REVIEW OF THE INTERNATIONAL
STATISTICAL INSTITUTE
Volume 30: 1, 1962

TWO BREAKTHROUGHS IN THE THEORY OF STATISTICAL DECISION MAKING[1]

by

J. Neyman
Statistical Laboratory, University of California, Berkeley

INTRODUCTION

The present paper was prepared as a contribution to the organized discussion at two sessions of the statistical meetings held in New York in December, 1961. Both sessions, with several speakers, were concerned with the same domain symbolized by reference to Bayes. The papers offered for discussion are by Allan Birnbaum [1], by Jerome Cornfield [2] and by Herbert Robbins and Ester Samuel [3]. These three papers are not isolated studies but members of three different streams of publications, each with a different philosophical outlook. During the year 1961 Birnbaum alone contributed at least four different papers on the subject.

The purpose of the present paper is to examine and to compare the three different schools of thought with reference to a certain historical perspective. In this historical perspective it would appear that we are now witnessing a very important development in statistical theory that, for some reason, failed to attract the attention of broad circles of statisticians.

CYCLES IN THE DEVELOPMENT OF SCIENTIFIC DISCIPLINES

In the development of particular scientific disciplines it is convenient to distinguish a number of different "fronts" and consider them separately. In mathematical statistics a number of such fronts may be distinguished, for example the front of sequential analysis, the front of experimental design, etc. In a study of the development of a given front it is interesting to distinguish certain recurring cycles of progress. As I see it, each cycle is roughly divided into three distinct phases.

Phase (i) is the phase of vague feelings experienced by a number of research workers that the contemporary stage of the domain under study is somehow unsatisfactory and that, in order to achieve real progress, a new start from some new angle is necessary. If the incipient cycle is of considerable importance, then the search for the new start is not easy, and the workers concerned may be observed wandering about, occasionally in circles, solving here and there a minor problem, but generally unsatisfied and apt to be dogmatic and to engage in fruitless polemics.

In pure mathematics, on the front of integration, a phase of this kind prevailed in the early part of this century. One of its marks was the disputes surrounding "pathological" functions which declined to have an integral or, even worse, while being continuous had no derivatives. The easy way out, adopted by some mathematicians of that period, was to postulate away the difficulty by ruling out the pathological items and by declining to call them functions. Similar disputes were frequent on the front of probability and, particularly, on the philosophical front of the relation between probability theory and the physical world.

[1] This paper was prepared with the partial support of the Office of Ordnance Research, U.S.A. Grant (DA-ARO(D)-31-124-G83).

Phase (ii) of the developmental cycle begins when someone manages to find a formulation of the problem and produces at least a partial solution. This is a breakthrough. The dogmas frequently advanced in phase (i) become pointless, and a new fruitful field of research is opened. One might expect the event to be greeted by general applause and by an immediate reorientation of research. In due course this ordinarily does happen, but not at once, and the new idea takes quite some time to break the established routine of thought and to penetrate the minds of the rank and file of scholars. On the integration front in pure mathematics the breakthrough was accomplished by Lebesgue. The majority of my own university teachers contempleted the new concept with suspicion and enmity. However, Lebesgue's ideas took hold and, after a decade or two, not only integration, but the whole body of mathematics began to look different than it did before Lebesgue. Similarly, in probability and on the front of its relation to reality, there was a tremendous change after Kolmogorov's axiomatization on the one hand and after von Mises' philosophical writings on the other.

The beginning of phase (iii) of the developmental cycle is difficult to pinpoint. The phase itself is marked with more or less general recognition of the basic idea produced by the earlier breakthrough and with a rush to solve the outstanding problems. Various generalizations and particularizations of the basic concepts are offered and various applications to other domains developed. The third phase of the cycle climaxed by the concept of Lebesgue integral lasted for a long time. In probability, the third phase of the Kolmogorov cycle is in full bloom right now, with Kolmogorov's own active participation.

BAYES' FRONT IN THEORY OF STATISTICS

In this section I shall sketch the essence of the "Bayes' front" in mathematical statistics. This will be done in two forms, first mathematically and next using several examples taken from applied fields.

Consider a set Θ of items to be described as parameter points. To each parameter point $\theta \varepsilon \Theta$ there corresponds a distribution P_θ over some space E described as the experimental space. It is given that an observable random variable X with values in E follows one of the distributions P_θ with unknown value of the parameter θ. In connection with some contemplated action it is important to know that value ϑ of the parameter θ that corresponds to the observable X. The appropriate experiment is performed yielding $X = x$. What can one say about ϑ? Denote by A the set of all possible action. What should be the rule $a(x)$ determining the action to be taken when the observed value of X is X? Still in other words, how best to define a function $a(x)$ on E, with values in A so as to optimize, in some sense, the selection of the action to be taken in accordance with the results of observations?

It will be noticed that none of the questions asked constitutes a mathematical problem. The PROBLEM of the Bayes' front consists precisely in reformulating these questions so as to obtain mathematical problems that can be solved using the basic datum: Θ, E, P_θ and A, with an appropriate formalization of what is intended to achieve.

Here are a few examples of applied problems which will be repeatedly referred to in further discussion. In order to avoid the suggestion of any judgment of the relative importance of the domains of application, the examples are presented in alphabetical order.

Example 1. Acceptance sampling. A big consumer organization purchases certain items delivered in lots. N is the number of item in a lot, n is the number of those items randomly selected for inspection, X is the number of defectives in the sample and θ the number of defectives in the lot. The parameter space Θ consists of $N+1$ integers 0, 1, 2, . . . , N. The set A of actions contemplated may include only two elements "accept lot" and "reject lot", or may be more comprehensive. The general set up is familiar to the readers.

Example 2. Astronomy. A given galaxy G may be a member of a physical system of multiplicity θ. In other words, G may be a single galaxy and $\theta = 1$. Or G may be a member of a physical pair in which case $\theta = 2$. Alternatively, G may be a member of a triplet, quadruplet, etc., with corresponding values of $\theta = 3,4, \ldots$ In this case the parameter space Θ is the set of positive integers. The experimental space E may be taken as comprising all possible combinations X of various measurements regarding G and its neighbour galaxies on the photographic plate. If a galaxy G is a member of a system of multiplicity $\theta = 2$, then G and its companion can be used, jointly with other similar pairs, for the estimation of the average mass of galaxies. Denote by M_2 the method of doing so applicable to physical pairs. If $\theta = 3$ so that the galaxy G is a member of a physical triplet, then method M_2 is not applicable and, instead, one may use method M_3, etc. We assume that, with the available observations, single galaxies cannot be used for estimating the average mass. The set A of contemplated actions regarding G is: (1) drop G as a single galaxy; (2) use G and its companion for study by method M_2, etc. The statistical problem is how best to use the available observations to compile lists of double, triple, etc. galaxies.

