DISCUSSION OF PAPERS READ AT THE MAY 22, 1950 MEETING

CREDIBILITY PROCEDURES—LA PLACE'S GENERALIZATION OF BAYES' RULE AND THE COMBINATION OF COLLATERAL KNOWLEDGE WITH OBSERVED DATA

ARTHUR L. BAILEY

Volume XXXVII, Part I

WRITTEN DISCUSSION BY *DR. RICHARD VON MISES, Harvard University

Comments on Statistical Theory of Inference.

1. The basis of any statistical (probability) theory of inference is supplied by the concept, due to Bayes, of a priori and a posteriori probabilities. These may perhaps better be called "over-all" and "inferred" probabilities.

2. To be sure, the generalization, ascribed to Laplace, according to which the over-all probability is not necessarily a constant, but an actual function of the variable parameter has to be incorporated in the Bayes theory.

3. Under certain conditions (which are fulfilled in most practical cases) a decisive supplementation of the Bayes theory is supplied by the following statement: The influence of the over-all probability upon the inferred probability decreases the more the number of trials on which the statistics is based increases; in the limit, for an infinite number of trials, the inferred probability becomes independent of the over-all probability.

4. The a priori or over-all probability should not be confounded with concepts like "credibility" or "degree of confirmation," or "strength of expectation," etc. Whenever it is assigned a numerical value and enters as such a computational formula, it is a frequency limit like any other probability.

5. The fact that in many cases the over-all probability is not exactly known does not preclude the application of Bayes' concepts. One has to introduce reasonable estimates for it and to study the extent (depending on the number of trials), to which the indeterminacy affects the results.

6. The various currently used inference methods must be judged according to their compatibility with Bayes' concepts. Some results are the following:

a) The method of confidence limits (or fiducial limits) is in agreement with Bayes' theory, but it does not answer the question what inference can be drawn from a definite observation.

b) The Nyman-Pearson method of testing hypotheses, if interpreted in the correct way, leads to a weak and in most cases insufficient answer.

c) The likelihood method solves the problem only if the over-all probability is supposed to be constant or some metaphysical principle of "insufficient reason" is applied.

d) The recently developed decision functions of A. Wald are in full agree-

*By invitation.

ment with Bayes' concepts; they solve a more elaborate problem connected with the original inference problem.

7. All so-called small sample theories which derive estimates, decisions, etc., from a small number of observations without taking into account the over-all probability are completely unfounded and unreliable.

8. Many more detailed investigations of the consequences of Bayes' theory would be in order, for instance, that initiated by Mr. Bailey, or the development of approximation formulas according to point 5, or the extension of the range of validity of the statement in point 3, etc.

9. It is to be hoped that those and similar problems will find the attention of competent statisticians, as the unjustified and unreasonable attacks on the Bayes theory, initiated by R. A. Fisher, will fade out.

References

To 3): The statement is first proved in my paper: *Mathematische Zeitschrift* 5 (1919), p. 83, later in my textbook *Wahrscheinlichkeitsrechnung* (now: reprint Rosenberg, New York 1945), p. 185. See also *Lecture notes on probability and statistics*, Harvard University 1946, Chapter VI, 3.

To 5): An example is given in my paper: On the correct use of Bayes' formula, *Annals of Math. Statistics* 13 (1942), pp. 156-165.

To 6a, b, c): See the *Notes* mentioned above, Chapter IX.

To 6a: See also my paper: On the foundations of probability and statistics, *Annals of Math. Statistics* 12 (1941), p. 200.

To 6b): See also my paper: On the problem of testing hypotheses, *Annals of Math. Statistics* 14 (1943), pp. 238-252.

WRITTEN DISCUSSION

BY

*E. C. MOLINA

The privilege of discussing Mr. Bailey's paper is indeed gratifying to one who is not a member of the Casualty Actuarial Society. Moreover, the paper under consideration is of particular interest to one who, for nearly four decades, has been applying inverse probability formulas to problems confronting another great industry.

Inverse, or a posteriori, probability is that branch of probability theory which enables one to draw conclusions regarding the antecedents or causes of observed events. Quoting from the first paragraph of Mr. Bailey's paper, one has recourse to inverse probability theory "to determine the weight to be given to the indications of actual observations in a combination of such indications with a priori expectations which were based on other actual data, on prior knowledge or on reasonable assumptions made before observations were available."

In the literature of probability theory great confusion exists because many authorities have failed to distinguish clearly between the original Bayes inverse theorem and its subsequent generalization by Laplace. The generalized

*By invitation.

theorem embraces, or brings together, both the data obtained from a series of observations and whatever "collateral" information exists in relation to the observed results. We are greatly indebted to Mr. Bailey for the emphasis he has placed on the Laplacian generalization. Its appearance as a sub-title to his "Credibility Procedures" gives one *ab initio* the kernal of Mr. Bailey's analysis.

One whose acceptance of Laplace's generalization for the solution of inverse probability problems is based on an extensive and intensive study of the classics beginning with Bayes' famous Essay and running through the works of Laplace, Poisson, Cournot, Bertrand, Poincaré, Czuber, Borel, Castelnuovo *et al*, finds it difficult to commend in restrained terms the paper on "Credibility Procedures" submitted by Mr. Bailey for your consideration.

WRITTEN DISCUSSION
ON THE UTILIZATION OF DIRECT AS WELL AS COLLATERAL INFORMATION IN THE PROBLEM OF STATISTICAL ESTIMATION

*JOHN E. FREUND
Associate Professor of Mathematics
Alfred University

> "To make a careful estimation means to utilize all relevant knowledge available and to reason well in deriving the estimate from this knowledge."
>
> R. Carnap[1]

Attempts to formulate a general theory of statistical estimation date back as far as the eighteenth century. It was only recently, however, that a method was developed, the method of Maximum Likelihood, which, although not general in the strictest sense, takes care of a relatively large class of problems of estimation. A good number of statisticians seem reluctant, however, to accept this method in general, questioning its appropriateness in specific applications, as well as doubting the soundness of its arbitrary choice of criteria.

It seems questionable to us whether it is at all possible to formulate satisfactory universal principles which define a "best" estimate, not necessarily the same in each case, for every problem of estimation and for every kind of direct or indirect evidence. Indeed, we are doubtful whether it is actually wise to follow the above quotation and consider *all* available information under *all* circumstances.

This does not mean that we are questioning the usefulness and importance of the most recent developments in the generalization of statistical theory, which are due mainly to John von Neumann and Abraham Wald.[2]

In the first part of this paper we shall discuss very briefly our own approach to the subject of credibility. Because of the magnitude of the problem, it is understandably impossible, to present to you today anything but a brief ab-

[1] R. Carnap, *Logical Foundations of Probability*, Univ. of Chicago, 1950.
[2] J. von Neumann, *Theory of Games and Economic Behaviour*, Princeton Univ., 1944 and A. Wald, *Statistical Decision Functions*, New York, 1950.
* By invitation.

stract of these views. The second part of this paper deals with two very short comments on Mr. Bailey's paper, "Credibility Procedures," which was presented to you this spring at the Stockbridge meeting of the Casualty Actuarial Society.

