measure of best agreement; another possibility would be to have P_θ centered at x in some sense. A second way is to think of the sample size as n, so that P_x places probability n_j/n on each point of $\mathcal{X}_1$ on which n_j of the x_i's fall. Since $\prod_j p_j^{n_j}$ is maximized, subject to $\sum_j p_j \leqq 1$, by the choice $p_j = n_j/n$, one measure of best agreement between P_θ and P_x is to make $\prod p_j^{n_j} = p_\theta(x_1) \cdots p_\theta(x_n)$ [in the discrete case] as large as possible. From this point of view it is even clearer that, while likelihood is a very reasonable measure of agreement, there is nothing necessary or unique about this choice. In fact, other measures are in common use. The least squares discrepancy measure $\sum(p_j - n_j/n)^2$ and the chi-square discrepancy measure $\sum(p_j - n_j/n)^2/p_j$ are examples, μ discrete.

When μ is not discrete and the function given by $p_\theta(x)$ is not continuous in both θ and x, likelihood is unreasonable and must be replaced by some such measure as $\limsup P_\theta(X_1 \in S)/\mu(S)$, the limit being over open sets containing x with $\mu(S) \to 0$. It is difficult to cover desired cases without covering undesired cases as well, and no reasonably general definition is in common use.

The use of likelihood as an agreement measure can be applied to any kind of inference problem—point estimation, testing hypotheses I and II, confidence set estimation:

(1a) Point estimation (Maximum Likelihood Estimation). The choice of likelihood $\prod p_\theta(x_i)$ as agreement measure specifies a point estimator z given by taking for $z(x)$ any value of θ maximizing $\prod p_\theta(x_i)$. This maximum likelihood estimator is

often easy to calculate, and is very widely used. From the inference point of view it is a best estimator in its own right and questions about losses are not relevant. From the decision point of view it is usually fairly good but can be very bad. Asymptotically as $n \to \infty$ and subject to certain regularity conditions (those of Wolfowitz [12] are not satisfied) it has some optimum properties. Lehmann [3] gives a list of references.

(1b) Testing Hypotheses I. For this problem the choice of likelihood as an agreement measure specifies the test

$$\text{Reject} \quad H \quad \text{when} \quad \sup_{\theta \in H} \prod p_\theta(x_i) \leqq c$$

where c is determined by the size condition. This test is sometimes used in practice, especially for a simple hypothesis H. No particular name seems to be attached to it.

(1c) Testing Hypotheses II (Likelihood Ratio Test). One way of making a specific choice of test based on likelihood as an agreement measure is the likelihood ratio test of Neyman and Pearson [6]:

$$\text{Reject} \quad H \quad \text{when} \quad \sup_{\theta \in H} \prod p_\theta(x_i) \leqq c \cdot \sup_{\theta \in H'} \prod p_\theta(x_i)$$

where c is determined by the size condition. From the inference point of view, the likelihood ratio test is a best test in its own right. From the decision point of view it is usually a good test (Neyman–Pearson Lemma shows it optimum for simple H and H'); it can be very bad. Asymptotically as $n \to \infty$ it has some optimum properties: Lehmann [4] page 16 gives references.

(1d) Confidence Set Estimation (Fiducial Limits). One way of making a specific choice, based on likelihood as an agreement measure, is

$$z(x) = \{\theta : \prod p_\theta(x_i) \geqq c\}$$

where the confidence level condition determines c. Kendall and Stuart, [2] vol. 2, page 136, call this the fiducial interval for θ (in cases where this set is an interval). This is the confidence set obtained by inverting the tests 1b of single point hypotheses. From the inference viewpoint it is a most reasonable set in its own right; from the decision viewpoint it may be good or bad.

Fiducial confidence sets do not appear to be much used in practice: they are often difficult to compute and have no decision-optimum properties that would justify heavy computing effort. For example, when X has binomial (n, p) distribution the maximum likelihood estimator is computed by maximizing $\binom{n}{x} p^x (1-p)^{n-x}$ which is very easy to do; while to compute the fiducial interval one must find the two solutions of $\binom{n}{x} p^x (1-p)^{n-x} = c$ for arbitrary c and determine c so that the confidence level condition is satisfied. This is too laborious—almost everybody uses the confidence set, tabled by Clopper and Pearson, that can be obtained with no work at all by inverting the equal probability tail tests.

Kendall and Stuart caution that some authors use the term fiducial for confidence sets other than the above one. Conceivably the term might be used for any confidence

set whatever that is based on likelihood as an agreement measure. Another example of such a set is that given by

$$z(x) = \{\theta : \prod p_\theta(x_i) \geqq c \sup_\theta \prod p_\theta(x_i)\},$$

obtained by inverting the likelihood ratio tests (1c).

(2) *Least squares methods.* In its most general form the method of least squares is just the general inference method as described in terms of the expectations, under P_θ and P_x, of a class $\mathscr{T}$ of real-valued functions on $\mathscr{X}_1$, with the important restriction that the class $\mathscr{T}$ should not depend on x. That is, having observed $X = x$, choose a class $\mathscr{T}$, *not* depending on x, of real-valued functions t on $\mathscr{X}_1$, and take for d some measure of discrepancy between $E_\theta t(X_1)$ and $\sum_{i=1}^n t(x_i)/n$ over the class $\mathscr{T}$.

