TRANSACTIONS OF SOCIETY OF ACTUARIES
1965 VOL. 17 PT. 1 NO. 47AB

BAYESIAN STATISTICS

DONALD A. JONES

INTRODUCTION

THE objective of this paper is to bring Bayesian statistics to the attention of the members of the Society of Actuaries. This may have been done superficially when *Fortune* published an article, "The Science of Being Almost Certain" (February, 1964), that gave equal photographic coverage to Bayesian statistics and to classical statistics in a broad brush look at the recent history of statistics. However, the purpose here is to present a more technical exposition of the rudiments of Bayesian statistics. The exposition will not be sufficient to make the readers practicing Bayesians, but I hope that it will induce them to share my optimism for the future applicability of statistics to actuarial problems under the Bayesian outlook.

The foundations of Bayesian statistics is laid on a personalistic definition of probability in contrast to the statistical (or relative frequency) definition, which is at the heart of classical statistics, that is, the general set of ideas which has dominated the statistics classrooms for the last thirty years. In Section II we shall see that these personalistic probabilities satisfy the same axioms, or laws, that statistical probabilities satisfy, so the Bayesians' calculus of probability will not be new. On the other hand, the definition of probability forms the bridge between mathematical probability and the applications. Thus, what the Bayesian calculates and how he interprets his calculations will be new and will often be in conflict with classical training, but I think that his interpretations will rarely be in conflict with intuition.

The purpose of the first section is motivation—the need for something like personalistic probability in graduation theory is indicated. A discussion of personalistic probability at the Society Examination II level then follows in Section II. Sections III and IV contain the statistical inference material of the paper by means of discussing two examples. Some of the criticisms of classical statistics which have induced statisticians to look for new methods are illustrated by the examples in Section III. A four-step procedure for Bayesian analysis is illustrated by these same examples in Section IV.

I. SOME HISTORY

As a point of departure for our look at the relationship between the statistician's definition of probability and his statistical methods let us use

E. T. Whittaker's introduction of his difference-equation method of graduation, which he first read before the Edinburgh Mathematical Society in 1919. Difference-equation graduation means the theory that is chapter 5 of M. D. Miller's *Elements of Graduation* [13]. More precisely, assume that we have n observed values, $u_1, u_2, \ldots, u_n$, from which we wish to obtain n graduated values. For the given observed values define a function of n variables, say,

$$L(y_1, y_2, \ldots, y_n) = F(y_1, y_2, \ldots y_n) + hS(y_1, y_2, \ldots, y_n), \quad (1)$$

where

$$F(y_1, y_2, \ldots, y_n) = \sum_{x=1}^{n} W_x(y_x - u_x)^2,$$

$$S(y_1, y_2, \ldots, y_n) = \sum_{x=1}^{n-k} (\Delta^k y_x)^2,$$

h is a positive number, k is a positive integer, and the W_x's are suitably chosen weights (e.g., exposures when graduating mortality rates). Of course, h, k, the u_x's, and the W_x's are variables too, but they are to be constant in the determination of *a* set of graduated values. Both F and S are commonly used measures of fit and smoothness, respectively, each of which decreases with "improvement." Thus, if

$$L(y_1^{(1)}, y_2^{(1)}, \ldots, y_n^{(1)}) < L(y_1^{(2)}, y_2^{(2)}, \ldots, y_n^{(2)}),$$

then the $y_x^{(1)}$'s are a better set of graduated values than are the $y_x^{(2)}$'s. *The* graduated (by the difference-equation method) values are the numbers $v_1, v_2, \ldots, v_n$ such that $L(v_1, v_2, \ldots, v_n) \leq L(y_1, y_2, \ldots y_n)$ for all y's. The analysis and algebra to find these values by desk calculator for $W = 1$ and $k = 2$ are given in Miller's monograph.

Whittaker's principal contributions to this method of graduation are a justification for adopting $v_1, v_2, \ldots, v_n$ (as defined above) as the graduated values and the derivation of the difference equation which the v's must satisfy in order that they will minimize L. The algebra needed to find the v's, which is given in Miller [13], is Robert Henderson's contribution to this graduation method—often called the "Whittaker-Henderson method." We will limit our consideration here to Whittaker's justification of the method as given in *The Calculus of Observations* ([22], p. 302) or in his "On a New Method of Graduation" [20].