Example 3. Diagnostic tests. The value of a parameter θ is an important characteristic of health (e.g. average blood pressure, blood sugar content, etc.) Accordingly, prior te being examined by a physician, each patient is routinely subjected to a series of diagnostic tests meant to provide data for the estimation of θ. I need not emphasize that what is observed at these tests is not the value of θ but only the value of a random variable X whose distribution depends upon θ. Here again θ has a certain range Θ of possible values and, at least in some cases, the distribution of X given θ may be assumed known.

Example 4. Screening chemical compounds for curative properties against cancer. In his contribution to the Fourth Berkeley Symposium Nathan Mantel described an impressive research program now going on in several centers intended to determine those chemical compounds that offer a promise of cure of cancer. The intention is to try each of a large number of compounds, and it is this large number of treatment that creates a particular difficulty: because of the space limitation, the number of replicates used must be limited. The procedure with each treatment may be visualized as follows.

A small number n of animals are exposed to some carcinogen and then are given a treatment T. Some experimental random variable X is then observed for each animal, perhaps the survival time or the number and sizes of tumors in a given organ observable after so many weeks. Here θ stands for the effect of treatment T, a parameter intervening in the distribution of X. The statistical problem consists in devising a method of using the observable X's in order to select for closer study those of the tested chemicals that offer promise as remedies against cancer.

The "very classical" solution of these and of thousands of similar problems is provided by the famous Bayes' formula. In addition to the data mentioned at the outset of this section, namely the parameter space Θ, the experimental space E and the conditional probability distribution P_θ of X, defined over E for every $\theta \in \Theta$, Bayes' formula requires an extra datum. This extra datum is the so called *a priori* probability distribution of θ in Θ. With deference to Bayes this distribution will be denoted by B.

If B is given then, for any observed $X = x$, Bayes' formula gives the *a posteriori* distribution of θ, or, in simpler language, the conditional distribution of θ given $X = x$, say

$$(1) \qquad dG(\theta \mid x) = \frac{P_X(x \mid \theta)\, dB(\theta)}{\int P_X(x \mid \theta)\, dB(\theta)}$$

When $X = x$ is observed, then formula (1) can be used in any of the familiar manners in order to answer the questions posed. Various hypotheses can be tested by computing their posterior probabilities. One may estimate θ by computing its most probable or its expected value, etc.

The trouble is that, in many cases met in practice, while the conditions of the problem considered imply the randomness of the observable variable X and, even imply the distribution of X given θ, they involve no such implication regarding θ, and formula (1) cannot be used. This is not an absolute rule, and the following counter-example may clarify the ideas.

Example 5. A Mendelian experiment. Consider an inheritable characteristic Γ known to depend upon a pair of genes A and a, of which the first is the dominant. As a result, the individuals with compositions AA and Aa are indistinguishable. Suppose that an individual I is the progeny of a pure recessive aa and of another individual that looked dominant. Suppose I bears the dominant aspect of Γ and, therefore, must be a hybrid Aa. I is allowed to selffertilize and produces a progeny I_1, which also looks dominant. The genetical composition of I_1 is uncertain and may be either $\theta_1 = AA$, or $\theta_2 = Aa$. The Mendelian Theory implies that

$$(2) \qquad P\{I_1 : AA\} = \tfrac{1}{3}, \qquad P\{I_1 : Aa\} = \tfrac{2}{3}$$

In order to estimate the genetical composition θ of I_1 an experiment is performed: I_1 is allowed to selffertilize and produces n progeny. Among these progeny the number of identifiable recessives is X and the same Mendelian Theory implies the distribution of X given θ. Specifically,

$$(3) \qquad P\{X = 0 \mid \theta_1\} = P\{X = 0 \mid I_1 : AA\} = 1,$$

$$(4) \qquad P\{X = k \mid \theta_2\} = P\{X = k \mid I_1 : Aa\} = \binom{n}{k}\left(\frac{1}{4}\right)^k\left(\frac{3}{4}\right)^{n-k}, \; k = 0, 1, 2, \ldots, n.$$

Here, then, the conditions of the problem, namely the simple Mendelian Law, that imply the distribution of X given θ as in formulas (3) and (4), imply also the probability distribution of θ on its space $\Theta = (\theta_1, \theta_2)$. As a result, Bayes' formula (1) can be applied without any extraneous assumption.

This is not so in the four earlier examples selected so as to illustrate the range of possibilities. Contrary to what prevails in Example 5, the earlier examples involve no single chance mechanism implying both the distribution of the observable X and of the unobservable θ. In the astronomical example 2 and in the example 3 concerned with diagnostic tests, an extra structure is easily visualized implying that θ is a random variable. In one case one might think of a chance mechanism being the origin of multiple systems of galaxies. In the other case, one may visualize a steady clientele of a clinic, recruited from a particular social stratum subject to a fixed system of various "risks". In both these cases I expect little reluctance in considering that the relevant θ is a random variable. However, in order to apply the Bayes' formula, the randomness of θ is not sufficient. One must also know the *a priori* distribution of θ and the datum in examples 2 and 3 does not contain anything regarding this distribution. Examples 1 and 4 are even more difficult, because any attempt at treating θ as a random variable seems to require not only a very extensive imagination, but also some violence against common sense.

In reviewing the five examples quoted, it will be seen that it is the first four that are illustrative of the usual practice. The fifth example is rather special, if not entirely artificial. It is true that in atomic and nuclear physics one can find somewhat similar situations, but these are rare.

The above circumstances were noticed long ago, and there developed what I call the Bayes' front in mathematical statistics. It is characterized by an indistinct, but frequently strong feeling that it is possible to devise a mathematical theory of using the observable X so as to provide justifiable selection of actions (or conclusions, or decisions) regarding θ even in those cases where the datum of a problem does not include the *a priori* distribution of θ. Innumerable minds, great and small, did their best to produce a solution. The remedies offered were many, and the following list is incomplete and is meant only for illustration.