The discrepancy measure always used is given by

$$d(P_\theta, P_x) = \sum_{t \in \mathscr{T}} \{E_\theta t(X_1) - \sum_{i=1}^n t(x_i)/n\}^2,$$

from which comes the name "least squares," our aim being to make d as small as possible by choice of θ. This includes weighted least squares—the weights can be included in the t's. However, if it seems more reasonable or if it leads to an easier minimization problem, one might want to use some other discrepancy measure such as

$$\sum_{t \in \mathscr{T}} \left|E_\theta t(X_1) - \sum_{i=1}^n t(x_i)/n\right| \quad \text{or}$$

$$\sum_{t \in \mathscr{T}} \left\{\left|E_\theta t(X_1)\right|^{\frac{1}{2}} - \left|\sum_{i=1}^n t(x_i)/n\right|^{\frac{1}{2}}\right\}^2$$

and still, somewhat improperly, use the name least squares.

Usually $\mathscr{T}$ is taken to be a finite set, commonly the set given by $X_1, X_1^2, \cdots, X_1^k$ for real X_1, and by $U_1, V_1, U_1^2, U_1V_1, V_1^2, \cdots, V_1^k$ for bivariate $X_1 = U_1, V_1$. However, any functions at all can be used, provided only that their expectations exist. There are some obvious guides on the choice of t's: (i) the expectations should be simple functions of θ making it easy to determine what θ values minimize d; (ii) the expectations should depend strongly on θ, preferably in a monotone manner —thus a function whose expectation does not depend on θ is no use at all, and the function given by $t(x) = 0$ for x negative, 1 for x nonnegative where X_1 is real valued, may be a good choice for a translation parameter θ but a bad choice for a scale parameter θ; (iii) because of the discreteness of P_x, continuous functions t seem preferable—small changes in the x_i's then will make only a small change in the value of the estimator; (iv) functions t that are bounded or at least do not grow too fast as $x \to \infty$ are preferable because any realization of P_x is a bounded approximation to the true P_θ and necessarily ignores the "tails" of P_θ.

The method of moments is a special case of the method of least squares, in which the class $\mathscr{T}$ (usually consisting of the powers and cross products mentioned above) is so chosen that a unique value of θ makes $d = 0$, i.e. makes $E_\theta t(X_1) \equiv \sum_{i=1}^n t(x_i)/n$.

The method of least squares could be used for any type of inference problem (point estimation, testing hypotheses, confidence set estimation) but is in common

use only for point estimation. From the inference viewpoint it is its own justification; from the decision viewpoint it has the optimum property of giving minimum variance unbiased linear estimators (the Gauss–Markov theorem, see Plackett [8]). It is a very useful method because one can almost always choose $\mathscr{T}$ and d so as to easily obtain an estimator even in problems where the usual methods lead to minimization problems too difficult to carry out. From the decision viewpoint this estimator may be good or bad, but at least it is *an* estimator, and resulting losses can be computed. Here are two very simple examples illustrating expediency as a criterion for choosing $\mathscr{T}$ and d.

EXAMPLE. Let X_1 have a Poisson (λ) distribution and let $n = 1$. Here it seems foolish to use the functions given by $X, X^2, X^3, \cdots$ when the functions given by $X, X(X-1), X(X-1)(X-2), \cdots$ have the very much simpler expectations $\lambda, \lambda^2, \lambda^3, \cdots$. Further, it seems foolish to use the d given by $\{x-\lambda\}^2+\{x(x-1)-\lambda^2\}^2+\{x(x-1)(x-2)-\lambda^3\}^2+\cdots$ when the d given by $\{x-\lambda\}^2+\{[x(x-1)]^{\frac{1}{2}}-\lambda\}^2+\{[x(x-1)(x-2)]^{\frac{1}{3}}-\lambda\}^2+\cdots$ is so much easier to minimize. In both cases, there is no obvious reason for thinking one choice more reasonable than the other.

EXAMPLE. Let X_1 have Cauchy density given by $(1/\pi\theta)/\{1+(x/\theta)^2\}$, with θ an unknown scale parameter. Here the usual methods (e.g. maximum likelihood) are hard to carry out. But it is easy to estimate θ using the method of moments—just use any convenient function t. One such function, given by $|x|^{\frac{1}{2}}$, has expectation $2^{\frac{1}{2}}\theta^{\frac{1}{2}}$; equating this to the P_x expectation $\sum|x_i|^{\frac{1}{2}}/n$ of the same function gives us the least squares estimator $\{\sum|X_i|^{\frac{1}{2}}/n\}^2/2$ for θ.