Whittaker, following George King ([11], p. 114), starts with the premise that the objective of a graduation is to find the "most probable" values. Thus to every set of possible true values (i.e., every ordered set of n num-

bers $y_1, y_2, \ldots, y_n$), he assigns a probability.[1] The first probability assigned is called the antecedent probability, and it answers "How probable?" before observations are considered. Now, according to Whittaker, before observed values are seen, some sets of y's are more probable than others due to conceptions of smoothness. Thus the antecedent probability should be a function of a measure of smoothness, for example,

$$S(y_1, y_2, \ldots, y_n) = \sum_{x=1}^{n-3} (\Delta^3 y_x)^2.$$

Now S is a sum of squares, so "by analogy to the normal frequency law" Whittaker assigned

$$\text{Prob. } (y_1, y_2, \ldots, y_n) = c \cdot e^{-(1/2)hS} \tag{2}$$

where h is an arbitrary constant and c is the constant to make the total probability 1.

To make inferences from observations containing random errors, we must have a probability distribution for the errors. Whittaker assumed each observed value minus the corresponding true value to be normally distributed with mean zero and known variance. Precisely, the conditional probability of observing u_1 when the true value is y_1 is

$$K_1 e^{-(1/2)W_1(u_1-y_1)^2},$$

where K_1 and W_1 are constants, like c and h above. Similar probabilities hold for each index. If, in addition, the errors of observation are assumed to be independent, then the conditional probability of observing u_1, $u_2, \ldots, u_n$ when the true values are $y_1, y_2, \ldots, y_n$ is

$$K e^{-(1/2)\sum_{x=1}^{n} W_x(u_x-y_x)^2}. \tag{3}$$

Whittaker's "most probable" values are to be the most probable after considering the observed values; thus he wants the conditional probability that the true values are $y_1, y_2, \ldots, y_n$, given that the observed values are $u_1, u_2, \ldots, u_n$. This conditional probability is the product of (2) and (3) divided by the marginal probability for the observed values at $u_1, u_2, \ldots, u_n$, say, $G(u_1, u_2, \ldots u_n)$,

$$\frac{K e^{-(1/2)(F+hS)}}{G(u_1, u_2, \ldots, u_n)}. \tag{4}$$

Whittaker's final step is to find those y's which maximize (4), or, since G and K are not functions of the y's, it is equivalent to find those y's which

[1] Today Whittaker would have talked of probability density for these continuous random variables.

minimize $F + hS$. Thus, we have Whittaker's justification of his method of graduation.

To use or not to use the difference-equation method has probably been decided on computational grounds rather than on the merits of its theoretical justification. However, G. J. Lidstone and Whittaker corresponded in the *Transactions of the Faculty of Actuaries* (XI, 233–37) concerning two criticisms which Lidstone had made of Whittaker's justification. In addition to historical interest, their correspondence illustrates a need for a definition of probability.

One of Lidstone's criticisms was of the antecedent probability given by (2) when applied to the usual problem of seeking graduated mortality rates. The maximum value of (2) is attained when, and only when, $S = 0$, and this is equivalent to $\Delta^3 y_x = 0$ for $x = 1, 2, \ldots, n - 3$. From this Lidstone argued (correctly) that the "most [antecedent] probable" mortality rates would lie on a second-degree polynomial. He then asserted that "this assumption [about the probability distribution] is simply not true, but rather one contradicted by general previous experience." Second, Lidstone criticized the theory's lack of a formula for the arbitrary h (in his presentation Whittaker had suggested that all the W's be taken equal for simplicity, which explains Lidstone's silence about no formula for the W's). He summarized: "Thus the method starts with a hypothesis which is not in accordance with experience, and ends with a constant which is not determinable from the data. . . . It is for this reason that it appears to me to be better (as it is simpler) to reach the expression $F + hS$ by more general and less theoretical considerations."

Lidstone's criticisms are much in the spirit of classical statistics; for example, his first criticism seems to imply the existence of common experience evaluated identically by all concerned and thus leading to unique probabilities, just as we have unique "areas." Whittaker's interpretation of his graduation method is not unlike that which would be given by a Bayesian statistician today except that a good formulation of personalistic probability did not exist at the time to put meaning in the probabilities assigned. In his published response to Lidstone, Whittaker ignored the h criticism and claimed that "most probable" was undefined before consideration of the observed values, so Lidstone's parabola of mortality rates was a red herring. Lidstone responded, "I used the term 'most probable' with the ordinary meaning 'having the greatest chance'—in this case the *a priori* chance discussed in Whittaker's hypothesis," and so ended the public correspondence. As we view the history of the difference-equation method of graduation up to today, it is clear that most actuaries have shared Lidstone's viewpoint. "Smoothness" is of the eye and

the mind, so perhaps today's personalistic definition of probability would have settled the misunderstanding between Whittaker and Lidstone and hence shaped the destiny of Whittaker-Henderson graduation in a different form.