I.

For reasons which will be explained later we shall treat the problem of credibility as a problem of multiple estimation. By multiple estimation we mean the problem of estimating the population parameters of a set of populations which have been chosen as a group because of some property or properties which they may have in common. Such a set of populations might, for example, consist of the various risks which belong to a given classification.

Let us denote the distribution which is associated with each of these populations (risks) by the symbol $f(x_{ij} \mid \theta_i)$. This distribution represents the conditional probability (probability density) of obtaining an observation x_{ij} from the ith population of our class, if the parameter θ_i is a certain fixed constant. The symbol x_{ij} stands for the jth observation taken from the ith population. This symbolism is convenient, if we have more than one observation from each population (risk).

Our problem is to estimate the population parameters θ_i, which, for each population must, of course, be a fixed constant, but which need not be the same for the various elements of our chosen class of populations. Consequently, if we consider the entire class of populations (risks), we can now speak of the distribution $f(\theta)$ of the parameters θ_i *within the chosen class of populations.* Whenever we treat the parameters θ_i as variables, *in this sense,* the subscript i will be omitted.

Given the distribution $f(x_{ij} \mid \theta_i)$ and the distribution $f(\theta)$, we can readily calculate the distribution $f(\theta \mid x_{ij})$ by means of the rule of Bayes-Laplace. The new distribution function $f(\theta \mid x_{ij})$ expresses the probability that a given observation x_{ij} has come from a population whose parameter equals the constant θ. In order to complete the symbolism which we shall use, let O_{n_i} stand for a random sample of n_i observations taken from the ith population and let θ_i' stand for an estimate of θ_i.

In the examples which we shall discuss, it will always be assumed that the direct information consists of random samples O_{n_i} from at least one of the populations. As a matter of convenience (it is by no means necessary), we shall also assume that the n_i are all equal to a constant n. The indirect, collateral, or antecedent information which may be available in our examples will consist of either complete or partial information concerning the distributions $f(\theta)$ and $f(x_{ij} \mid \theta_i)$. Estimates which are based on partial or complete knowledge of the distribution $f(\theta)$ will be called *Credibility Estimates.*

Before we can estimate the parameters θ_i, we must first establish a criterion which defines what we mean by a "good", "best", or "preferred" estimate. This is essential because we can estimate the θ_i in infinitely many ways. As a matter of fact, the method of estimation is completely arbitrary unless we specify some sort of criterion, on the basis of which we can distinguish between the various kinds of estimates with reference to some desirable properties.

This situation is quite similar to the customary problem of fitting a straight line through a given set of points. We can draw, of course, infinitely many

of these lines, and unless we define what we mean by a "good fit", we have no basis for expressing a preference for any one of these lines. Therefore, also in our original problem, we must establish such a criterion before we can estimate the θ_i. It is important to note that a chosen criterion must be such that whatever information is available or can be obtained will be sufficient to perform the method of estimation which has thus been defined. Furthermore, we shall, in general, base our criteria on pragmatic considerations, such as minimizing certain quantities relating to errors or maximizing certain probabilities.

We shall now proceed to discuss a few of the credibility estimates which may be obtained under several conditions regarding the collateral information and under correspondingly different criteria which will define our "preferred" estimates. It must be understood, of course, that by knowledge, collateral or otherwise, we mean empirical and not a priori knowledge. As we shall show later on, we are *not* justified in using the rule of Bayes-Laplace, unless we have an empirical basis for the type of distribution which is to be used for $f(\theta)$.

Case A. *We have complete knowledge of both* $f(\theta)$ *and* $f(x_{ij} \mid \theta_i)$. In this case, where we have the maximum amount of collateral information, short of actually knowing the θ_i, we might suggest two alternative criteria which define our "preferred" estimates. *Criterion 1*:

> "The estimates should be such that if we were to apply this method of estimation to all members of our class of populations (to all risks within the given classification), the direct information being identical in each case, then the error variance $\Sigma(\theta_i' - \theta_i)^2$ should be a minimum."

It can easily be shown that the estimate which is thus defined is simply the mean of the distribution $f(\theta \mid O_{n_i})$ and we shall consequently call this type of estimate a *Mean Estimate.*[3] Therefore in this case

$$\theta_i' = \int \theta . f(\theta \mid O_{n_i}) d\theta$$

and the actual form of the estimate will, of course, depend on the distributions which are being used.

An alternative solution may be obtained by means of *Criterion 2*:

> "The estimate θ_i' should be the value of the parameter θ of the population (within our chosen class of populations) from which the given sample is most likely to have come."

It is important to note that this estimate is *not* a Maximum Likelihood estimate. The estimate which has thus been defined is simply the mode of the distribution $f(\theta \mid O_{n_i})$ and we shall therefore call it a *Modal Estimate.* Consequently in this case

$$\theta_i' = \text{the mode of } f(\theta \mid O_{n_i})$$

and the form of the estimate will again depend on the nature of the distributions which are used in the computation of $f(\theta \mid O_{ni})$. It is an interesting fact that if both of the original distributions are normal distributions, the estimates resulting from the two different criteria will be identically the same.

[3] The term "Mean Estimate" was suggested to us by Prof. H. Reichenbach of the University of California at Los Angeles.

Case B. *We know the distribution* $f(x_{ij} \mid \theta_i)$, *but we have only partial information concerning* $f(\theta)$. *For example, we might know its first two moments, namely* $\overline{\theta}$ and σ_θ^2. The criterion which we shall employ in this case is an adaptation of the "Theory of Adjustment",[4] originally developed for problems in surveying, where we estimate parameters like the sides and angles of a triangle, which, as we know, must satisfy certain trigonometric identities. The criterion, in this case, is the following, *Criterion 3*:

> "The set of estimates θ_i' should be the values of the θ_i which maximize the probability of obtaining the given sample of n observations from each population within our classification, under the condition that the θ_i' must satisfy a given number of functional restrictions."

This means that we must maximize the probability P, where P is given by

$$f(O_{n_1} \mid \theta_1) \cdot f(O_{n_2} \mid \theta_2) \ldots\ldots\ldots\ldots f(O_{n_k} \mid \theta_k)$$

under the condition that, for example, $\Sigma\theta_i' = c$ and $\Sigma\theta_i'^2 = d$, where c and d are known constants. As a suitable name for estimates of this type we would like to suggest *Restricted Maximum Likelihood Estimates*. It is important to note that the objections which Mr. Bailey raised against the method of Maximum Likelihood (on page 6 of his paper) do not apply in this case. The resulting estimates θ_i', which can easily be obtained using the technique of Lagrange Multipliers, will be weighted estimates, very much like those which are obtained with the use of the other criteria.

An alternative approach, in this case, might be to disregard some of the collateral information concerning $f(x_{ij} \mid \theta_i)$ and treat the problem as if it belonged to the case which we shall discuss next.