(3) *Chi-square methods.* These are characterized by the following choice of discrepancy measure, due to Karl Pearson. We partition the space of X_1 into k regions $S_1, \cdots, S_k$ (this partitioning does not depend on x) and ask for best agreement between probabilities and observed proportions for these regions in the sense of the discrepancy measure given by $d(P_\theta, P_x) = n.\sum_{i=1}^{k}\{P_\theta(S_i)-P_x(S_i)\}^2/P_\theta(S_i)$. This is just the method of least squares with $\mathscr{T}$ taken to be the indicator functions of the sets $S_1, \cdots, S_k$ and with sum of squares (possibly with constant weights) replaced by the above weighted sum of squares with weights $1/P_\theta(S_i)$ depending on θ.

This discrepancy measure can be used in any kind of inference problem; it is in common use for point estimation (minimum chi-square estimation) and testing hypotheses I (the chi-square test). From the inference viewpoint it is its own justification; from the decision viewpoint it has optimum properties as $n \to \infty$. These are discussed and references given in [4]. An asymptotic property of great convenience is that as $n \to \infty$ the distribution of $d^*(H, P_X)$ tends to a noncentral chi-square distribution, making it easy to approximate the cut-off point and the power function of the test. In hypothesis testing, the regions $S_1, \cdots, S_k$ are chosen so that $d(P_\theta, P_x)$ would be expected to be large when θ has values against which we want the test to have high power.

(4) *Kolmogorov–Smirnov methods.* In these methods the discrepancy measure used is given by $d(P_\theta, P_x) = \sup_{x'}|F_\theta(x')-F_x(x')|$ where F_θ, F_x are the cumulative

probability functions corresponding to the probability measures P_θ, P_x respectively. In least squares language this amounts (for real-valued X_1) to taking for $\mathcal{T}$ the indicator functions of the sets $(-\infty, t]$, $-\infty < t < \infty$, and using the discrepancy measure given by $\sup_t \left|E_\theta\, t(X_1) - \sum_{i=1}^n t(x_i)/n\right|$.

The use of this measure, and asymptotic results, are discussed in [1] for hypothesis testing I problems. Wolfowitz uses this measure as an example in point estimation and Wolfowitz's regularity conditions can be used to show that for $n = 1, 2, 3, \cdots$ it gives a consistent sequence of estimators.

From the inference viewpoint this discrepancy measure is its own justification. Like other measures differing from likelihood it has the drawback that the θ value minimizing it for a particular x can make this value of x impossible while other θ values give this x a higher probability.

EXAMPLE. Let $X_1, \cdots, X_n$ be independent, each with Rectangular $(\theta - \frac{1}{2}, \theta + \frac{1}{2})$ distribution, $n = 100$. If $X_1 = 0$, $X_2 = 1, \cdots, X_{100} = 1$ we know for sure that $\theta = \frac{1}{2}$ but the value $\theta = .99$ minimizes the discrepancy measure, and this θ value makes the observed sample impossible.

At first sight this seems to be a serious drawback (from the decision viewpoint it *is* serious, leading to inadmissibility) but actually it may be of little importance. It may happen only with negligibly small probability. Besides, the probability model is intended only as an approximation to reality, and it is common to use probability models that ascribe positive probability to x's known to be impossible (for example, the normal distribution as a model for measurements known to be positive) or on the other hand that assign zero probability to values known to occur. In the above example it might well be argued that the outlying X_1 value should be rejected as an accident and that $\theta = .99$ is a very reasonable guess indeed. That example can be made more realistic, less simple, by spreading out the X_i's that are concentrated on the point 1.

Also discussed in [1] are the Cramer–von Mises tests of hypotheses I which use the discrepancy measure $\int_{-\infty}^{\infty} \{F_\theta(t) - F_x(t)\}^2\, dK(t)$ where K can be more or less arbitrarily chosen, one choice being $K = F_\theta$.

(5) *Nonparametric methods.* There are a very large number of nonparametric hypothesis testing I methods in which the rather arbitrarily chosen discrepancy measure (chosen so that we would expect it to be small when $\theta \in H$, and large for $\theta \in H'$ values against which we want high power) is distribution-free, i.e. has the same known distribution for all $\theta \in H$. This enables us to determine the cut-off point for the test I. Furthermore, it often happens that this distribution approaches a simple limit as $n \to \infty$, giving an easy approximation to the cut-off point for large n. Some examples are: the chi-square test, the Kolmogorov–Smirnov test, the sign test, the Wilcoxon test.

REFERENCES

[1] DARLING, D. A. (1957). The Kolmogorov–Smirnov, Cramér–von Mises tests. *Ann. Math. Statist.* **28** 823–838.

[2] KENDALL, M. G. and STUART, A. (1961). *The Advanced Theory of Statistics*. Griffin, London.

[3] LEHMANN, E. (1950). *Notes on the theory of estimation.* Univ. of California. mimeo.
[4] LEHMANN, E. (1959). *Testing Statistical Hypotheses.* Wiley, New York.
[5] NEYMAN, J. (1962). Two breakthroughs in the theory of statistical decision making. *Rev. Inst. Internat. Statist.* **30** 11–27.
[6] NEYMAN, J. and PEARSON, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika* **20A** 175–240 and 263–294.
[7] NEYMAN, J. and PEARSON, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philos. Trans. Roy. Soc. London, Ser. A,* **231** 289–337.
[8] PLACKETT, R. L. (1949). A historical note on the method of least squares. *Biometrika* **36** 458–460.
[9] WALD, A. (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist.* **10** 299–326.
[10] WALD, A. (1947). *Sequential Analysis.* Wiley, New York.
[11] WALD, A. (1950). *Statistical Decision Functions.* Wiley, New York.
[12] WOLFOWITZ, J. (1957). The minimum distance method. *Ann. Math. Statist.* **28** 75–87.