Before proceeding to this list, a remark regarding the terminology adopted in this paper is in order. Some of the colleagues who read the first draft of this paper suggested that the term "Bayes' front" may be misleading and that a better term may be "problem of statistical inference." My reason for preferring "Bayes' front" is that the word "inference" appears associated with a particular method of approach to the "Bayes' front," contrasted with another method symbolized by terms "inductive behavior" and "statistical decision making," a distinction that appears important. The terminology adopted will make it convenient to speak of the "inference approach to the Bayesian problem" contrasted with the "behavioristic" or "decision making approach" to the same problem.

ATTACKS ON THE BAYES' FRONT

I wish to begin with a particular attach on the Bayes' front for which I personally have great respect. This attack was made, independently and practically simultaneously, by two great scholars, Serge Bernstein in Russia and Richard von Mises in Germany. I do not know when Bernstein published his result, but I learned it in Bernstein's lectures in 1915 or 1916. The result of von Mises was published in 1919 [4]. Both results give a theorem to the general effect that, if the *a priori* probability density of θ is continuous then, as the number of independent observations is increased, the *a posteriori* distribution of θ given by (1) tends to a calculable limit, independent of the *a priori* distribution. Thus, if in a given case there is no reluctance to consider θ as a random variable with a smooth probability density, and if one performs a large number of observations, it is more or less immaterial what function one substitutes instead of B. Through this theorem Bernstein and von Mises provided interesting information on the behavior of Bayes' formula with reference to a certain set of conditions.

From the applied statistical point of view, it is obvious that the Bernstein-von Mises approach solves some of the difficulty on the Bayes' front, but is far from removing it. First of all, in an overwhelming number of cases the circumstances limit the number of observations severely. In the astronomical example 2, there is possible just one observation on X characterizing a given galaxy (assuming that this observation is errorless). In medical practice it is rare that the technicians take more than one single sample of blood, etc. But then what about those frequent cases where one is reluctant to consider θ as a random variable?

The next category of attacks on the Bayes' front I am going to mention is "the easy way out". This category is composed of innumerable attempts to postulate away the difficulty. The postulates proposed are of two different kinds. One is exemplified by the so called principle of insufficient reason. This is to the general effect that, if one is ignorant of the *a priori* distribution of θ, one "has the right" to assume that it is uniform over Θ. A modification of this principle adopted by certain authors consists in substituting in (1) instead of B an arbitrarily selected function other than that implying uniformity over Θ. In certain cases, similar procedures are surrounded by an impressive array of axioms, terminology and deductions. However, in all cases of the kind discussed, one is led to Bayes' formula with a function B not implied by the nature of the problem considered.

The characteristic feature of the works of the kind described is that, contrary to what we have seen with regard to the Bernstein-von Mises theorem, they do not provide new information. Instead, the writings of particular writers, so to speak, "authorize" us to use certain substitutes for the prior distribution when this is either unavailable or nonexistent. Some other writers go farther than that and more or less prescribe the use of substitute *a priori* distributions. As I see it, whether the form of the dictum is just permissive or normative, it is dogmatic and solves no problem.

Another kind of the "easy way out" treatment of the problem advanced before World War II, and still being advanced now, begins by giving up the use of Bayes' formula whenever the problem at hand does not imply the *a priori* distribution of θ, which I consider to be a step in the right direction. However, the next step made appears to me questionable. It consists in the invention of certain functions, say $L(x, \theta)$, and in adopting them as "new measures of confidence or diffidence" in the given value of θ when the observations give $X = x$. Here again some of the authors concerned do not stop at using the invented functions themselves, but emphatically prescribe the use of these functions by others. When someone asks the question "Why?" the reply is usually lengthy, may involve references to "Rational mind" and various axioms, but studiously avoids giving information as to any verifiable consequences of adopting the given receipe or of declining to do so. Here again, I see distinct marks of dogmatism.

Traditionally, the two approaches to the Bayesian problem just described are connected with the term "inference". This connection can be traced to the pre-von Mises interpretation of probability as a measure of confidence rather than as a conceptual counterpart of relative frequency. The sensations of confidence and uncertainty being subjective and complex, it is not unnatural for different individuals to formalize the measure of confidence or uncertainty in a variety of ways and to come up with different sets of axioms or postulates as described above.

In the present paper the term "inferential theory" and its derivatives will be used to describe the attempts to solve the Bayes' problem with a reference to confidence, beliefs, etc., through some supplementation of the usual basic datum of the problem the parameter space Θ, the sample or the experimental space E, and the distribution of the observable X defined over E for every $\theta \varepsilon \Theta$, with the possible further specification that θ itself is a random variable. The supplementations I have in mind are either a substitute *a priori* distribution of θ or a new measure of uncertainty.

As already mentioned, in my opinion, the inferential theory solves no problem but, given a degree of earnestness of the authors, may create the illusion that it does.

The breakthrough of the prewar developmental cycle occurred in the 1930's with the following leading ideas.

(a) If one takes the trouble of performing experiments leading to $X = x$ and then doing some calculations, this presupposes the desire to avoid errors.

(b) In the set of conditions described, specifying Θ, E, P_θ, and A, but nothing else, errors in judgments (or actions, or decisions) regarding ϑ are unavoidable.

(c) Any judgment (or action, or decision) taken as a result of observing $X = x$, is a function of X, which is a random variable. Therefore, this action itself, say $a(X)$, has the character of randomness, with a distribution computable from the P_θ specifically included in the conditions of the problem.

(d) With reference to (b) it appears interesting to inquire about the relative frequency, or probability, with which a given function $a(X)$ will lead to erroneous conclusions regarding ϑ. And this leads to the final question:

(e) Given Θ, E, P_θ and A, how to define a function $a(X)$ from E to A so that it will be a random variable for all $\theta \in \Theta$ and so that its use would guarantee small probabilities of errors in judgments (or decisions or actions) regarding ϑ?

Of the five points listed it is point (c) that is the decisive one. It is the realization that $a(X)$ is a random variable that broke the stagnation of the preceding period and the routine of thought established by the continuous preoccupation with the Bayes' formula. The random variables commonly considered at the time were numerical random variables or their sets and an explicit consideration of a random variable with "values" described as "assertions" or "decisions" or "actions" was a novelty that many contemporary statisticians found difficult to assimilate.

As is well known, the period marked by the above considerations culminated in the memorable works by Abraham Wald [5], [6]. The characteristic feature of studies in the particular direction described is a complete absence of any reference to "confidence". The principal subject of study became the functions $a(X)$, the so-called statistical decision functions, and their properties ("performance characteristics") with reference to the variously understood and variously evaluated "errors", or "losses" incurred due to "errors".