Case C. *The only knowledge which we have about* $f(x_{ij} \mid \theta_i)$ *is, in this case, its standard deviation* σ *which is assumed to be the same for all populations of our class.*[5] *The only knowledge which we have concerning* $f(\theta)$ *consists of its mean and standard deviation,* $\overline{\theta}$ *and* σ_θ *respectively.* This leads us to Mr. Bailey's "Best Linear Regression". The resulting estimates, which we shall call *Best Linear Estimates*, are defined by the following criterion, *Criterion 4*:

> "The estimates should be such that if we were to apply this method of estimation to all members of our class of populations (risks) for all possible samples from these populations (randomization), the error variance $\Sigma(\theta_i' - \theta_i)^2$ should be a minimum, under the condition that the θ_i' be of the form $A\bar{x}_i + B$, where $\bar{x}_i$ is the sample mean of the ith population."

The properties of this type of estimate are well known, having been developed in detail by Mr. Bailey in one of his earlier papers.[6] It is important to note the distinction between Criterion 4 and Criterion 1. We are now summing on the x_{ij} as well as on θ, whereas we kept the direct information, i.e. the x_{ij}, constant in our formulation of Criterion 1.

Case D. *The only collateral information which we have in this case consists of the mean and standard deviation of* $f(\theta)$, *i.e.* $\overline{\theta}$ *and* σ_θ. A possible estimate

[4] See N. Arley and K. R. Buch, *Introduction to the Theory of Probability and Statistics*, New York, 1950, esp. chapter XII.

[5] This assumption is modified, for example, in the multiplicative case, where the σ's are proportional to the θ's.

[6] See A. Bailey, "A Generalized Theory of Creditibility," *Proceedings* of the Casualty Actuarial Society, Vol. XXXII, 1945.

which suggests itself in this case is, what might reasonably be called a *Restricted Least Square Estimate.* Its properties are defined by *Criterion 5*:

"The estimates θ_i' should be such that they minimize the expression

$$\sum_{i=1}^{k} \sum_{j=1}^{n} (x_{ij} - \theta_i')^2$$

under the restriction that $\Sigma\theta_i'/k = \bar{\theta}$ and $\Sigma\theta_i'^2/k = \sigma_\theta^2 + \bar{\theta}^2$ where k is the number of populations belonging to our chosen classification."[7]

Applying this criterion, we obtain (again with the aid of Lagrange Multipliers) the final result that

$$\theta_i' = C \cdot \bar{x}_i + (1 - C) \cdot \bar{x} + (\bar{\theta} - \bar{x})$$

where $\bar{x}$ is the over-all mean of the sample values of all the populations, and where the constant C is given as

$$C = \sigma_\theta / \sigma_{\bar{x}_i}$$

The criterion used in this case demands that certain conditions which are satisfied by the θ's must also be satisfied by the θ''s. In other words, we have transferred certain properties of the population parameters to their estimates.

It must be evident, that the analysis which we have given in the above discussion is far from being an exhaustive study of the subject of credibility. Indeed, it has been our purpose rather to indicate by means of a few special cases the approach which can be used in obtaining credibility formulae, i.e. the formulae for credibility estimates. The steps to be taken can be summarized as follows:

1. We must specify precisely the nature of the collateral information which might be available or which might be obtained.
2. We must then formulate a principle which defines the preferred properties which we want our estimates to have. It is important that these conditions must be such that they can be satisfied by whatever collateral information is available, and they must also be such that they can be translated into mathematical terms.
3. The final step consists of computing the actual formulas, on the basis of the given criterion, using the collateral information which was specified in step 1. This last step may involve a good deal of mathematical detail, but once the criteria have been established in step 2, the problem is, logically speaking, straightforward.

We have denoted the estimates, which we have developed, as "preferred estimates", rather than as "good" or "best" estimates, because this term seems to be more descriptive of the actual situation. An estimate can be "best" in a variety of different ways, depending on whatever we happen to mean by the word "best". (We could, for instance, call an estimate "best" if its formula looks the "prettiest".) The term "preferred" estimate brings out very clearly

[7] It is necessary, in this example, to have a sample from *each* of the populations. The symbol $\sigma_{\bar{x}_i}$ stands for the standard deviation of the sample means, as computed from the given data.

that the estimate is based on conditions which express a preference which may be based on pragmatic or other considerations.

II.

It seems to us that the basis which Mr. Bailey chose in his development of the theory of credibility in his paper, "Credibility Procedures", is equivalent to what we have described very briefly in Case A, Criterion 1.

In spite of the fact that this equivalence may not be immediately apparent, we feel that the meaning of Mr. Bailey's distribution $K(x)$ must be interpreted in the sense of our $f(\theta)$.

The quantity x, the "true" expected losses of a particular risk (or whatever parameter we are trying to estimate) must be a constant as long as we are speaking about a specific risk. As a matter of fact, it must be *defined* as a constant which, incidentally, belongs to what is called the mathematical model. We, therefore, cannot speak about the probability or certainty of its existence. In his general discussion, Mr. Bailey makes the statement:

> ". . . The actuary knows that there is more than one possible value of x and is willing to assume that he can approximate the a priori probabilities of the existence of such possible values."

We cannot believe that Mr. Bailey means to imply that x can be anything but a constant. If, however, we speak about more than one risk, then the corresponding values of x may, of course, be different, and in this sense we can say that there is more than one possible value of x.[8] Therefore, in order to treat x as a variable, we must embed a given risk within a class of other risks. Indeed, we cannot speak about the probability of obtaining a certain value of x, unless we specify such a class of similar risks, which in probability theory is commonly referred to as the reference class.

The embedding of an event within a class of similar events for the purpose of making predictions or estimations is a common procedure in scientific methodology. Let us suppose, for example, that we wish to predict whether it will rain tomorrow or not. The meteorologist, whom we consult, tells us that the probability that it will rain tomorrow is .65. As it is quite evident that one or the other has to happen, we must interpret his statement as saying that: "In a large class of similar situations, we can expect it to rain about 65 percent of the time." In order to make a meaningful prediction, we had to embed the given situation within a large class of similar situations. This, incidentally, is precisely what is being done when a risk is given a manual rate at the time when it is first insured and when no direct information is available.

In establishing the criterion for his "best" estimate (in the sense of Least Squares), Mr. Bailey says that the error variance is summed

> ". . . for all of the possible cases for which H may occur."

To speak of "all possible cases" is meaningful only if we specify a definite reference class. It seems to us, therefore, that Mr. Bailey's criterion is identical with our criterion 1 in which we also summed the error variance over the entire class within which we have classified the risk.

It is true, of course, that the question of how to formulate a problem

[8] R. von Mises, "On the Correct Use of Bayes' Formula," Annals of Math. Statistics, 1941, p. 191.

and how to state the criteria is a matter of taste and expedience, so long as the formulations are equivalent. The reason why we prefer our own development as presented in the first part of this paper is that it seems to us to be a logically more precise formulation which is a good deal easier to understand.

It is important to note that although the criterion is based on the entire class, this does not mean that we must estimate every element of that class. This can most easily be understood if we refer to the example which we gave before. The prediction that the probability that it would rain is .65, specifies the "best" odds, even though we may be interested in the weather only on one particular day. In the same sense, we have a "best" estimate, even though we may estimate only one of the risks which belong to the chosen class.