COMMENTS ON BLYTH'S PAPER

BRADLEY EFRON:[2] Students of statistics who venture into applied work after years of study are often surprised to find out that they have been taught the wrong subject. Faced with a messy batch of actual data, their more experienced colleagues usually will not concern themselves with admissibility, minimax properties, subjective prior distributions, etc.; instead they will get right down to the business of "inferring" what is going on, probably using a combination of the methods Professor Blyth outlines in Section 5 of his paper.

Of course any good applied statistician must keep in mind the principles of theoretical statistics ("power", "Bayes Law", "sufficiency", "unbiasedness", etc.), and bring these principles to bear whenever they are appropriate and feasible. This truism should not be used to conceal the fact that there is a much greater distance between theoretical statistics and applied statistics than between, say, theoretical and applied physics.

Professor Blyth's paper is an attempt to describe mathematically the vulgar form of statistics in common use (which he calls the "inference model"), and show how it differs from the classroom and journal variety of statistics (which he calls "the decision model," a term embracing both the objectivist decision theory of Wald and the subjectivist world of the Bayesians).[3] Since statistical inference is, unlike the weather, something which everyone does and no one talks about, Professor Blyth deserves a great deal of credit for his lucid venture into this difficult subject.

The basic point of Blyth's paper seems incontrovertible; the inference model is a

---

[2] Stanford University.

[3] Equating "theoretical" with "decision" and "applied" with "inference" is a crude oversimplification, which I have used, nevertheless, in order to preserve Professor Blyth's terminology in this note. I hope that the discussion which follows will make my meaning clear.

much looser collection of ideas than the decision model, fundamentally because real life problems usually present themselves to the statistician in a vague and amorphous form. In Blyth's formulation the necessary vagueness of the inference model is obtained by deleting the specific loss functions of the decision model, and substituting an almost arbitrary concept of "closeness" or "agreement" with the observed data. (Hence the often-heard contention of the applied statistician that he believes in "working close to the data").

Although I agree with Professor Blyth that the inference model usually does not have a specific loss function built into it, I cannot accept this as its only distinguishing feature from the decision model. The theoretical and applied models of any discipline customarily differ in just this way, with the latter being a deliberately relaxed version of the former.

In my mind the difference between the theoretical and applied ("decision" and "inference") models of statistics lies at a deeper and more disturbing level. An example, in which the choice of a loss function is not the crucial difficulty, will be helpful in making this point.

A new type of stellar object ("Raysars"), is discovered at Mt. Palomar. A total of 10 Raysars are sighted after a year of careful investigation, and an important quantity, the logarithmic intrinsic brightness L_i, is measured for each one. These measurements are made independently by a technique with a known normal error law

$$L_i \sim \mathcal{N}(\lambda_i, 1),$$

where λ_i is the true log intrinsic brightness of Raysar i. The observed values $L_i = l_1, L_2 = l_2, \cdots, L_{10} = l_{10}$ have a total sum of squares $d^2 \equiv \sum_1^{10} l_i^2 = 20$. An answer is desired to the following question: is $\delta^2 \equiv \sum_1^{10} \lambda_i^2$ less than 20 or greater than 20?

Letting $L = (L_1, L_2, \cdots, L_{10})$, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_{10})$ and $l = (l_1, l_2, \cdots, l_{10})$, we know that $L \sim \mathcal{N}(\lambda, I)$, in the usual multivariance normal notation. An objectivist will argue as follows: Given λ, $D^2 \equiv \sum_1^{10} L_i^2$ has a noncentral χ^2 distribution, $D^2 \sim \chi'^2_{10}(\delta^2)$, with mean $\delta^2 + 10$ and standard deviation $(4\delta^2 + 20)^{\frac{1}{2}}$. The observed value $d^2 = 20$ is unlikely if $\delta^2 \geqq 20$, since even in the case $\delta^2 = 20$, D^2 has mean 30 and standard deviation 10. An exact calculation yields

$$P(D^2 \leqq 20 \,|\, \delta^2 \geqq 20) \leqq .156.$$

The objectivist would decide that δ^2 was probably less than 20, and, if pressed for a quantitative assessment, might estimate δ^2 at 10, the minimum variance unbiased estimate.

Next, let us consider the problem from a Bayesian (subjective) point of view. Considering the novel nature of Raysars, and the fact that log intrinsic brightness varies from -10 to $+25$ for ordinary stars, the Bayesian should not have strong a priori opinions on λ. This leads to the approximate a posteriori distribution

$$\lambda \sim \mathcal{N}(l, I) \quad \text{given} \quad L = l.$$

The a posteriori distribution of δ^2 given $L = l$ is $\chi'^2_{10}(d^2)$, which has mean $d^2 + 10 = 30$ and standard deviation $(4d^2 + 20)^{\frac{1}{2}} = 10$. The subjectivist decides that δ^2 is greater than 20 (with a posteriori probability .844) and perhaps estimates δ^2 as 30, the a posteriori mean.