With reference to the possibility that the parameter θ may be a random variable with an unknown distribution B, statistical decision functions were sought (e.g. confidence intervals) with properties independent, or approximately independent, of B. However, the main preoccupation of the period was the case where θ is not a random variable. This particular point will be referred to later.

The findings of the theory motivated by considerations (a) to (e) may be symbolized by the following: in the given conditions Θ, E, P_θ, A, the probabilities of errors (or the upper bounds of such probabilities, etc.) cannot be less than a stated limit and this limit can be achieved by using the specified decision function, say $a^*(X)$.

Because of the explicit references to the consequences of adopting this or that statistical decision function, that is of following this or that rule of adjusting decisions to the observations the theory described may be called "behavioristic" or "theory of decision making". It is hoped that this terminology will emphasize the contrast with the "inferential" approach defined earlier.

Currently, many of the concepts of the behavioristic theory of statistics are in frequent use in many domains. However, their acceptance was not immediate. The reverberations of the now old disputes have been recently reviewed by Pitman [7].

NEW DEVELOPMENTAL CYCLE ON THE BAYES' FRONT AND THE FIRST BREAKTHROUGH

The breakthrough of the 1930's just described created a new field of study with a number of "fronts" of its own. Since World War II we have witnessed several important "breakthroughs" on these new fronts. The first seems to be due to Charles Stein [8] whose two-sample procedure insures a predictable power function in cases where this was thought to be impossible. Next came the breakthroughs of Tukey [9] and Scheffe [10] on the front of the multiple comparison problem. There were others. Several priority surprises occurred. The possibility of the behavioristic treatment of the point estimation problem was traced to Laplace and Gauss [11]. Confidence intervals for the binomial p were found to have been discovered by E. B. Wilson [12], who used a conceptual structure of his own. A faultless statement of the problem of confidence regions, and its solution, has been found in a short paragraph of an early paper by Hotelling [13].

However, on the Bayes' front proper it was not "all quiet".

For example, at the Washington meeting of the International Statistical Institute of 1947 and, two years later, at the Paris International Congress of Philosophy of Science, the Italian statisticians headed by Corrado Gini complained that the "Anglo-Saxon School of Statistics" neglects probabilities *a priori*. They were joined by others who protested that, in many cases similar to my Example 3, previous experience creates in the experimenter a strong feeling that some of the values of the parameter θ are distinctly more probable than others and that this experience is entirely ignored by tests of hypotheses and by confidence intervals. The complaints of the above kind were frequently quite loud, but for a long time no specific solution was offered. We lived through phase (i) of a new developmental cycle on the Bayes' front.

The breakthrough relating directly to the utilization of previous experience occurred in the paper by Herbert Robbins published in 1955 [14]. With reference to situations exemplified by my Examples 2 and 3 (diagnostic tests in a clinic, multiple galaxies), Robbins considered as given that the parameter θ is a random variable but with an unknown distribution B. In addition, Robbins assumed that the statistician has to deal with a sequence of cases characterized by independent and identically distributed pairs of random variables (θ_i, X_i), $i = 1, 2, \ldots$. Each θ_i was assumed to have the same (unknown) marginal distribution B and, given θ_i, the corresponding variable X_i was assumed to have the known conditional probability density $P_{X_i}(x \mid \theta_i)$. The next assumption was that the values of $\theta_1, \theta_2 \ldots, \theta_n$ are all unobservable and that the observations are limited to the values of $X_1, X_2, \ldots, X_n$. The problem was to use all these observations, combined with the postulated existance of B, in order to estimate θ_n. Robbins' term to describe this problem is the Empirical Bayes Problem. It will be seen that the conditions contemplated correspond exactly to the situation in a clinic described in my example 3.

In order to see this consider a physician P working simultaneously at two different clinics A and B, meant for different clientele. Suppose that the two clinics use the services of the same testing laboratory. Consider a case where on the same day the physician P has two patients, one at A and the other at B, with identical outcomes of a certain diagnostic test for which a decision theoretical estimate of θ was calculated. It is easy to visualize two different situations developing depending upon whether the physician is new to the two clinics or has a few years' experience in both. In the first case the estimate of θ handed to the physician is likely to affect his diagnosis in about the same manner in clinic A and in clinic B. However, if the previous experience of the physician indicated that among the clientele of A a certain disease for which the estimated θ is an important indicator is much more frequent than among the clientele of B, then there is little doubt that the role of the same estimate of θ in the physician's thinking in clinic A will be different from that in B. In other words, it may be expected that the previous experience of the doctor will combine somehow with the outcome of the test to form a diagnosis of the patient's state of health. Here, then, we have a situation where, in a non-systematic way and perhaps subconsciously, a decision theoretical estimate of θ is, so to speak, subjected to a distinctly Bayesian correction. This correction is performed using a rough estimate of the distribution *a priori* that the physician forms using his earlier experience. The question arises whether this same operation can be performed consciously and in a systematic manner. This is exactly the question that was asked, and answered, by Robbins. I should mention that an almost identical question was earlier asked by Richard von Mises [15]. However, von Mises' answer, in the form of certain inequalities, appears much less satisfactory than Robbins', somewhat in the style of phase (i).

Approaching the problem as one of point estimation, Robbins adopted the usual mean square error as his measure of accuracy (or as his risk function). If the *a priori* distribution B of θ were known, then the estimate of θ_n minimizing this risk would have depended only on the value x assumed by X_n and would have been given by, say

$$(5) \qquad \hat{\theta}(x) = \frac{\int \theta\, p_X(x \mid \theta)\, dB(\theta)}{\int p_X(x \mid \theta)\, dB(\theta)}$$

which is the conditional expectation of θ given $X = x$. Robbins' formulation of the Empirical Bayes' Problem is: Find a function of $X_1, X_2, \ldots, X_{n-1}$, and of $X_n = x$, say $\tilde{\theta}_n(X_1, X_2, \ldots, X_{n-1}, x) = \tilde{\theta}_n(x)$ which is a consistent estimate of $\hat{\theta}(x)$. If more than one such function is available, determine the one which is in some sense best.