With this look at a point in actuarial history where there existed a need for a precise definition of probability, let us turn to the two primary competing definitions today.

II. PERSONALISTIC PROBABILITY

What meaning do you give to: "The probability of getting a head on a single toss of this coin is 0.5"? I would expect answers which could be classified in one of the following four categories: (*a*) "head" is one of two equally likely outcomes due to symmetry; (*b*) in a long sequence of repeated tosses, the proportion of outcomes that are heads is 0.5; (*c*) "I'd bet even money for a head, and I'd bet even money against a head on a given toss of the coin"; and (*d*) "I can't say exactly, as I have always left that to my intuition." The first two of these responses are nonpersonal in the sense that they say something about the coin (and the flips) but nothing about you. In this same sense the latter two are personal—in fact, the last one says something about you and nothing about the coin. To pursue the first and the last responses would be digressions here because the controversy today is primarily between classical statisticians, who would give response (*b*), and the Bayesians, who would give response (*c*).

Response (*b*) is in the spirit of the *statistical definition*, that is the "long-run frequency" concept of probability which is found in most beginning (classical) statistics textbooks. All the textbooks listed in the 1964 Syllabus for Part II give this definition (e.g., [8], p. 10: "Similarly, for each event A we have a probability, $P(A)$, representing the proportion of times the event A occurs in a long sequence of repetitions of the random experiment"). This statement follows an informal postulate as to the existence of random experiments: "experiments repeatable under essentially constant conditions," so that the "proportions are found to be stable for large n."

Let us not halt for either the debate on the tenuous representation of the abstract "random experiment" by the real experiments of applications or to inquire how one "would find" $P(A)$. Both of these are important and unsettled challenges for adherents of the statistical definition, but neither is the real issue here. The chief objection of Bayesians is raised against the limited domain of the definition.

Bayesians believe that the role of statistics is to furnish a guide for consistent behavior in the presence of uncertainty; thus they seek a prob-

ability which can serve uncertainties of all origins, not just those of "random experiments." Indeed, they believe that the exclusion of uncertainty about matters of fact from the domain of (statistical definition) probability has foiled the development of statistical inference. Edwards, Lindman, and Savage put it this way:

> With rare exceptions, statisticians who conceive of probabilities exclusively as limits of relative frequencies are agreed that uncertainty about matters of fact is ordinarily not measurable by probability. Some of them would brand as nonsense the probability that weightlessness decreases visual acuity; for others the probability of this hypothesis would be 1 or 0 according as it is in fact true or false. Classical statistics is characterized by efforts to reformulate inference about such hypotheses without reference to their probabilities, especially initial probabilities.
>
> These efforts have been many and ingenious. It is disagreement about which of them to espouse, incidentally, that distinguishes the two main classical schools of statistics. The related ideas of significance levels, "errors of the first kind," and confidence levels, and the conflicting idea of fiducial probabilities are all intended to satisfy the urge to know how sure you are after looking at the data, while outlawing the question of how sure you were before. In our opinion, the quest for inference without initial probabilities has failed inevitably [(6), p. 196].

Now let us turn to the alternative definition adopted by Bayesians.

The *personalistic definition* of probability, which is based on a postulated consistent behavior of the individual relative to his state of information, experience, and opinions, may be given the following formulation in terms of gambling. "His [the individual's] probability for the event A, denoted by $P(A)$, is the amount that he is willing to pay if A does not obtain, in return for a promise to receive $1 - P(A)$ if A does obtain (betting on A), *and* he is willing to accept the reverse gamble." The notation is not adequate, for it reflects neither the individual nor his current state of information, etc. However, this inadequacy is more tolerable than cumbersome notation. To think of the bets in terms of money is an approximation to more realistic and rigorous formulations which include a theory of utility. The reader who seeks rigor and completeness may find a formulation of that order and many of the arguments for and against such a definition in the first five chapters of L. J. Savage's *Foundations of Statistics*, which also has an excellent annotated bibliography.

"Consistent behavior" means that the individual will not set up a series of gambles (viz., personal probabilities) which will result in no gain for all outcomes and a loss for at least one outcome. Abstractly, this means that

he must assign his probabilities so they satisfy certain rules or axioms. Fortunately, these are the familiar ones:

(i) $P(A) \geq 0$;

(ii) $P(S) = 1$;

(iii) $P(A + B) = P(A) + P(B) - P(AB)$.