Where does all this leave our inference statistician, who actually has to answer to the astronomers? If he tries to "stay close to the data", he may report the maximum likelihood estimate of δ^2, which is 20, the likelihood ratio statistic for testing $\delta^2 \leqq 20$ vs $\delta^2 \geqq 20$, which is 1, and conclude that the data is not decisive on the question of whether or not δ^2 is greater than 20. This conclusion is apt to leave our statistician in an uneasy mood, particularly if he has, as a point of comparison, gone through the objectivist and subjective arguments above. Even if he does not believe in minimum variance unbiasedness, for instance, he may note that the maximum likelihood estimate of δ^2 based on D^2 alone is also near 10,[4] and wonder why the seemingly irrelevant extra information provided by the direction of the vector L is changing his estimate so drastically.

The nub of the difficulty, which this example[5] is designed to exacerbate, is that both the objective and subjective decision theorists have very definite and very different "frameworks of replication" in mind when they use probability models. The objectivist asks, "how well would my decision rule work, on the average, if I were to replicate this experimental situation a great many times, always with the same ("true") value of the unknown parameter, but allowing the observations to fluctuate randomly according to the given probability laws?" The Bayesian thinks of his potential replications in the opposite manner: "how well would my decision rule work, on the average, if I were to replicate this experimental situation a great many times, always with the same (observed) values of the measured statistics, but allowing the unknown parameter to fluctuate randomly according to the a posteriori distribution?"

Neither of these frameworks is necessarily applicable to a given experimental situation. The Bayesian concept is often patently inappropriate, particularly in scientific experimentation. The assumptions of the objectivist model are minimal

---

[4] The m.l.e. of δ^2 based on D^2 is 10.6. This is obtained by numerically maximizing the $\chi'^2_{10}(\delta^2)$ density, evaluated at $D^2 = 20$, as a function of δ^2. The representation of a noncentral χ^2 as a Poisson mixture of central χ^2 distributions is useful here. [4]. The m.l.e. of δ based on *all* the data is of course 20, since this is the length of l, the m.l.e. of the vector λ.

[5] Readers will recognize the objectivist portion of the example as a part of C. Stein's construction of an improved estimator for the mean of a multivariate normal distribution [6]. The Bayesian analysis is naive, in that a practical Bayesian would recognize that his seemingly innocuous prior distribution was actually strongly prejudiced against small values of δ^2. If he assumes an improper prior with density $g(\lambda) \propto \|\lambda\|^{-(n-2)}$ he will get approximately the same answers as an objectivist to questions involving δ^2. If the definition of a "practical Bayesian" is one who always chooses his priors to give good objectivist proportions, then there is little pragmatic difference between the objective and subjective viewpoints. Most modern Bayesians would object to this last statement, and give examples where the different viewpoints resulted in different conclusions. In the Raysar example this difference has been deliberately aggravated. For a nice example of Bayesianism shading into objectivism, see Section 3.3.3 of Reference [3].

and usually acceptable in practice, but the conclusions they lead to may be irrelevant. (A blatant example is the confidence interval $(-\infty, +\infty)$ that can arise in estimating the ratio of two means. Another example is given below).

Both the objective and subjective schools have attempted to broaden the philosophical basis of their respective models. These attempts at greater applicability range from the ingenuous ("in his lifetime the statistician would err 5% of the time") to the heroic (subjective probability). Nevertheless it is a fact of life that in many practical situations, *neither* frame is satisfactory. In such situations the difficulty in choosing an appropriate loss function is dwarfed by the fundamental problem of choosing an applicable framework in which to view the experiment and its results.

I consider the "inference model" of statistics to be an amorphous collection of methods which attempts to fill the void lying beyond the philosophical scope of decision theory. The adjective "amorphous" is pejorative only in the sense that no cohesive theoretical framework has yet been able to contain this body of methods (despite such formidable attempts as fiducial inference.) Many of the most fruitful ideas in modern statistics have their origins in the inference model, with a later history of decision-theoretic justification. Examples include maximum likelihood estimation, the analysis of variance, and to some extent nonparametric methods. A very pragmatic definition of inference model statistics is the following: all statistical statements are comparative; in any given experimental situation, no statistical technique can treat both the unknown parameters and the observed values of the statistics as unique. The types of comparative statements possible in the objective or subjective models are often irrelevant or inapplicable to the problem at hand. The inference model is an attempt to make statistical statements anyway.

The next paragraphs are devoted to an example of one important technique inference statisticians have developed for making statistical statements outside of the decision theoretic framework.