Naturally, the solution of the problem must depend on the conditional distribution of X given θ. Robbins' paper is chiefly concerned with the case where the observable variable is discrete. In this case the estimation of the denominator in (5) is trivial. This denominator is simply the marginal probability $P\{X = x\}$ that $X = x$. Let $N(n, x)$ denote the number of those variables $X_1, X_2, \ldots, X_{n-1}$ that assume the specified value x. Then the quotient $N(n, x) / (n-1)$ is an easy consistent estimate of $P\{X = x\}$. Consistent estimation of the numerator in (5) depends directly on the distribution $p_x(x \mid \theta)$. When X is a Poisson variable with expectation θ, we have

$$(6) \qquad \int \theta\, p_x(x \mid \theta)\, dB(\theta) = \frac{1}{x!} \int \theta^{x+1} e^{-\theta}\, dB(\theta) = (x+1) P\{X = x+1\}.$$

Thus, (6) can be consistently estimated by $(x + 1)\, N(n, x + 1) / (n - 1)$. Combining the two estimates, we obtain the following Empirical Bayes Estimate of θ_n

$$(7) \qquad \tilde{\theta}_n(x) = (x + 1) \frac{N(n, x + 1)}{N(n, x)}$$

The familiar results assert that, as n is increased, $\tilde{\theta}_n(x)$ converges to $\hat{\theta}(x)$ with probability one. Thus, with a considerable number of earlier observations $X_1, X_2, \ldots, X_{n-1}$, the estimate (7) will be almost as accurate as the estimate $\hat{\theta}(x)$ calculable only with complete knowledge of the *a priori* distribution B. The gain in precision, compared to the non-Bayes estimate of θ_n, depends upon the nature of B and may be very considerable.

In order to illustrate the working of the Empirical Bayes Estimate (7) return to the example of two clinics A and B. Assume that the diagnostic test applied to the two patients gave $X = 5$, where X is a Poisson variable with unknown expectation θ. The decision theoretical non-Bayesian estimate (minimum variance, unbiased) of θ for these two patients will be the same $\hat{\theta} = 5$. However, the records of the two clinics may show that among the thousands of cases where the same test was applied to their clientele there were

$$N(5) = 100 \text{ and } N(6) = 200$$

cases with $x = 5$ and $x = 6$, respectively, in clinic A and

$$N(5) = 1000 \text{ and } N(6) = 500$$

similar cases in clinic B. The corresponding Empirical Bayes' Estimates of θ for the two patients will then be

$$\tilde{\theta}(5) = (5 + 1)\frac{200}{100} = 12$$

in clinic A and

$$\tilde{\theta}(5) = (5 + 1)\frac{500}{1000} = 3$$

in clinic B. The essential point regarding this procedure is that, if the amount of previous experience is large, then the Bayes' estimate of θ that could be computed with full knowledge of the *a priori* distribution of this parameter can not be much better than the one computable from Robbins' formula.

The Poisson distribution of the observable random variable is just one of the examples considered by Robbins in his paper of 1955. Similar formulas are deduced for the case where the observable variable has the geometric and the binomial distribution. Also the paper contains indications regarding similar treatment of other distributions. In each case studied Robbins produces one Empirical Bayes' Estimate. This proves their existence. However, Robbins emphasizes that the problem of selecting the estimate that is, in some sense, best is still outstanding. Here, then, is a fruitful new field for theoretical studies.

There are also theoretical problems of a different kind immediately concerned with particular fields of application. For example, in Example 2 concerned with multiple galaxies, there is the question as to what characteristic X of the configuration of galaxies surrounding the given galaxy G should be considered, what is the distribution of X given the multiplicity θ of the system to which G belongs and finally, how to obtain an Empirical Bayes' Estimate of θ. It is likely that the solution of this triple problem will be rather difficult to attain. While this probably will be the case in a number of other applied problems, the initial solutions given by Robbins for the Poisson, the geometric and the binomial distributions, with the hypergeometric being easily added to the list, can be applied directly in many domains and, being very simple, should be included in elementary courses of instruction. Depending upon the nature of the uncontrollable unknown *a priori* distribution of θ, the gain in precision of the estimate, in the sense of mean square error, may be very large indeed.

Before proceeding any further, let us record the exact nature of Robbins' first breakthrough. As is usual in such cases, it consisted in breaking a routine of thought.

In the earlier period, the behavioristic statisticians were preoccupied with solving outstanding problems where the parameter θ is not a random variable. This preoccupation created a routine of ignoring those cases where there is no reluctance of treating θ as a random variable, but with an unknown distribution B. It seemed that no behavioristic problem could be formulated in such cases other than by ignoring the randomness of θ, or by seeking solutions that, in a sense, are independent of B. Robbins' achievement consists in breaking this routine of thought. He isolated a category of situations where the datum includes details ("previous experience") sufficient for an approximate solution of a behavioristic problem that admits an exact solution only if the *a priori* distribution of the parameter θ is completely specified.

THE SECOND ROBBINS' BREAKTHROUGH ON THE BAYES' FRONT

In the preceding pages I emphasized that the first breakthrough accomplished in Robbins' paper of 1955 is applicable to cases, exemplified by my Examples 2 and 3, where the research worker finds it appropriate to consider that the unobservable parameter θ is itself a random variable with a fixed but unknown *a priori* distribution. Examples 1 and 4 were introduced in order to illustrate cases where at least some individuals, myself included, are reluctant to postulate the randomness of θ. In fact, the number of defective pairs of shoes per lot manufactured in San Francisco is likely to have nothing to do with such number in a lot manufactured in New York, and a postulate that both are independent sample values from the same distribution appears to be violently "stretching" the point. Similar considerations apply to the problem of screening chemical compounds for curative effects on cancer.

In each of these two examples we have a number, say N, of disconnected problems, either the problems of testing hypotheses or of estimation. For example, in acceptance sampling of lots of shoes, there may be $N = 1000$ lots each with its own number θ of defectives and what we face is the problem of deciding whether the ith lot should or should not be accepted as conforming with the agreed specifications, for $i = 1, 2, \ldots, N$. Similarly, in the problem of screening chemical compounds, there may be N of them undergoing a preliminary test, each with its own θ characterizing its curative effect on cancer, and for each, say the ith of them, the problem is to decide whether further studies of the ith compound should be continued or abandoned.