S, as used in equation (ii), is the entire sample space or the certain event.

As an illustration of how the necessity of the axioms follows from consistent behavior, let us look at the argument for (iii), the so-called addition theorem. Let A and B be two events of the consistent individual's world. Suppose that he bets on A, bets on B, bets against AB, and against

TABLE 1

GAINS

Bet	Outcomes			
	AB (1)	$A\bar{B}$ (2)	$\bar{A}B$ (3)	$\bar{A}\bar{B}$ (4)
On A	$1-P(A)$	$1-P(A)$	$-P(A)$	$-P(A)$
On B	$1-P(B)$	$-P(B)$	$1-P(B)$	$-P(B)$
Against AB	$-[1-P(AB)]$	$P(AB)$	$P(AB)$	$P(AB)$
Against $A+B$	$-[1-P(A+B)]$	$-[1-P(A+B)]$	$-[1-P(A+B)]$	$P(A+B)$

$A + B$ [the motivation for this choice is by viewing (iii) as $P(A) + P(B) = P(AB) + P(A + B)$]. His possible gain for each bet and outcome is shown as an entry in Table 1.

The net gain from all bets for a given outcome is the total of the entries of the appropriate column. Observe that each column total is $P(AB) + P(A + B) - P(A) - P(B)$; in other words, for this combination of gambles the individual's net gain is the same for all outcomes. Consistent behavior requires that this net gain not be negative, so we have

$$P(AB) + P(A + B) - P(A) - P(B) \geq 0 .$$

On the other hand, the individual must be willing to accept the reverse of each of the four gambles, and thus he must choose his probabilities to satisfy also

$$-[P(AB) + P(A + B) - P(A) - P(B)] \geq 0 .$$

These two inequalities imply (iii).

These properties of personalistic probabilities are sufficient to validate the familiar methods of calculation with classical probabilities—except those calculations which involve conditional probabilities. For the adherent to the personalistic definition, the conditional probability of A, given B, denoted by $P(A\,|\,B)$, is word for word like that of $P(A)$, with the additional provision that all payments are canceled if B does not obtain. The consistent-behavior postulate requires that the "multiplication law" of probability holds for the personalistic conditional probabilities, that is

(iv) $$P(A\,|\,B) \cdot P(B) = P(AB) .$$

If $P(B) = 0$, then $P(AB) = 0$, and $P(A\,|\,B)$ is undefined, so (iv) should be interpreted as having a zero on each side.

The four properties (i), (ii), (iii), and (iv) describe, mathematically, both the personalistic probability and the statistical probability. All the familiar identities that we learned for statistical probabilities hold for the personalistic probabilities. Thus, if the assignment of both kinds of probabilities "is the same" in an application, then any derived probabilities will be the same. For example, suppose that a personalistic probabilist and a statistical probabilist are observing a crap game. If they both assign the probability of 1/36 to each of the 36 possible throws with a pair of dice, then they would follow the same calculating rules, as outlined in Table 2, to reach 244/495 as the probability of winning (i.e., throwing a 7, an 11, or "making the point"). The personalistic probabilist would say: "Since I am prepared to bet 1 to 35 on any one of the 36 outcomes for a single roll of the dice, I must be prepared to bet 244 to 251 on winning in a single crap game." The statistical probabilist would say: "If these dice were rolled a large number of times, about 1/36 of the outcomes would be of each type; therefore, if the dice were rolled for a large number of crap games, then the shooter would win about 244/495 of them."

Bayes' theorem is one of the identities which is a consequence of properties (i)–(iv); therefore, it is a theorem in both theories of probability. For two given events, say, A and B, we have from (iv)

$$P(A\,|\,B)P(B) = P(AB) = P(B\,|\,A)P(A) .$$

If $P(A) > 0$, we may write

(v) $$P(B\,|\,A) = \frac{P(A\,|\,B)P(B)}{P(A)} ,$$

which is the fundamental form of Bayes' theorem.