Since the type of inference discussed in this problem involves the highly controversial rule of Bayes-Laplace, we would like to add a very brief comment on the justification of this rule. This formula, commonly called simply the "Rule of Bayes", can be derived from the axioms of probability in two or three simple steps. Consequently, the arguments which have been raised against the application of this rule consist basically of the claim that we can never actually know all of the distributions which are involved. If we do not know the distribution $K(x)$, but merely assume its form a priori, we are guilty of distributing our ignorance in some arbitrary fashion, uniformly or otherwise.[9] The important consideration, therefore, is that we must have an *empirical basis* for the type of distribution to be used for $K(x)$; and casualty actuaries are indeed privileged because this type of information is seldom available in other fields of scientific inquiry. Consequently, it seems to us that Mr. Bailey is unnecessarily asking for criticism in his statement (on page 6) that he is considering the weights to be given to

> ". . . observed data in its combination with collateral *or with a priori knowledge.*"

All the indirect information *must* consist of collateral data or of reasonable inferences drawn from such collateral knowledge.

We are certain that it will please Mr. Bailey to hear that a good number of statisticians are disturbed by, what Professor Carnap calls the startling spectacle of unsolved controversies and mutual misunderstandings that appears in most standard treatises on probability and statistics. It is our sincere hope that Mr. Bailey's pioneer work in the field of credibility may lead to the elimination of some of these controversies and to a better understood and more general approach to the problem of statistical estimation.

WRITTEN DISCUSSION BY

*M. V. JOHNS, JR., National Bureau of Casualty Underwriters

Mr. Bailey's very interesting and important paper presents a novel departure from the conventional philosophies of statistical estimation. Although Mr. Bailey has concerned himself primarily with the derivation of estimation procedures through the application of Bayes' Theorem it seems to me that he

[9] If, for example, we put a priori $K(x)$ equal to the Gamma distribution, we might justifiably be accused of employing the principle of "Gamma—distributed ignorance," analogous to the principle of "Equally distributed ignorance" which is mentioned several times in Mr. Bailey's paper.

* By invitation.

has made an even more basic contribution in his recognition of the fact that the parameters characterizing a group of related probability distributions may properly be considered as stochastic variables under certain conditions. As Mr. Bailey has pointed out this concept has heretofore been almost completely ignored by the recognized authorities in the field of mathematical statistics.

It should be noted that the stochastic variation of a group of distribution parameters will be of a somewhat different sort than that of the variables characterized by the individual distributions. If we consider a group of variables such that broadly similar causal factors apply to all of the members of the group we may expect the distributions of the variables comprising this group to exhibit certain similarities. However, the parameters characterizing these distributions need not be identical for all members of the group since there may be specific influences operating to produce differences among the members. Once the individual variables comprising the group are defined the values of the parameters characterizing the distributions of each of these variables are fixed. However, these values certainly will not be evenly distributed throughout the range of all possible values since their variation is restricted by the underlying casual similarities existing among the variables. Thus we may say that the set of parameters so defined constitutes a sample from the statistical population composed of the values of the parameters of all possible distributions having the same underlying similarities. It is in this sense that we may consider the parameters as stochastic variables.

Since the credibility procedures discussed by Mr. Bailey are basically estimation processes it might have been more logical to derive them from some criterion of accuracy rather than from considerations explicitly involving inverse probabilities. In the following section I will present an outline of such a derivation which does not explicitly involve inverse probabilities and which assumes that *all* of the parameters (rather than just the expected values) of the various distributions are stochastic variables in the sense described above.

II.

This section will be devoted to the development of an estimation procedure predicated on the principles outlined above and to the application of this procedure to insurance statistics. The notation $E[A \mid B_i, (i = 1, 2 \cdots)]$ will be used throughout the following exposition to indicate the conditional expectation of A given the quantities B_1, B_2, B_3, etc. where these quantities B_i represent various parameters of the probability distribution of A.

Let x_{ij} be the ith variable of the jth class of variables where the number of variables in the jth class is n_j and there are N classes in all and where the x's are all mutually independent. The criteria for determining the arrangement of the variables into these classes will usually be such as to insure that each variable in a particular class will have properties more similar to those of other members of the class than to the properties of variables in other classes. The class of all the x's must be determined so that the probability distributions of the x's have certain general characteristics in common. Specifically, this class of variables must be defined so that every possible value of each parameter of the various probability distributions will be associated with a definite probability of occurrence in the sense described in the preceding section.

If all of the moments of a probability distribution are given then the dis-

tribution is completely determined so that we may define the parameters of any distribution in terms of its moments.

Let the parameters t_{ijk}, $(k = 1, 2, \cdots)$ of the distribution of x_{ij} be defined by

$$E[x_{ij}^k \mid t_{ijk},\ (k = 1, 2, \cdots)] = t_{ijk} \quad \text{for } k = 1, 2, \cdots.$$

Now each of these quantities t_{ijk} is associated with a probability determined by the general character of the class of all the x's. Therefore, we may define parameters of the *joint* distribution of all the t's as follows:

$$E[t_{ijk}^s \mid T_{ks},\ S_r,\ (k, s, r = 1, 2, \cdots)] = T_{ks} \text{ for } k, s = 1, 2, \cdots$$
$$\text{and } E[t_{ijr} \cdot t_{vjr} \mid T_{ks},\ S_r,\ (k, s, r = 1, 2, \cdots)] = S_r \text{ for } i \neq v$$
$$\text{and } r = 1, 2, \cdots.$$

These parameters T_{ks} and S_r do not entirely determine the joint distribution of the t's since the higher order product moments are not considered but they will be sufficient for the present investigation.

We will assume that it is desired to estimate the expected value of the arithmetic mean of the n_j variables comprising the jth class. This expected value will be represented by $t_{j1} = \frac{1}{n_j}\sum_i t_{ij1}$, and the estimate of t_{j1} will be represented by t'_{j1}. In order to determine the "best estimate" of t_{j1} we may set up a criterion of accuracy in terms of a minimum error variance. That is, we may minimize

$$\sigma_{j1}^2 = E[(t'_{j1} - t_{j1})^2 \mid T_{ks},\ S_r,\ (k, s, r = 1, 2, \cdots)]$$

with respect to t'_{j1} where t'_{j1} is considered as a function of $\bar{x}_j = \frac{1}{n_j}\sum_i x_{ij}$ and does not involve the t's. If no further restrictions are placed on t'_{j1} we may minimize σ_{j1}^2 as follows:

$$\sigma_{j1}^2 = \int_R \int_{R'} (t_{j1} - t'_{j1})^2\, P dt_{j1} d\bar{x}_j, \qquad \text{where } P \text{ is}$$

the joint probability density function of t_{j1} and x_j, R is the region containing all values of $\bar{x}_j$ and R' is the region containing all values of t_{j1}. For simplicity the variance is shown here as an ordinary double integral as if P were continuous throughout R and R'. Since t'_{j1} is independent of t_{j1} we may write:

$$\sigma_{j1}^2 = \int_R \left[\int_{R'} t_{j1}^2\, P dt_{j1} - 2t'_{j1} \int_{R'} t_{j1}\, P dt_{j1} + t'^2_{j1} \int_{R'} P dt_{j1} \right] d\bar{x}_j$$