The only thing that links the N decision problems in each of these examples is that they are N identical problems of testing hypotheses. Another essential point of the situation is that these N identical problems may be treated simultaneously. For example, the sampling of lots of shoes may be going on in various cities for a month or so, but it may be agreed that the decisions about particular lots will be taken some time later, simultaneously. The same possibility occurs with the problem of screening chemical compounds where the decision of abandoning some compounds and continuing the study of some others may be taken at the conclusion of a phase of study. A similar possibility occurs also in the problem of identifying physical systems of galaxies, although in this instance it appears natural to contemplate a chance mechanism determining the multiplicity of systems. On the other hand, no possibility of a simultaneous decision regarding patients of a clinic appears possible. The patients arrive when they feel in need of medical help, and each hopes to obtain immediate service based on all the information available at the particular moment. Thus, the ith patient must be treated using the experience provided by the earlier i–1 patients.

The above represents my own point of view on the particular problems contemplated. Where to draw the line between cases where θ is to be treated as a random variable and those where such an assumption would be extraneous, is a subjective matter. The interesting question of behavioristic statistical theory is what can be done to control errors in those cases illustrated by Examples 1 and 4 where there are N identical problems and the research worker decides against considering θ as a random variable? To all appearances, we are in the circumstances of now-classical decision problems, and the best that one can do appears to be: Treat each of the N separate problems as best we can, perhaps using most powerful tests, if such are available.

This appears to have been the situation before the publication of Robbins' paper of 1950 in which he made a breakthrough that I am describing as his "second". Historically, of course, this "second" breakthrough, of 1950, occurred earlier than the one of 1955, I describe as the first. My reason for this interchange in ordering is that the breakthrough of 1955 is easier to follow intuitively than that of 1950 and that, besides, some of the solutions attained for particular problems are, more or less, ready for practical applications. As of now, the situation with the breakthrough of 1950, which I call the second, appears to be less ripe.

The essence of Robbins' second breakthrough (of 1950) is the information that, if a statistician deals simultaneously with a large number N of identical problems of testing hypotheses and wants to diminish the overall expected frequency of errors of both kinds, then, at least in certain circumstances, he can bring this frequency below the level attainable through N independent applications of the most powerful test. The resulting gain may be very impressive.

Thus far, the possibility of such solutions has been proven for the case of testing a simple hypothesis against a simple alternative, and this limits severely the field of applications. For example, it is not to be expected that among the N lots of shoes the number of defective pairs can have only two known values, say either 5 or 50. Similarly, the curative effect of a drug may be characterized by a number of values of θ. In addition, in a number of problems there are likely to be nuisance parameters. However, these difficulties do not detract from the importance of Robbins' discovery; they add to the interest in the new field of study which, before Robbins' paper, no one dreamed of.

Here is a sketch of the example treated by Robbins in his original paper. The observable variable X is assumed to be known to have a normal distribution with known variance, say $\sigma^2 = 1$, and with an unknown mean θ. However, it is assumed that θ can have one of only two values, either $\theta = -1$ or $\theta = +1$. We have N unrelated problems to be considered simultaneously, and for the ith problem there is the unknown $\theta_i =$ either -1 or $+1$ and the corresponding $X_i = N(\theta_i, 1)$. The statistician considers that the two errors possible regarding the ith θ are of equal importance and seeks ways and means to make the frequencies of these errors equal and to reduce their common value to as low a level as practicable.

When the two kinds of error are of the same importance, the most powerful test of the hypothesis that $\theta_i = -1$ against the alternative $\theta_i = +1$ consists in the rule of adopting $\theta_i = -1$ when $X_i \leqslant 0$ and of adopting $\theta_i = +1$ when $X_i > 0$. The probability of being wrong in either case is simply

$$P\{0 < X \mid \theta = -1\} = P\{X < 0 \mid \theta = +1\} = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-(x+1)^2/2}\, dx = 0.1587. \tag{8}$$

Thus, if the statistician decides to use the most powerful test independently N times, the expected number of erroneous decisions will be $0.1587 \times N$ and, if N is large, the actual proportion of errors will be close to 16 percent.

Now consider an alternative procedure invented by Robbins which, to some minds, may easily appear revolting. Denote by $\bar{X}$ the arithmetic mean of $X_1, X_2, \ldots, X_N$ observable in the N unrelated problems. Robbins' procedure is:

$$\text{If } \bar{X} \leqslant -1, \text{ assert } \theta_i = -1 \text{ for all } i = 1, 2, \ldots, N.$$

$$\text{If } -1 < \bar{X} < +1 \text{ and } X_i \leqslant \frac{1}{2} \log \frac{1 - \bar{X}}{1 + \bar{X}}, \text{ assert } \theta_i = -1\,.$$

$$\text{If } -1 < \bar{X} < +1 \text{ and } X_i > \frac{1}{2} \log \frac{1 - \bar{X}}{1 + \bar{X}}, \text{ assert } \theta_i = +1.$$

$$\text{If } +1 \leqslant \bar{X}, \text{ assert } \theta_i = +1 \text{ for all } i = 1, 2, \ldots, N.$$

The "revolting" aspect of this rule is that, for example, the decision regarding a lot of shoes delivered by a given manufacturer would depend not only on the number of defectives in the sample drawn from this lot, but also, to a considerable extent, on the number of defects found in samples from lots delivered by the competitors.

The frequency of errors in following Robbins' rule is represented by a fairly complicated formula depending upon N and also upon the proportion $p(\theta)$ of those cases where $\theta = +1$. The asymptotic values of the expected frequency of errors, reasonably approached when $N = 100$, is as in Table 1.

TABLE I

ASYMPTOTIC EXPECTED FREQUENCIES OF ERROR IN USING ROBBINS' TEST

Proportion of $\theta = 1$		0 or 1	0.1 or 0.9	0.2 or 0.8	0.3 or 0.7	0.4 or 0.6	0.5
Expected error	Robbins	0.00	0.07	0.11	0.14	0.15	0.16
frequency	M.P.T.	0.16	0.16	0.16	0.16	0.16	0.16

The figures in the first line are asymptotic values, as $N \to \infty$, of the expected frequencies of errors in using Robbins' test. For finite N they are slightly larger. Thus, with $N = 100$ and $p(\theta) = 0.5$, the Robbins' frequency of error is 0.1628 as against 0.1587 attainable by the most powerful test. It will be seen that Robbins' procedure is the more effective the farther is the frequency $p(\theta)$ from one-half. Here the expected frequency of errors may be reduced by a factor of 2 or more.