In frequent applications of (v), B is one event in an exhaustive set of mutually exclusive events, say, $B_1, B_2, \ldots$; hence

$$P(A) = \sum_i P(AB_j) = \sum_i P(A \mid B_j)P(B_j)$$

may be substituted in the denominator of (v) to have

$$P(B_r \mid A) = \frac{P(A \mid B_r)P(B_r)}{\sum_i P(A \mid B_j)P(B_j)}. \tag{v a}$$

For those of us who pride our Hall and Knight training, we may change notation by $P(B_r \mid A) = Q_r$, $P(A \mid B_r) = p_r$, $P(B_r) = P_r$ to have the form given on page 393 of Hall and Knight's *Higher Algebra;*

$$Q_r = \frac{p_r P_r}{\Sigma(pP)}. \tag{v b}$$

If B is one of a family of events for which one may calculate $P(B \mid A)$, the form

$$P(B \mid A) \propto P(A \mid B)P(B), \tag{v c}$$

where you should read "is proportional to (with respect to the variable B)" for the symbol $\propto$, may be sufficient. For example, in some applications one looks for the maximum (over a set of B's) of $P(B \mid A)$. All these forms of Bayes' theorem assume positive probabilities to give meaningful results. After notation is given for continuous probability distributions, we can give the continuous analogues of (v) and (v c).

The essence of (v) first appeared in 1763 in a paper written by Thomas Bayes [3] and published posthumously by Richard Price, who is famous to actuaries for the construction of the Northampton Tables. During the nineteenth century, Bayes' theorem was at the center of many fiery controversies concerning inverse probability. One such discussion, upon the occasion of E. T. Whittaker's address to the Faculty of Actuaries in 1920, is given in the *Transactions of the Faculty of Actuaries* [19]. Statistical analysis in the period from 1920 to 1960 was dominated by the work of R. A. Fisher, E. S. Pearson, J. Neyman, and others who could not find a place for Bayes' theorem in their analysis. Thus, when I looked in the Syllabi (1950–64) of the Society for a convenient reference to Bayes' theorem, the only books which gave the theorem were Cramér's book [5], which is mentioned as a secondary text, and the third edition of Hoel's book [10], which was first placed on the 1963 Fall Syllabus. As documen-

tation of statisticians' increasing interest in Bayes' theorem, we may observe that the first two editions of Hoel's book did not mention Bayes' theorem—but the third edition devotes seven pages to it.

For the benefit of those who followed the Syllabus too closely and are in contact with Bayes' theorem for the first time, we may calculate $P(B_k|W)$ for $k = 2, 3, \ldots, 12$, in the crap-game example given above. $P(B_k|W)$ is the conditional probability that a k was rolled on the first roll, given that the dice roller won the game. These probabilities are in column (4) of Table 2. Formulas (v a) and (v c) are illustrated by the relationships between column (4) and columns (1), (2), and (3).

TABLE 2

B_k is the event that a k is rolled on the first roll.
W is the event that the shooter wins the crap game.

k	$P(B_k)$ (1)	$P(W\|B_k)$* (2)	$P(WB_k)$ (3) (1)×(2)	$P(B_k\|W)$ (4) (3)÷(244/495)
2.....	1/36	0	0	0
3.....	2/36	0	0	0
4.....	3/36	1/3	1/36	55/976
5.....	4/36	2/5	2/45	88/976
6.....	5/36	5/11	25/396	125/976
7.....	6/36	1	1/6	330/976
8.....	5/36	5/11	25/396	125/976
9.....	4/36	2/5	2/45	88/976
10.....	3/36	1/3	1/36	55/976
11.....	2/36	1	1/18	110/976
12.....	1/36	0	0	0
Total..	1	$P(W)=244/495$		1

* $P(W|B_k)$ equals the probability of rolling a k before a 7. Therefore $P(W|B_k) = P(B_k) + [1 - P(B_k) - P(B_7)]\ P(W|B_k)$, or, $P(W|B_k) = P(B_k)\ [P(B_k) + P(B_7)]$.

We have observed that Bayes' theorem is true for both probabilities defined in this section. It seems reasonable to ask why Bayes' theorem is so useful in a statistical theory based on personalistic probability that such is called Bayesian statistics and yet classical statisticians have found so few applications for Bayes' theorem that it is not discussed in their basic textbooks. The purpose of statistical analysis is to transform knowledge on the basis of observed data (usually experimental)—and hence to guide action. According to personalistic probability, one's knowledge is described by his probabilities, so a statistical analysis based on personalistic probability must transform probabilities on the basis of observed data. The postulated consistent behavior requires that these transformed prob-

abilities be the conditional probabilities that were held before observing the data. That is, if $P(B|A)$ to $1 - P(B|A)$ are your odds for B with bets canceled if A does not obtain—then, after learning that A obtains, your odds must be $P(B|A)$ to $1 - P(B|A)$. The importance of Bayes' theorem then is to provide the algorithm to calculate the conditional probabilities. This will be more clearly illustrated by the two examples discussed in Section IV.