The following problem is based on real data which arose in connection with a disease prevalence study. In thirty-six cities a standardized disease prevalence statistic Z_i was obtained by random sampling. The model

$$Z_i \sim_{\text{ind}} \mathcal{N}(\theta_i, 1) \qquad i = 1, 2, 3, \cdots, 36$$

was used as an approximation, which could be assumed to be very good from binomial considerations and the sample sizes involved. Here θ_i was city i's actual disease prevalence rate as compared to the national average, and by definition $\sum_1^{36} \theta_i = 0$. It was desired to know whether or not the prevalence rate was related to the rainfall r_i in each city. The values of r_i were constants obtained from city records. (These were also measured from the nationwide average, so that $\sum_1^{36} r_i = 0$).

The standard "one degree of freedom" test statistic [4] for this situation is

$$S_r = [(\textstyle\sum_1^{36} r_i Z_i)/(\sum_1^{36} r_i^2)^{\frac{1}{2}}]^2$$

which under the null hypothesis of no relation between rainfall and disease prevalence rate has a χ_1^2 distribution. (The test which rejects for large values of S_r

is easily seen to be a uniformly most powerful unbiased test for $\sum_1^{36} r_i\theta_i = 0$ versus $\sum_1^{36} r_i\theta_i \neq 0$.) S_r had an observed value of 13.06 so that the null hypothesis was definitely rejected.

The trouble with this analysis is that it is irrelevant: The value of $\sum_1^{36} z_i^2$ was 144.2, which implied that the vector $\theta = (\theta_1, \theta_2, \cdots, \theta_{36})$ had an estimated length of $(144.2-36)^{\frac{1}{2}} = 10.4$ (with a standard deviation of about one for this estimate). Given that the vector θ has length 10, one can show that it is very difficult to choose the constants r_i so that the statistic S_r is *not* significant.

A much more pertinent question is the following: among all statistics of the form

$$S_R = [(\textstyle\sum_1^{36} R_i z_i)/(\sum_1^{36} R_i^2)^{\frac{1}{2}}]^2$$

is S_r unusually large? Here the observed values z_i are thought of as fixed while the vector $R = (R_1, R_2, \cdots, R_{36})$ takes on all possible values satisfying $\sum_1^{36} R_i = 0$. Without loss of generality we can also assume that $\sum_1^{36} R_i^2 = 1$, since S_R is homogeneous. This suggests a natural measure on the vectors R, namely the measure of "area" on the surface of the 35 dimensional sphere determined by $\sum_1^{36} R_i = 0$ and $\sum_1^{36} R_i^2 = 1$ (normalized so that the measure of the whole sphere is 1). Calling this measure λ, for "Lebesegue," the question becomes, "What is the λ measure of those vectors R having $S_R > S_r$?" A standard geometric argument [1], gives the answer to this question in terms of the Studentized version of S_r,

$$T_r = [(\textstyle\sum_1^{36} r_i z_i)/(\sum_1^{36} r_i^2 \sum_1^{36} (z_i - \bar{z})^2)^{\frac{1}{2}}]^2,$$

which we also recognize as the square of the sample correlation coefficient between r and z. We compute $T_r = .029$ for the data given and note that this is the upper 32% point of a Beta distribution with parameters 1/2 and 34/2. (This computation can also be done using t or F tables. As we have pointed out, the length condition $\sum_1^{36} R_i^2 = 1$ is actually irrelevant since S_R (or equivalently T_R) is homogeneous in R. To answer the question posed, we can choose any distribution of the vector R having a uniform distribution of direction in the hyperplane $\sum_1^{36} R_i = 0$. In particular a multivariate normal distribution in that hyperplane, with mean vector 0 and covariance matrix I, allows us to use standard normal theory as shown.)

We conclude that rainfall does not have any discernible linear relationship with disease prevalence, since if we chose the coefficients R_i completely at random we would get a value of S_R exceeding S_r 32% of the time (the same argument also shows that no low order polynomial function of rainfall has an interesting relationship to disease prevalence in this problem).

This example was intended to show a familiar inference technique in an unfamiliar setting. It is a close cousin to the two-sample permutation test introduced by Fisher. "Randomization as a Basis for Inference" is a familiar and apt name for the basic idea, which is to compare the observed value of the statistic not with its theoretical distribution under the null hypothesis, but rather with the distribution of values which would arise if the null hypothesis were selected in a random manner. Rejecting the null hypothesis is interpreted to mean that the experimenter

actually knew some relevant information about the structure of the experiment when he set up the null hypothesis.

Although the randomization model can be forced into a decision-theoretic mold, the type of comparative statements it makes seems to me to be of a fundamentally different nature than those of either the objective or subjective models. I believe that future breakthroughs in the science of statistical inference will be made at this fundamental level of increasing the catalogue of useful frameworks in which to view statistical data.

REFERENCES

[1] HOTELLING, H. (1967). The behavior of some standard statistical tests under nonstandard conditions. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1** 319–360.

[2] RAIFFA, and SCHLAIFER, R. (1961), *Applied Statistical Decision Theory*. Harvard Univ. Press.

[3] RAO, L. R. (1965). *Linear Statistical Inference And Its Applications*. Wiley, New York.

[4] REISOL, O. (1954). Tests of linear hypotheses concerning binomial experiments. *Skand. Aktuarietidskt.* **37** 38–59.