The decision problem representing simultaneous treatment of N identical but unrelated decision problems, as just described, has been named by Robbins the compound statistical decision problem. Over the decade since the publication of the Robbins' initial paper of 1950, there appeared a few other publications on the same or on an allied subject. Robbins' first paper is specifically concerned with testing simple statistical hypotheses against a simple alternative. His own and his students' subsequent papers [16], [17], [18] are similarly oriented. However, it was found possible to drop the requirement of simultaneity of N decisions and to substitute for the original solution an appropriate sequential procedure. In 1955 Stein studied the compound problem of point estimation of the mean of a normal distribution [19]. Somewhat to his admitted surprise he found that, with $N > 2$, the usual estimator is not the best. Apparently, the discussion with Tukey and Robbins helped to break down Stein's originally contrary conviction.

A paper [20] of a substantially greater generality was presented by Blackwell at the International Congress of Mathematicians held in Amsterdam in 1954. It is directed somewhat "diagonally" between the Empirical Bayes problem and the compound problem as described above. Briefly, Blackwell considers a sequence of decision problems, somewhat like in the Empirical Bayes' problem. However, he does not assume the existence of either known or unknown *a priori* distribution of the parameter θ. Instead, it is assumed that, after the decision regarding θ_n the true value of θ_n becomes known. (This corresponds to the situation in a clinic of my example 3 where it may be known for a number of earlier patients whether they were suffering from the cancer of the lung or not.) Here again the compounding of the many identical problems proves a great advantage.

Short as the above list of novel results is, it indicates that in 1950–1955 we entered into phase (ii) of the new developmental cycle on the Bayes' front. A new broad domain has been opened for study, very interesting theoretically and promising considerable gains in the applied fields. However, before these gains are achieved, a period of controversy is to be anticipated. Here the representatives of applied domains may be very helpful. For examples, in spite of foreseeable opposition of the manufacturers, the large consumer groups are likely to welcome the information about the possibility of reducing sampling inspection errors by compounding the many separate identical problems. Similary, the scholarly bodies concerned with screening chemical compounds for curative effects on cancer, may be expected to be interested in methods insuring a decreased over-all frequency of errors.

What is the essence of Robbins' second breakthrough? As I see it, the basic idea is best understood heuristically with reference to the non-sequential case where θ can have only values of a finite set, say $\vartheta_1, \vartheta_2, \ldots, \vartheta_s$. If the statistician is interested in reducing to a minimum the overall frequency of errors in N disconnected problems, then the order in which the particular values ϑ occur in the sequence $\theta_1, \theta_2, \ldots, \theta_n$ is immaterial. All that is relevant is how many of the $\theta_1, \theta_2, \ldots, \theta_N$ are equal to ϑ_1, how many are equal to ϑ_2, etc. In these circumstances, the N values of the parameter θ, namely $\theta_1, \theta_2, \ldots, \theta_N$, play a role very similar to that of a sample of N independent observations on a single random variable with a fixed distribution. Thus, the Compound Decision problem appears akin to an Empirical Bayes' problem, in which the possibility of utilizing the accumulated experience is quite intuitive.

While the above heuristic description appears plausible, the mechanism of Robbins' discovery is likely to have been different. In fact, the reduction of the Compound Decision problem to an Empirical Bayes' problem suggests that, if both problems are formulated and solved by the same individual, then the solution of the Empirical Bayes' problem would have been reached first. However, Robbins' intuition brought him to the Compound Decision problem in 1950 and to the Empirical Bayes' problem five years later.

TWO ALTERNATIVE APPROACHES TO THE BAYESIAN PROBLEM

The papers by Allan Birnbaum and Jerome Cornfield exemplify two different "inferential" approaches to the Bayesian problem, sharply different from the approach by Robbins.

Birnbaum's paper [1] is "On the foundations of statistical inference" and is concerned with "two main general problems of informative inference: the problem of finding an appropriate *mathematical characterization* of statistical evidence as such; and the problem of *evidential interpretation*, that is, of determining concepts and terms appropriate to describe and interpret the essential properties of statistical evidence." The terminology used in treating the two problems is rich and includes "statistical inference", "informative inference", "experimental evidence", "statistical evidence" and "evidential meaning". The symbol to denote the latter term is $Ev(E, x)$. In addition, there are terms "intrinsic significance level", "intrinsic confidence set" and "intrinsic confidence coefficient". The method of treating the two problems consists in introducing three principles, of which the first two imply the third. The first principle is

> "*The principle of sufficiency* (S): If E is specified experiment, with outcomes x; if $t = t(x)$ is any sufficient statistic; and if E' is the experiment, derived from E, in which any outcome x of E is represented only by the corresponding value $t = t(x)$ of the sufficient statistic; then for each x, $Ev(E, x) = Ev(E', t)$ where $t = t(x)$ ... (S) may be described informally as asserting the 'irrelevance of observations independent of a sufficient statistic'."

The principle S, and also other principles of Birnbaum, have a normative character. Birnbaum considers them "compellingly appropriate" and the 38 page article is concerned with the relationships, discussed on examples, between the principles and the various current statistical techniques.

The principles of Birnbaum appear as a kind of substitutes for known theorems. For example, various authors proved theorems to the general effect that the use of sufficient statistics will minimize the frequency of errors. An elegant theorem of this kind, proved by C. R. Rao and, independently by David Blackwell, asserts that the minimum variance unbiased estimate of a parameter must be a function of the sufficient statistic if such exists. Birnbaum does not mention any such theorems but, instead, advances his principle S. It would seem that, in cases where theorems of this kind are applicable, the principle S is superfluous. If it is introduced at all, as a compelling principle, cases must be visualized where it would be applied in spite of the absence of proved propositions indicating the advantages of such applications. If a case arises where a method, inconsistent with Birnbaum's principles, is proved to have properties that someone considers advantageous, Birnbaum is likely to disapprove.

An example that occurs to me is that of Joseph Berkson's finding [21] that, if one drops the requirement of unbiasedness and insists only on the smallness of the mean square error, estimators may be found that are better than those computable from the existing sufficient statistic. This is an important information opening a new field of study. Yet, if Birnbaum's principle S were adopted, this field of research would have been dogmatically closed.

There is another aspect of the situation that may be useful to mention. Birnbaum's principles may be regarded as an attempt to summarize the different ways in which the statisticians approach their problems. In particular, the principle S may be thought of as providing a shortcut in various problems in which a sufficient statistic exists. The example of Berkson's estimators shows that the current research in statistics extends beyond the area encompased by Birnbaum's principles.