Notation

Usually the events in an application of probability are conveniently defined in terms of "random variables," that is, the A's and the B's of our previous equations are given by statements of the form $x = 2$, $y \leq 3$, $v < V < v + dv$, etc. For a discrete random variable we usually say, "Let $p(x)$ be the probability function for X," which means $p(x) = P(X = x)$ for all numbers x. Thus you must remember that p is the probability function for X when you see $p(5)$, say. Since personalistic probability produces more random variables (the classical statistician's parameters have personalistic probability distributions), I prefer to show this probability function "ownership" by a subscript. Thus, for discrete random variables

$$p_X(a) = P(X = a)$$

and

$$p_{X|\theta}(a|t) = P(X = a|\theta = t).$$

and for continuous random variables

$$P(a < X \leq b) = \int_a^b p_X(x)\,dx$$

and

$$P(a < X \leq b \mid \theta = t) = \int_a^b p_{X|\theta}(x \mid t)\,dx.$$

In this notation Bayes' theorem may take the forms

$$\text{(va')} \qquad p_{\theta|X}(t|x) = \frac{p_{X|\theta}(x \mid t) \cdot p_\theta(t)}{\sum_i p_{X|\theta}(x \mid i)\, p_\theta(i)}$$

$$\text{(vc')} \qquad p_{\theta|X}(t|x) \propto p_{X|\theta}(x \mid t)\, p_\theta(t).$$

III. CRITICISMS OF CLASSICAL STATISTICS

While there has never been a single methodology of statistics acceptable to all statisticians, there has been a general set of ideas, based on the statistical definition of probability, which has dominated the classrooms of this continent, and to a slightly lesser degree those of the rest of the world,

for the past thirty years. These ideas will be called "classical statistics"; however, the reader is warned that they pertain to a heterogeneous group of thinkers who show sharp disagreement on some points.

The procedures of classical statistics which receive criticism from many statisticians, including Bayesians, might be termed "averaging procedures." These procedures, such as confidence intervals, unbiased estimation, significance tests, and randomization, promise to do well (on the average) in a long run of identical inference situations, but they sometimes produce anomalous results on the way. A confidence interval example and an estimation example will be given to illustrate these anomalies and criticisms. These examples will be analyzed again by the methods of Bayesian statistics following the description of those methods in the next section.

Example CI (Confidence Interval).—An interval estimate for a parameter, θ, is an interval in which θ is asserted to lie and a percentage which describes the quality of the assertion. The most common answer given by classical statisticians for an interval estimate is a confidence interval. (Some statisticians would use the late R. A. Fisher's fiducial interval, which may or may not coincide with the confidence interval.) A *confidence interval* is defined as follows: *Let* X *denote the outcome of the random experiment to be performed, with the understanding that* X *may be a vector of observations, say,* $X = (X_1, X_2, \ldots, X_n)$. *If* L(X) *and* R(X) *are two functions such that*

$$P[L(X) \leq \theta \leq R(X) \mid \theta \text{ is the true value}] = p,$$

for all θ, *then* [L(X), R(X)] *defines a set of 100 p per cent confidence intervals for* θ. The classical statistician asserts that if he uses L and R to compute confidence intervals, then the proportion of times that his intervals will contain θ in a long run of such inference problems is p. One of the most familiar confidence intervals is for the unknown mean, θ, of a normally distributed population with an unknown variance (see [10], pp. 275–76). X is the vector of $n (n > 1)$ independent observations on the population, and

$$L(X) = \bar{X} - t\sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n(n-1)}}$$

and

$$R(X) = \bar{X} + t\sqrt{\frac{\sum_i (X_i - \bar{X})^2}{n(n-1)}},$$

where t is chosen to obtain a desired p. While the probability assertion of the confidence-interval construction is correct, it is sometimes obtained by assigning long intervals almost certain to contain the parameter to some data points and by assigning short intervals with little chance of containing the parameter to other data points. The anomalies produced by such a construction are not present for the normal distributions, which possess certain symmetries, but they are vivid for the uniform distributions of the following example.