Setting the partial derivative with respect to t'_{j1} of the quantity in brackets equal to zero and solving for t'_{j1} we have

$$t'_{j1} = \frac{\int_{R'} t_{j1} P dt_{j1}}{\int_{R'} P dt_{j1}} \qquad \text{as the value of } t'_{j1} \text{ which}$$

minimizes σ_{j1}^2. This is equivalent to $t'_{j1} = E[t_{j1} \mid \bar{x}_j, T_{ks}, S_r, (k, s, r = 1, 2, \cdots)]$ as in Mr. Bailey's derivation from Baye's Theorem. Suppose, however, that we wish to approximate this value of t'_{j1} by a polynomial of mth degree in $\bar{x}_j$. If we represent such a polynomial by $t'_{j1} = a_o + a_1\bar{x}_j + a_2\bar{x}_j^2 + \cdots + a_m\bar{x}_j^m$ we may evaluate the coefficients $a_o, a_1, a_2, \cdots a_m$ in terms of the T's and S's by minimizing σ_{j1}^2 with respect to $a_o, a_1, a_2, \cdots a_m$. The case of particular interest here is that in which the estimate is linear in $\bar{x}_j$. When dealing with insurance statistics a linear estimate is the most practical since the data available is usually insufficient for the evaluation of the constants involved in an estimate of higher degree. Thus, letting $t'_{j1} = a_o + a_1\bar{x}_j$ and minimizing σ_{j1}^2 with respect to a_o and a_1 we have

$$a_o = (1 - a_1)\, T_{11},$$

$$a_1 = \frac{T_{12} - S_1 + n_j\,(S_1 - T_{11}^2)}{T_{21} - S_1 + n_j\,(S_1 - T_{11}^2)}.$$

If a sufficient number of observed values of the variables x_{ij} were available it would be possible to estimate the parameters T_{11}, T_{12}, T_{21} and S_1 and hence obtain values for a_o and a_1. Unfortunately, since the data available to the insurance statistician are not sufficiently detailed for this purpose, further assumptions must be made in order to derive a workable procedure.

The insurance problem is essentially that of the estimation of pure premiums, so that we may define x_{ij} to be the total losses produced by the ith unit of exposure of the jth risk (or territory or manual classification depending on whether the estimate is being made for experience rating or for manual ratemaking purposes). The pure premium for the jth class will then be represented by $\bar{x}_j$. In order to obtain estimated values for a_o and a_1 we must first assume that each of the N classes designated by $j = 1, 2, 3, \cdots N$ is internally homogeneous. This means that $t_{ijk} = t_{vjk}$ for any i and v and for all values of k within the jth class. The parameter S_1 will then be replaced by T_{12} and we will have

$$a_1 = \frac{n_j}{\dfrac{T_{21} - T_{12}}{T_{12} - T_{11}^2} + n_j}$$

It can easily be shown that if the exposure basis is varied, the quantities T_{21}, T_{12}, and T_{11} will vary in such a way that a_1 will remain constant for any particular class. This is in accord with our intuitive conviction that the credibility coefficient associated with an insurance pure premium should be independent of the exposure basis used.

Since the values of x_{ij} are usually not available separately but only in the form of the average, $\bar{x}_j$, it will be necessary to make some assumptions regarding the form of the probability distribution of x_{ij} if we are to obtain some sort of estimates of T_{12} and T_{21}. The assumptions adopted henceforth are as follows:

(a) The probability density function of the losses x_{ij} will be non-negative and will have a discontinuity at zero since the probability that a unit of exposure will produce no losses is a definitite positive quantity.

(b) The claim frequency will follow a Poisson distribution so that the probability that x_{ij} will be zero (i.e., that the ith unit of exposure will produce no losses) will be given by e^{-c_j} where c_j = the expected claim frequency of the jth class.

The probability density function of x_{ij} under these assumptions is

$$f(x_{ij}) = \begin{cases} e_j^{-c}; \text{ for } x_{ij} = 0 \\ (1 - e_j^{-c})\, g(x_{ij}); \text{ for } x_{ij} > 0 \end{cases}$$

where

c_j = the expected value of the claim frequency, and
$g(x_{ij})$ = the probability density function for all losses greater than zero.

The variance of x_{ij} is then given by

$$t_{ij2} - t_{ij1}^2 = t_{ij1}^2 \left[\frac{k_j^2 + e^{-c_j}}{1 - e^{-c_j}} \right]$$

where

k_j^2 = the coefficient of variation associated with $g(x_{ij})$, i.e., the coefficient of variation of the losses greater than zero, and
t_{ij1} = the expected value of x_{ij} as before.

Since the data necessary for the direct evaluation of k_j are usually not available this expression for the variance of x_{ij} must be modified somewhat in order to reduce it to a form more useful for estimation purposes. In order to accomplish this we may make the following assumptions:

1) The coefficient of variation of the distribution of claims by size of claim is constant for a given type of coverage, and
2) The claim frequency is independent of the claim size.

Then it may be shown directly that

$$k_j^2 = \frac{(1 - e^{-c_j})\,(k_o^2 + 1)}{c_j} - e^{-c_j}$$

where k_o is the coefficient of variation of the distribution of claims by size of claim. The expression for the variance of x_{ij} then becomes

$$t_{ij2} - t_{ij1}^2 = t_{ij1}^2 \left[\frac{k_o^2 + 1}{c_j} \right]$$

We now have a foundation which makes possible the estimation of T_{11}, T_{12} and T_{2i} from available data. The value of k_o may be estimated from the appropriat observed distribution of claims by size of claim, and observed values of the pure premiums $\bar{x}_j$ and claim frequencies c_j are usually available.

In order to obtain the necessary relationships we may first note that

$$E\,[x_j \mid T_{ks}, (k, s = 1, 2, \cdots)] = T_{11},$$

$$E\,[x_j \mid T_{ks},\ (k, s = 1, 2, \cdots)] = \frac{T_{21}}{n_j} + \frac{n_j - 1}{n_j} \cdot T_{12}, \text{ and}$$

$$E\,[t_{i2} - t'_{i1} \mid T_{ks},\ (k, s = 1, 2, \cdots)] = T_{21} - T_{12}$$

Now, replacing these expected values by the corresponding observed averages and representing the observed values of x_j, c_j and k_o by X_j, C_j and K_o respectively and the estimates of T_{11}, T_{12} and T_{21} by T'_{11}, T'_{12}, and T'_{21}, respectively, we may derive the following equations for evaluating T'_{11}, T'_{12} and T'_{21}:

$$(1) \qquad \frac{\sum_j n_j \overline{X}_j}{\sum_j n_j} = T'_{11}$$

$$(2) \qquad \frac{1}{N}\sum_j n_j \overline{X}^2_j = T'_{21} + T'_{12}\left[\frac{\sum_j n_j}{N} - 1\right]$$

$$(3) \qquad (K_o^2 + 1)\frac{\sum_j \frac{n_j \overline{X}^2_j}{C_j}}{\sum_j n_j} = T'_{21} - T'_{12}$$

These estimates may be neither unbiased nor efficient but they probably represent the best that can be done with the available information.