[5] STEIN, C. (1966). An approach to the recovery of interblock information in balanced incomplete block designs, in *Festschrift for J. Neymann*. Wiley, New York, 351–366.

COMMENTS ON BLYTH'S PAPER

KEI TAKEUCHI[6]: The comparison between "inference" and "decision" in mathematical statistics is a theoretically interesting and practically important problem. But as the past controversies for many decades over this problem have shown, it requires much subtlety for the discussion of this delicate issue to be fruitful.

At the very beginning of the discussion, the term "inference" should be defined carefully. Sometimes futile controversy started simply from the misunderstanding or misuse of such terminology. And the logico-methodological aspects other than abstract mathematical formulation of the concepts should be carefully examined. In a pure mathematical model, "inference" can be reduced to a special kind among a wider class of statistical decision problems, regarding the problem as a choice among a set of possible statements and considering the possible "loss" due to wrong statements. The motivation of the first paper of A. Wald on statistical decision functions [12] was to give a unified approach to several forms of inference like testing and estimation by such a formulation. The pertinent problem here is not whether this formulation is really appropriate or not, but whether there exists any way of looking at the problem of inference which might clarify logical difference between inference and decision, which is missed by the decision function approach. In this sense the essential difference between "inference" and "decision" should not be characterized by the contrast between a set of possible statements against

[6] Courant Institute of Mathematical Sciences.

concrete practical decisions, or between error probability against monetary loss or utility. The distinction between accepting or rejecting a hypothesis and taking this or that decision, or between the errors of the first and the second kind and the expected loss due to the wrong decision is not a basic difference from this standpoint.

The problem of "inference" proper has been formulated, rightly in my opinion, as the problem of reasoning from the given observed sample to the population or the parameter, whereas the "decision" problem has been considered as the choice among possible decision procedures or rules characterized by their probabilistic behaviors or average performances under possible probability distributions. R. A. Fisher, especially in his last book [5], emphasized this point very strongly, and contended that statistical inference has nothing to do with the "long-run frequency," that is the probabilistic behavior of the procedure. He tried to construct a system of the logic of statistical inference which was to be entirely different from the Neyman–Wald type theory of statistical decisions. As it has turned out, Fisher's viewpoint has not been widely accepted by mathematical statisticians, nor has his effort been entirely successful. There is an appreciable group of statisticians who deny the possibility of such a logical system, including, notably, J. Neyman [8], [9]. There are some others who recognize the necessity of such a logic, but are not satisfied with Fisher's approach. And some of the people who call themselves Bayesian statisticians have in some sense unified inference and decision by introducing subjective (or personal) probability.

This is not a proper place to discuss such problems fully, but briefly summarizing my own viewpoint, I think that inference proper in the sense above is a real problem in almost all of the applications of statistical methods in scientific research and even in most of the more practical applications in management or engineering problems, where the issues are an understanding of the actual processes and interpretation of the data other than simple choices among a predetermined set of actions. I am not really satisfied with the approach of R. A. Fisher, because his logical concepts are difficult to interpret and to justify, nor with the Bayesians because actually what is required in scientific research is not the consistency of subjective probability but rather objective statements, whatever the term "objective" may mean. However, it cannot be denied that R. A. Fisher and leading Bayesian statisticians, such as Savage and Lindley, have contributed much to clarify important points in this problem area [7], [10], [11].

The author seems to consider "inference" in the sense above, i.e. reasoning from the observed data to the parameter, but he is not consistent in that he chooses as the criteria of "inference" not only the distance from the observed frequency distribution to the possible population distribution, which is a function of given data and the theoretical distribution only, but also such characteristics as the size of the test, which is based on the probabilistic property of the procedure calculated for the set of points including non-observed sample points. The last point was vigorously objected to by Fisher, who insisted that the "inference" should not depend on the nature of unobserved points (cf. also Barnard [1]). A more sophisticated formulation of this viewpoint is the so-called "conditionality" principle

as discussed by Birnbaum [2] and others, which implies that if there exists a subset of the sample space which is clearly distinguishable in some sense, including the sample point, the "inference" should refer to this subset instead of the whole sample space. Although the "conditionality" principle sometimes leads to a contradiction, and there is no wide agreement among statisticians how to solve the problem, it is, I believe, generally admitted through many examples that unwarranted applications of unconditional probability may lead to absurd conclusions if they are considered as the solutions of inference problems (Cox [4]). And it is clear that "inference" methods cannot be based directly on probabilistic properties of the procedure; or at least it is necessary to supplement them with some other considerations or principles.

It seems to me a little strange to read a paper which discusses the problem of "inference" but does not mention anything, even negatively, about R. A. Fisher, the Bayesians, or others who have discussed the problem for many years, at times quite heatedly. A more serious defect is that the author does not pay attention to a few principles of "inference" which are widely accepted, or at least thought to be worthy of consideration. The first among these is the principle of "sufficiency" which is agreed upon by almost all statisticians interested in "inference" even though they may differ about its precise meaning and interpretation. The second is the "likelihood" principle which is accepted by Bayesians and some others, but not by many belonging to the Neyman–Pearson school. The principle of "invariance" may also be mentioned, which sometimes justifies Fisher's fiducial approach, as was proved by Fraser [6]. Anyone may of course reject any or all such "principles" but, he should not do so simply by ignoring them. I feel that the author should have discussed the problem more carefully before violating the sufficiency principle in order to defend randomization in the inference, a choice which is rejected by most statisticians as a legitimate technique of "inference".