Suppose that there exists a population uniformly distributed over an interval of length 2, with the center of the interval (the mean), θ, unknown. Graphically, the probability density function for a random observation on this population is shown in Figure 1. Let X (an n-dimensional

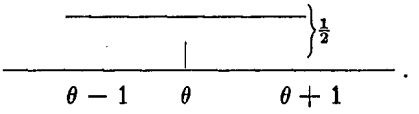

FIG. 1

vector) denote the outcome of obtaining n independent observations on this population. A confidence interval for θ is given by defining $L(X)$ and $R(X)$ to be, respectively, the smallest and the largest of the n observations. We must calculate the confidence level, $P[L(X) < \theta < R(X)]$, to complete the description of the intervals produced.

$$P[L(X) < \theta < R(X)] = 1 - P[L(X) < R(X) < \theta] - P[\theta < L(X) < R(X)]$$

$$P[L(X) < R(X) < \theta] = P \text{ (all } n \text{ observations will be left of } \theta) = (\tfrac{1}{2})^n$$

$$P[\theta < L(X) < R(X)] = P \text{ (all } n \text{ observations will be right of } \theta) = (\tfrac{1}{2})^n.$$

Therefore

$$P[L(X) < \theta < R(X)] = 1 - (\tfrac{1}{2})^{n-1},$$

and $[L(X), R(X)]$ will produce $100[1 - (\frac{1}{2})^{n-1}]$ per cent confidence intervals for θ. To see the promised anomalies, suppose that the smallest observation is 1.5 and the largest is 3.0, so the $100[1 - (\frac{1}{2})^{n-1}]$ per cent confidence interval is [1.5, 3.0]. This interval certainly contains θ, since these two extreme observations cannot be on the same side of θ. Again, if the smallest is 1.4 and the largest is 1.5, the $100[1 - (\frac{1}{2})^{n-1}]$ per cent confidence interval is [1.4, 1.5], yet it should, in fact, instill little confidence. This example is not known to have any immediate practical value, but it illustrates the lesson well. The percentage for a confidence interval is the probability of a "successful" experiment and is not the probability that this

experiment was successful. In short, the classical statistician's security in his long-run average is little consolation for his client with [1.4, 1.5].

Example E (Estimation).—The problem of giving a point estimate for a parameter, θ, is analyzed in classical statistics by the concept of estimators. *An estimator is a function defined over the set of possible outcomes of the contemplated random experiment; the value of the function at a particular outcome is to be the estimate of θ should that outcome obtain.* Each estimator, as a function of the possible outcome, is a random variable which derives its probability distribution for each θ, from the corresponding probability distribution over the possible outcomes. The statistician chooses his estimator in a given problem by requiring the chosen estimator (or its probability distribution) to satisfy one or more arbitrary conditions. Much of the classical theory of estimation used to be based on the "unbiased" condition. An estimator, $g(X)$, is unbiased if, and only if, $E[g(X)] = \theta$, for all θ, where θ is the parameter value of the probability distribution of X used to calculate the expected value. Figuratively, the probability distribution of $g(X)$ is always "centered" on the parameter to be estimated. Today, the justifications for using an unbiased estimator are not forceful nor often taken seriously; however, use of the condition persists among classical statisticians because the mathematics of finding an unbiased estimator which satisfies certain other conditions is often simple. Unbiased estimators which are also "functions of the existing sufficient statistic" and which have "minimum variance among such estimators" will be used in the example in the next paragraph. Those readers interested in the technical development of the estimators will find a general development in the Syllabus textbook ([8], pp. 221–22).

Suppose two actuaries wish to estimate the one-year survival rate, θ, following surgical treatment of a certain disease by observing a sequence of treated patients. Actuary A instructs the clinic to send him the follow-ups on the first n patients treated. Actuary B instructs the clinic to send him the follow-ups on all cases up through the yth death. Assuming independence and no "withdrawals," both actuaries are observing a Bernoulli sequence with parameter, θ, but each is observing a different random experiment. In A's experiment each outcome consists of n follow-ups, and the number of deaths is a random variable. In B's experiment each outcome contains exactly y deaths, the last follow-up being a death, and the number of follow-ups is a random variable. In statistical language, A and B are using different "stopping rules." If each of A and B decides to use the minimum-variance, unbiased, and sufficient estimator for his experiment, then A's estimator would be the observed survival rate and B's estimator would be the number of observed survivals divided by the num-

ber of follow-ups minus one. Thus, if the nth patient is also the yth death, A and B will receive the same n follow-ups, but A's estimate would be $(n - y)/n$ and B's estimate would be $(n - y)/(n - 1)$. Critics view this discrepancy between the estimates as a defect in the classical theory because it is not due to a difference in information received but rather to a difference in what information might have been received.