The estimation formulae may now be put into the form:

$$(4) \qquad t'_{j1} = Z_j \overline{X}_j + (1 - Z_j)\, T'_{11}$$

$$(5) \qquad Z_j = \frac{n_j}{\frac{T'_{21} - T'_{12}}{T'_{12} - T'^2_{11}} + n_j}$$

where Z_j is the "credibility" of $\overline{X}_j$.

Equations (4) and (5) are of the form usually associated with experience rating credibility procedures but they could easily be adapted for use in manual ratemaking and should give more accurate results than the present ratemaking credibilities based solely on the observed number of claims.

Many experience rating plans have incorporated credibility tables based on relationships very similar in form to (5) except that the premium volume is substituted for the risk exposure n_j which, of course, is what would be obtained by multiplying the numerator and denominator of (5) by [T'_{11} ÷ permissible loss ratio]. However, because of the maximum single loss provision usually incorporated in experience rating plans, equations (4) and (5) are not strictly identical with the operations performed in experience rating. The quantity corresponding to $\overline{X}_j$ in experience rating is derived from a truncated distribution in which the losses are not allowed to exceed a certain prescribed value so that the expected value of these modified observations is no longer equal to

the quantity being estimated. If we let $\overline{Y}_j$ be the observed pure premium derived under the maximum single loss provision, equations (4) and (5) become:

$$(6) \qquad t'_{j1} = Z_j(\overline{Y}_j + W_j) + (1 - Z_j)\, T'_{11}$$

$$(7) \qquad Z_j = \frac{n_j}{A_j + B_j n_j}$$

where W_j represents a correction for the bias introduced by the maximum single loss provisions and where A_j and B_j are functions of the T's and also depend on the manner in which the maximum single loss is determined for the jth risk.

In practice the difference between (4) and (6) may be offset to some extent by the fact that the quantity corresponding to T'_{11} will actually be based on the average experience of all risks of a particular type, whereas there may be a selection in favor of the group of risks which are experience rated. Thus, if the average experience for the group of experience rated risks is better than that for all risks of the same type a certain upward bias will be produced by using the experience for all risks (as reflected in the manual rates) as the estimate of T_{11}.

The quantities of A_j and B_j in (7) cannot easily be evaluated from available information and about all that can be said at present is that A_j approaches $\frac{T'_{21} - T'_{12}}{T'_{12} - T'^{2}_{11}}$ and B_j approaches 1 as the allowable maximum loss is increased indefinitely.

I would like to emphasize the fact that this whole approach to the problem of credibility procedures is predicated primarily on considerations of accuracy and does not take note of the stability requirements which are surely necessary from the point of practicability. In fact, previous derivations of credibility procedures have been concerned mainly with obtaining sufficiently stable estimates with considerations of accuracy being strictly secondary. Since credibilities designed to produce maximum accuracy do not bear any close relationship to the expected relative amount of chance variation of the individual pure premiums, the use of such credibilities might not produce a set of estimates which could be readily used to establish a set of stable rates.

In this connection I would like to suggest that since the expected number of claims may be shown to be directly related to the expected chance fluctuation of the corresponding pure premium, perhaps credibility tables based jointly on the observed number of claims and the exposures (or premium volumes) would give results consistent with both accuracy and stability.

WRITTEN DISCUSSION BY *WILFRED PERKS

Assistant Actuary, Pearl Assurance Co. Ltd. of London

As a convinced supporter of the principles of inverse probability my sympathies are naturally with Mr. Bailey's approach. Whether a particular problem of statistical estimation involves prior ignorance or prior knowledge the one system of Bayes' theorem meets the requirements of the problem.

*By invitation.

With appropriate invariant rules to express prior ignorance, the results of Bayes' theorem in certain important cases are identical with those of confidence intervals and associated techniques. In cases where the prior knowledge is a precise statement of a prior probability distribution all schools would, I suggest, use Bayes' theorem, although these cases have been labelled "trivial" by certain statisticians. It is in the cases where the prior knowledge is imprecise that serious difficulties arise, both in principle and in practical application and it is with cases of this kind that Mr. Bailey's paper is concerned. There is much to be said for Professor Jeffreys' judgment that vague prior knowledge might well be ignored and an appropriate indifference rule applied. I judge, however, that Mr. Bailey's problems involve rather more than "vague" prior knowledge, although it is still "imprecise".

I am in complete agreement with Mr. Bailey that we should, if it is appropriate, try to express, even if only approximately, our prior knowledge in the form of the hypothetical results of a set of hypothetical past trials. This leads at once to the use of the beta distribution for prior probabilities in the problem of estimating the binomial parameter and to the use of the gamma distribution (a limiting case of the beta distribution) for the problem of estimating the Poisson parameter (a limiting case of the binomial) i.e. to formulae (12) and (17) respectively of the paper.

I am, however, troubled about three things:—

1. Is the prior knowledge that we are assumed to have prior knowledge about a super-population from which a particular population is supposed to have been selected at random? That is to say are we estimating the parameter of the particular binomial distribution selected from a known distribution of binomial populations?
2. Or is the prior knowledge that we are assumed to have, prior knowledge about the particular population? That is to say, have we made a prior estimate of the parameter?
3. Have the underlying conditions of operation and observation remained unchanged throughout and as between the circumstances applicable to the prior knowledge and those applicable to the past and future observations? That is to say are there any reasons to suppose that there is a secular or other systematic variation in the parameter concerned?

Even if we have no "knowledge" of kind (1) above, there must be a starting point for the prior probabilities to be used in the application of Bayes' theorem, although any significant amount of "knowledge" of kind (2) would tend to swamp the importance of the particular form of "knowledge" of kind (1). I can understand that in practice we may have good reason for assuming a particular value for the mean of the prior probability distribution but Mr. Bailey's processes call for an assumption about the standard deviation of the distribution or, what is the same thing, an assumption about the total number of hypothetical observation as well as the proportion of hypothetical successes i.e. we need to know the value of the indices in the beta distribution as well as their ratio. It is the basis upon which this standard deviation can suitably be judged that I am not clear about. This was the difficulty that long confused

the problem of a self-consistent indifference rule. A mean value of 1/2 was satisfactory but the standard deviation arising out of a uniform distribution led to trouble. The invariant rule independently devised by Prof. Jeffreys and myself has now got over this difficulty for the indifference case, but the problem still remains in Mr. Bailey's case. This standard deviation is, of course, the vital factor in determining the weights for combining the prior estimate of the parameter with the observed frequency ratio.

If there is a systematic variation of the kind mentioned under (3) above, the use of Bayes' theorem is inappropriate. In practice, unfortunately, the situation is all too often complicated in this way.