Cornfield's paper [2] also advances a principle, the principle of initial impartiality. However, the general character of this principle and its use are very different from those of Birnbaum. Cornfield begins by emphasizing that he is concerned with inferences in those cases where, for one reason or another, the unknown parameter is treated as a constant, not as a random variable. The system of inferences developed is described as an "objective" Bayesian calculus. It consists in computing Bayes' formula (1) with an especially devised, or deduced, "initial function" as a substitute of the *a priori* distribution. It is this "initial function" that is provided by the principle of initial impartiality.

Apart from terminology (for example "initial function" rather than "prior probability", etc.) Cornfield's new "measure of uncertainty" is reminiscent of Jeffreys' posterior probability. The practical effects of computing either are certainly the same. If for a given hypothesis or proposition the computed quantity is small, then the truth of the hypothesis or proposition is subject to doubt. Cornfield notes cases where his measure of uncertainty coincides with Jeffreys' posterior probability. However, there are cases of disagreement and the question may arise which of the two sets of principles should be given credence. Even more interesting is the question about a general framework of concepts within which the problem of difference between the two methodologies can be resolved. My own attitude would be to inquire which of the two methods will lead to fewer errors, but considerations of this kind seem to have been excluded. In fact, if considerations of errors and their frequencies were admissible, then there would have been no need for the principle of initial impartiality or for any other principle. Cornfield could simply use methods developed in the nineteanthirties that guarantee a low level of frequency of errors, occasionally the lowest possible level. As things are, the possibility of resolving the contradictions between Cornfield and Jeffreys, and also similar contradictions between other proponents of the use of Bayes' formula calculated with a substitute of B not implied by the conditions of the problem appears to depend upon the relative persuasiveness of the authors concerned.

As I see it, the papers by Birnbaum and by Cornfield belong to phase (i) of a developmental cycle, not necessarily the same as the cycle marked by the paper of Robbins and Samuel.

SOMMAIRE

Dans ses deux papiers, l'un publié en 1951 et l'autre en 1956, M. Herbert Robbins a formulé deux catégories nouvelles de problèmes statistiques. Aussi il a donné des solutions de ces problèmes dans certains cas particuliers. Il semble que les deux catégories de problèmes auront une grande importance pour les applications.

La première catégorie concerne le fameux problème de Bayes où les conditions impliquent qu'un paramètre θ à estimer est une variable aléatoire dont on ignore la distribution. M. Robbins indique une méthode d'utilisation des observations antérieures pour estimer l'estimateur de θ qu'on pourrait calculer si on connaissait la distribution *a priori* de θ.

La deuxième catégorie de problèmes concerne le cas où on a à tester un grand nombre d'hypothèses statistiques $H_1, H_2, \ldots, H_h$ du même type. M. Robbins montre que, s'il est permis de faire le test d'une hypothèse H_i dépendre des observations relatives à toutes les hypothèses $H_1, H_2, \ldots, H_n$, même si ces hypothèses n'ont pas de relations logiques, la fréquence totale des erreurs peut être diminuée.

REFERENCES

[1] Birnbaum, A. On the foundations of statistical inference. To be published in the Journal of the American Statistical Association.

[2] Cornfield, J. An objective Bayesian calculus. Paper prepared for presentation at the December, 1961, meeting of the Institute of Mathematical Statistics.

[3] Robbins, H. and Samuel, E. Testing of statistical hypotheses: the "compound" approach. Paper prepared for presentation at the December, 1961, meeting of the Institute of Mathematical Statistics.

[4] Mises, R. von. Fundamentalsätze der Wahrscheinlichkeitsrechnung. Mathematische Zeitschrift, 4 1919, p. 1–97.

[5] Wald, A. Contributions to the theory of statistical estimation and testing hypotheses. Annals of Mathematical Statistics, 10 1939, p. 299–326.

[6] Wald, A. Statistical Decision Functions. New York, Wiley, 1950.

[7] Pitman, E. J. G. Statistics and science. Journal of the American Statistical Association, 52 1957, p. 322–330.

[8] Stein, C. M. A two-sample test for a linear hypothesis whose power is independent of the variance. Annals of Mathematical Statistics, 16 1945, p. 243–258.

[9] Tukey, J. W. Quick and dirty methods in statistics. Part II: Simple analyses for standard designs. Proceedings of the Fifth Annual Conference of the American Society for Quality Control, 1951, p. 189–197.

[10] Scheffé, H. A method for judging all contrasts in the analysis of variance. Biometrika, 40 1953, p. 87–104.

[11] Gauss, C. F. Abhandlungen zur Methode der kleinsten Quadrate. *German translation from Latin by A. Borsch and P. Simon, Berlin, 1887.*

[12] Wilson, E. B. Probable inference, the law of succession, and statistical inference, Journal of the American Statistical Association, 22 1927, p. 209–212.

[13] Hotelling, H. The generalization of Student's ratio. Annals of Mathematical Statistics, 2 1931, p. 360–378.

[14] Robbins, H. An empirical Bayes' approach to statistics. Proceedings of the Third Berkeley Symposium on Statistics and Probability, 1 1955, p. 157–164. Berkeley and Los Angeles, University of California Press, 1956.

[15] Mises, R. von. On the correct use of Bayes' formula. Annals of Mathematical Statistics, 13 1942, p. 156–165.

[16] Robbins, H. Asymptotically subminimax solutions of compound statistical decision problems. Proceedings of the Second Berkeley Symposium on Statistics and Probability, 1950, p. 131–148. Berkeley and Los Angeles, University of California Press, 1951.

[17] Hannan, J. F. and Robbins, H. Asymptotic solutions of the compound decision problem for two completely specified distributions. Annals of Mathematical Statistics, 26 1955, p. 37–51.

[18] Samuel, E. On the compound decision problem in the non-sequential and the sequential case. Columbia University Dissertation, 1961.

[19] Stein, C. M. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proceedings of the Third Berkeley Symposium on Statistics and Probability, 1 1955, p. 197–206. Berkeley and Los Angeles, University of California Press, 1956.

[20] Blackwell, D. Controlled random walks. Proceedings of the International Congress of Mathematicians, 3 1954, p. 336–338. Amsterdam, North Holland Publishing Company, 1956.

[21] Berkson, J. Estimation by least squares and maximum likelihood. Proceedings of the Third Berkeley Symposium on Statistics and Probability, 1 1955, p. 1–11, Berkeley and Los Angeles, University of California Press, 1956.