More particularly, in the discussion of testing problems, it is hard to understand why the author did not spare even one sentence for the difference between Fisher's and Neyman–Pearson's approaches to this problem, because not only has it been the object of notable controversy over the years, but it really touches upon the basic issue of the "inference" and has serious practical significance. Should the "size" of the test be always preassigned or can the "level of significance" be determined from the sample? This problem is not purely an academic one, and although no satisfactory mathematical theory for the latter approach has yet been established, in most of the real applications, it is usually considered to be, I think, the more appropriate approach.[7] Anyway, when one discusses the problem of testing hypotheses from the viewpoint of "inference", this is one of the most basic problems, which cannot be evaded without mention.

There may be serious doubt whether any clear-cut systematic theory of

[7] Even Neyman himself allowed for the sample significance levels in applied problems: Statistical problems in science. The symmetric test of a composite hypothesis (mimeographed) Lecture, 1969 IMS Annual Meeting.

"inference" proper could be constructed. If one is not willing to submit to arbitrariness, perhaps the whole theory must remain somewhat vague in its implications, as was shown in Birnbaum's discussion [3] of the concept of statistical "evidence." But the whole situation is not so vague or chaotic as to warrant the extremely wide range of arbitrariness allowed in Professor Blyth's paper.

For the theory of "inference", logical consistency and coherence are more important than for "decision" theory because it must stand on its own system of logic as well as on the considerations of general use for many kinds of decision problems, and logical coherence is the minimum requirement for such a system. For example, in the decision case, one can take an intermediate point as an estimate between two possible parameter points when it is difficult to decide; although it is certain that the estimate cannot be the true value, it may give a smaller expected error than the one which chooses only between two possible cases. But for inference such a procedure may give seriously misleading information about the situation; one should consider one or another of the possible cases, or possible inadequacy of the evidence. From this viewpoint the author's treatment of the example for the case of uniform distribution seems to be unfortunate, because the logical incoherence should not be lightly accepted, and the smallness of the probability of such an occurrence and the inexactness of the model has nothing to do with it here.

The inference procedures should be efficient in utilizing the information contained in the sample; more strictly so than the decision case, because if one uses only a part of the information contained in the sample, the rest of the information in the sample may provide logically contradictory conclusions, and also because the inference is for general use but not for a specific decision problem, where some part of the information could safely be discarded when it hardly affects the expected loss. From this viewpoint, the Wolfowitz type minimum distance method does not necessarily provide us with an efficient or even nearly efficient procedure. Consistency in large samples is a very weak condition, and when the asymptotic efficiency is not high, which is usually the case for various distance methods, no consistent procedure can be regarded as a good method of "inference". Also the large sample situation sometimes obscures the logical differences or difficulties of several approaches, since under a set of regularity conditions, asymptotic normality and/or asymptotic efficiency of maximum likelihood methods leaves little doubt about the desirable procedures to be used either in decision or inference. The real test lies, as it has at least since the 1900's, in the small sample situation.

REFERENCES

[1] BARNARD, G. A. (1949). Statistical inference. *J. Roy. Statist. Soc. Ser. B.* **11** 115–149.

[2] BIRNBAUM, ALLAN (1962). On the foundations of statistical inference. *J. Amer. Statist. Assoc.* **57** 269–326 (with discussion).

[3] BIRNBAUM, ALLAN (1969). Concepts of statistical evidence, in *Philosophy, Science and Method: Essays in Honor of Earnest Nagel.* St. Martin's Press, New York.

[4] COX, D. R. (1958). Some problems connected with statistical inference. *Ann. Math. Statist.* **29** 357–372.

[5] FISHER, RONALD A. (1959). *Statistical Methods and Scientific Inference*. Oliver and Boyd, Edinburgh.
[6] FRASER, D. A. S. (1961). The fiducial method and invariance. *Biometrika* **48** 261–280.
[7] LINDLEY, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, 2 vols. Cambridge Univ. Press.
[8] NEYMAN, JERZY (1957). Inductive behavior as a basic concept of philosophy of science. *Rev. Inst. Internat. Statist.* **25** 7–22.
[9] NEYMAN, JERZY (1962). The two breakthroughs in the theory of statistical decision making. *Rev. Inst. Internat. Statist.* **30** 11–27.
[10] SAVAGE, L. J. (1954). *Foundations of Statistics*. Wiley, New York.
[11] SAVAGE, L. J. *et al.* (1962). *The Foundations of Statistical Inference*. Methuen, London.
[12] WALD, ABRAHAM (1939). Contributions to the theory of statistical estimation and testing hypotheses. *Ann. Math. Statist.* **10** 299–326.