At the end of the paper Mr. Bailey refers to the "unsolved problem" of the frequency distribution of claims losses. This is essentially a multinomial problem which can perhaps be formulated in several ways. It is, however, the problem of estimating the $p_i(\Sigma p_i = 1)$ in a multinomial distribution.

If the p_i are linked by a mathematical formula the problem becomes one of estimating the parameters in the formula. Otherwise, the whole set of values of p_i have to be estimated jointly. I have examined the indifference problem in this case (J.I.A. LXXIII, 285) and R. E. Beard and I (J.I.A. LXXV, 75) have indicated the relative insignificance in practice of the correlation effect referred to by Mr. Bailey. It is usually sufficient in practice to assume that each p_i gives rise to an independant Poisson variable.

I should make it clear that I am not familiar with the rather extensive specialized literature in America on Credibility Procedures. My comments arise out of a reading of Mr. Bailey's paper alone and I realize that the points I have made may not be new and may have been answered already in the literature. Indeed, I cannot be sure that I am not misconceiving the problem altogether.

WRITTEN DISCUSSION BY L. H. LONGLEY-COOK

The author is to be congratulated not only on a most interesting and stimulating paper on credibility procedures in casualty actuarial work but also on an important contribution to the subject of inverse probability. Inverse probability has been considered in relation to actuarial work on a number of occasions and when Mr. Perks presented a paper to the Institute of Actuaries on the subject a few years ago a most interesting discussion resulted. My remarks, however, will be limited to the discussion of credibility procedures.

I fear this paper will be found difficult by most students and I have been wondering if the principal results can be brought out in a more simple manner without the use of inverse probability with all its pitfalls or too much loss of rigour. I hope the following demonstration will be of some assistance in this respect.

Following the author's development and using his notation, we first consider the case of the proportion of losses where an investigation shows H "successes" out of n "trials". It is desired to make the best estimate of the true loss frequency taking into account the prior knowledge but ignoring all question of trends, that is giving equal weight to all data. In the simplest form the prior knowledge will be H successes out of $\overline{n}$ trials and the best estimate of the loss frequency is clearly

$$\frac{\overline{H} + H}{\overline{n} + n}$$

which can be written

$$Z\frac{H}{n} + (1 - Z)m$$

where

$$Z = \frac{n}{\overline{n} + n} \quad \text{and} \quad m = \frac{\overline{H}}{\overline{n}}.$$

Since Z increases as n increases relative to $\overline{n}$, this shows that in the usual credibility formula greater weight should be given to larger volumes of observed data. It will be noted however that in this case if $\overline{n}$ is large compared to n practically no weight should be given to the current knowledge.

In practice the data making up a class are not homogeneous and we can imagine the prior knowledge being split up into a number of sub-groups with loss frequencies x_1, x_2, x_3, etc. Let the mean of these values be m and the variance σ^2. Although it may be unreasonable to assume that the distribution of the x's will follow any law, the best estimate which can be made, on the basis of prior knowledge alone, of the true loss frequency for some new sub-group is m subject to a variance σ^2. Also if the observed loss frequency of a new sub-group is H/n, the best estimate which can be made, on the basis of current knowledge alone, of the true loss frequency, q, is H/n with a variance, on the assumption of a Poisson distribution of $\{\sqrt{nq}/n\}^2$.

For rate making purposes we can use a combination of these two estimates

$$Z\frac{H}{n} + (1 - Z)m$$

The variance of this combination is

$$Z^2\{\sqrt{nq}/n\}^2 + (1 - Z)^2\sigma^2$$

Differentiating with respect to Z we find the condition for minimum variance is

$$2Z\{\sqrt{nq}/n\}^2 + (-2 + 2Z)\sigma^2 = 0$$

Using the approximation $q = m$, this becomes

$$Z = \frac{n\sigma^2}{n\sigma^2 + m}.$$

Hence we see that, even when the prior knowledge is large compared to the current data, if the current data consist of the experience of a sub-group and the sub-groups are not homogeneous one with another then more weight should be given to the experience of larger sub-groups.

Turning to the case where we are concerned with the dollar amount of losses instead of their number only, we can subdivide the total number of losses into groups according to size. Taking first the simple case of homogeneous data for losses of amount t, we have the proportion of claims of this size in the current data is H_t/n, and in the prior knowledge $\overline{H}_t/\overline{n}$. Hence the weight to be given to current knowledge is $n/(\overline{n} + n)$ whatever the size of the loss and no more weight should be given to the frequently occurring small losses. The

position is different when the prior knowledge can be divided into a number of sub-groups each with a slightly different experience. The formula is then

$$Z = \frac{n\sigma_t^2}{n\sigma_t^2 + m_t}$$

where σ_t^2/m_t is the ratio of the variance to the mean of the number of claims of size t. σ_t^2/m_t will normally decrease as t increases and hence, since for any sub-group n will be constant, Z will decrease as t increases. From this we see that in these circumstances more weight should be given to frequently occurring small losses.

It seems desirable to warn students that while standard credibility procedures are both necessary and desirable when a routine practice can be introduced, as for instance in Workmen's Compensation rate making, it is not generally practicable to replace actuarial judgment by credibility rules of thumb. If the actuary will see that he has a real knowledge of the data he is handling, how compiled, possibility of errors, changes in conditions which have occurred, etc., and will keep before him as yardsticks the square root of the number of claims and an approximate frequency distribution of claims by size, he will usually obtain a more satisfactory estimate of the rate he may expect in the future than by the blind application of any credibility formula.

REPLY TO DISCUSSIONS BY ARTHUR L. BAILEY

Dr. von Mises has provided us with a commentary on the theory of inference that only one with his broad knowledge of the many proposed solutions to the problems of statistical inference could state so concisely yet completely. It should be read and read carefully, preferably before reading my original paper.

Mr. Molina has been very kind in his comments. His contributions to the literature of mathematical statistics are almost unique because they evidence a determination to mold the mathematics to the practicalities of the case; instead of the reverse. His refusal to discard prior knowledge or collateral information in the analysis of observations has made him an outstanding advocate for inverse probabilities. I, as you should know, am personally very indebted to him for his kindness in going over an early draft of my paper and for the contributions he made to it, especially to the historical background of inverse probability theory.

The comments of Messrs. Perks, Longley-Cook and Freund and my recent reading of "Theory of Probability" by Harold Jeffreys, has convinced me that an estimate of x based on observations O_n (Freund's use of O_n instead of H' for the n observations is a distinct improvement in symbolism) should be symbolized as $E(x \mid O_n, K, L, C)$ where K represents the degree of prior knowledge as to the prior probability function $K(x)$, or the hypothesis substituted for such knowledge; where L represents the degree of prior knowledge as to the likelihood probability $P(O_n \mid x)$ or the hypothesis substituted for such knowledge; and where C represents the criteria selected as the basis of the estimate of the conditions imposed on the estimate.

Dr. Jeffreys has used a symbolism that expresses every probability in terms of the hypotheses made and has stressed the need for completely specifying all such hypotheses. He shows clearly that any evaluation of a posterior probability must be proportional to the product of the prior probability and the

likelihood probability. Similarly it could be shown that any use of an estimate based on an observation must involve either knowledge of or hypothesis as to both the prior probabilities and the likelihood probabilities as well as acceptance of the criteria utilized. It appears that much of the past confusion as to the relative merits of estimation procedures would have been avoided if the hypotheses regarding these probabilities as well as the criteria on which the estimates were based were always specified.

Mr. Freund calls attention to the fact that a chosen criterion must be such that whatever information is available or can be obtained will be sufficient to perform the method of estimation which has been defined. I would like to add another note of caution to this. Criteria should be avoided if they impose any conditions over and above what is necessary to provide the estimate. If the conditions are too broad, they may prevent the statistician from employing certain reliable and justifiable prior knowledge or collateral information.

Freund's use of the term "preferred" instead of "good" or "best" brings out only that tastes differ—and rightly so. For example, the "restricted" estimates produced by the criteria Freund suggests in Cases B and D would appeal to me in much the same way that "restricting" square pegs to round holes would. The condition that the variance of the estimates equal the variance of the thing being estimated, is in my opinion, an unsound one, especially when the correlation between the estimate and the thing being estimated is low. Although I have expressed this repeatedly to Freund, he still likes it proving that tastes differ and that "preferred" has no more useful meaning than "best". Let us then simply state what the criteria for an estimate is without characterizing it.

The real heart of the problem is that, to whatever extent knowledge is lacking as to the prior probabilities or the likelihood probabilities, the lack must be made up by hypotheses. One of the difficulties has been that criteria have been selected at times so as to completely hide the hypotheses implicitly made but not expressed. Take Freund's Case C as an example. His statement of the scope of the assumed knowledge, of the criteria applied, and of the results obtained are correct; but the simple condition that the estimate be a linear function of $\bar{x}i$ implied the hypothesis that the prior probabilities followed one specific distribution when the likelihood probabilities followed another specific distribution as I have shown in the paper now under discussion. The Beta and Binomial, the Gamma and Poisson, and the Normal and Normal were shown to be such paired hypotheses produced by the condition that the estimate be a linear function. An important reason for my writing the paper was to show what hypothesis as to the prior probabilities was involved in that apparently innocuous condition.

One of the most easily lost hypotheses is that implied in the use of a maximum likelihood estimate. The procedure is one that completely disregards the prior probabilities but produces an estimate in a form that requires the user of the estimate to assume that $K(x) = k$ for all possible values of x. The statistician refuses to make the hypothesis, but forces his client to make it.

Running throughout my paper and the discussions is the confusing difference in the concept of probability when we are dealing with a heterogeneous instead of a homogeneous population. Most probability theory and most statistical methods assume a homogeneous population, for each individual of

which the probability is the same—some constant value, known or unknown. In casualty insurance our basic assumption is that the insurance hazards differ from risk to risk as well as from classification to classification. We have only heterogeneous populations, for each individual of which the probability is different—a variable whose value is never known although we frequently wish to estimate it. When I deal, as I do, with the probability of a probability, I am dealing with a concept that never occurs in homogeneous populations and, therefore, with a concept that is disturbing to any newcomer to the field of heterogeneous populations.

Mr. Johns actually was the original cause of my paper. In the fall of 1949 we jointly undertook to determine the most effective split of losses between "normal" and "excess". We were stymied in that project by a philosophical snag. His training, which not only exceeded mine but was twenty years more up to date, would not permit treating a parameter as a variable and required that he impose as a condition, $E(x/B) = B$, to obtain an "unbiased" estimate of B. To proceed along the lines of my previous work on credibility violated his training. To follow his training meant that no split of losses was justifiable. We deserted the original project to study the philosophies of probability theory.

In textbook after textbook the only acknowledgment of the prior knowledge or collateral information that actuaries recognize in the credibility formulas was to be found in the one or two paragraphs covering the theory of inverse probability. Starting with this, guided by Mr. Molina's paper showing a practical application of inverse probability theory, and fortified with a recent paper by Mr. Freund showing that the generalized Bayes' Rule was still alive, my paper evolved.

If I had then had the 1948 edition of Jr. Jeffreys' book I could have shown Mr. Johns that the generalized Bayes' Rule (Mr. Jeffreys' theorem 10) was the basis of all evaluations of probabilities from observations, all tests of significance, and of all estimates. It would have been quite apparent that all of the accepted procedures taking up 99.8 per cent of the space in statistical texts are based on, or can be derived from, the theory of inverse probability in combination with one of the following three assumptions:

(1) The number of observations is so great that the effect of any prior knowledge or collateral information is trivial and can be disregarded.

(2) There is no prior knowledge or collateral information of any value and the theory of equal ignorance, for which $K(x) = k$, or the theory of equal indifference recently devised by Mr. Perks, for which $K(x) = k/x$, is applicable.

(3) We are dealing with a homogeneous population so that x has only one value, say A, and $K(x) = 0$ except that $K(A) = 1$.

Each of these three assumptions produce a credibility of 100 per cent for the indications of the observations; but, they are the only ones that will.

Mr. Johns has made two very substantial contributions. First, he has completely generalized the development of estimation procedures when parameters of sub-populations are treated as variables. Secondly, although he has dealt again with the case when the observation, H, is the product of the parameter, x, which is to be estimated, and an independent variable, h, with the restriction that the variance of h is constant for all values of x, his

procedure is such that the restriction can readily be removed. To do so would be of especial importance to us because it would produce a credibility formula for classification pure premiums or risk experience rating modifications not heretofore available.

In my paper I developed the procedures without regard to the source of the knowledge as to the prior probabilities or of the values of m and σ^2 into which that knowledge was to be concentrated. Mr. Perks has pointed out, and rightly so, that we should be much concerned with the source of such knowledge in any particular application. I have indicated in previous papers the general sources of such knowledge and will try to summarize them briefly here.

When we have no prior knowledge as to the values of x or of the values m and σ^2 it is contemplated that, if we have made observations for each of N individuals, we select the values of m and σ^2 which would lead us to expect to obtain the observed mean and variance of H. Such a selection will even evaluate the prior probabilities if we accept the suggested functional forms. When we do have previous estimates of the values of x, say y, it is intended that a new unknown be estimated namely $x^1 = x/y$ from adjusted observations of $H^1 = H/y$. The mechanics of performing such evaluations from collateral information and prior estimates is by no means settled and considerable work needs to be done along those lines.

Mr. Longley-Cook has indeed simplified the presentation of a demonstration that credibility procedures should be used in dealing with heterogeneous populations. Both as to his closing remarks on the desirability of being bound to the use of a mechanically applied credibility formula, and as to Mr. Johns' remarks on the desirability of maintaining stability in rates, I can only comment that the present matter under study is how to evaluate the indications of the statistical experience and not how to use such evaluations in making rates for the future. The combination of the indications of the past with actuarial judgment, or even with biased opinion, is another and very different study involving personal rather than mathematical equations.