A familiar, alternative, and competing theory of estimation which does not exhibit the above discrepancy is maximum likelihood estimation. To examine the maximum likelihood estimates for this example, let X and Z denote the number of survivors in A's and B's experiments, respectively. Then, for $\theta = t$,

$$p_{X|\theta}(x \mid t) = \binom{n}{x} t^x (1-t)^{n-x} \qquad x = 0, 1, \ldots, n \quad (5)$$

and

$$P_{Z|\theta}(z \mid t) = \binom{z+y-1}{z} t^z (1-t)^y \qquad z = 0, 1, \ldots. \quad (6)$$

Let

$$L_A(t, x) = \binom{n}{x} t^x (1-t)^{n-x} \qquad 0 < t < 1,$$

and

$$L_B(t, z) = \binom{z+y-1}{z} t^z (1-t)^y \qquad 0 < t < 1,$$

where x and z are considered fixed. This duplication of notation for these two pairs of functions of two variables is solely to emphasize the fixed variable. The definition of A's (or B's) maximum likelihood estimate of θ, when x (or z) survivors obtain, is the value of t which maximizes $L_A(t, x)$ [or $L_B(t, z)$], the *likelihood function* for the observed data of the experiment. Now in the case when the nth patient is also the yth death,

$$L_A(t, n-y) = \binom{n}{n-y} t^{n-y}(1-t)^y$$

$$= \frac{n}{y}\binom{n-1}{n-y} t^{n-y}(1-t)^y = \frac{n}{y} L_B(t, n-y),$$

that is, when these two experiments produce the same data, their likelihood functions are proportional. It follows that, in this case, their maximum likelihood estimates are equal.

In general, the maximum likelihood estimate for θ will depend only on the relative shape of the likelihood function. This principle for the estimation problem may be generalized to all statistical inference by replacing "the maximum likelihood estimate for" in the preceding sentence with "any inference about." More precisely, this general principle, the *likelihood principle*, is: When two experiments, indexed by the parameter θ,

result in outcomes such that their respective likelihood functions are proportional, then they yield the same information about θ. Thus, those who adopt the principle work and talk in terms of the *likelihood*, the class of all proportional likelihood functions. The likelihood principle is a "theorem" in those theories of statistics which use Bayes' theorem, $p_{\theta|x}(t|x) \propto p_{x|\theta}(x|t)\, p_\theta(t)$, because two proportional likelihood functions will yield the same posterior distribution, $p_{\theta|x}(t|x)$. This includes those theories based on a personalistic definition of probability which provides a meaning for $p_\theta(t)$. Some other statisticians—R. A. Fisher [7], G. A. Barnard [2], and A. Birnbaum [4]—find the likelihood principle compelling for other reasons and thus are among the critics of the averaging procedures of classical statistics.

Bayesian statisticians disagree with classical statisticians on one more fundamental point. Classical methods treat experimental data as if they were isolated from other relevant experience of the experimenter. For example, if one wishes to estimate a probability of survival, classical statisticians consider it legitimate to use prior experience and information to form the opinion that a Bernoulli model is appropriate for the experiment but is illegitimate to use prior experience and information to form a prior opinion about the probability of survival. This is the dictum regardless of the size of the experiment relative to the experience and information. Bayesian statistics goes two steps beyond the classical dictum above, not only holding it legitimate to form an opinion based upon the prior experience and information, but mandatory to do so. The Bayesian statistician views the experimental data as evidence to be assimilated into the experience and knowledge of the experimenter.

IV. ANALYSIS BY BAYESIAN STATISTICS

Bayesian analysis of a statistics problem may usually be divided into four steps. This is not a universal pattern, but one that is applicable to many problems. Step 1: Formulate the prior distribution for the parameters of the problem. Step 2: Define the model for the experiment. Step 3: Perform the computations according to Bayes' theorem to determine the posterior distribution. Step 4: Analyze this posterior distribution in accordance with the objectives of the problem.

The prior distribution for the parameters forms a measure of the information and experience available prior to the anticipated data. It is interpreted by means of the personalistic definition of probability as discussed earlier. Thus, when one has formed his prior distribution, he is prepared, in principle, to enter wagers about events defined by the parameters. Conversely, by considering a family of such hypothetical

wagers, he can determine his prior distribution. More practically, by considering a sufficient number of such wagers, he can determine enough characteristics of his prior distribution to enable him to reach a satisfactory approximation to it. One common approximation is the usual analytical procedure of setting bounds for the prior distribution which are transformed by the analysis of Bayes' theorem to bounds for the posterior distribution. Another approximation is to adopt some member of some convenient family of probability distributions which reasonably closely fits the characteristics determined above.

The model for the experiment is the probability distribution of the data, given the parameter values. For Bayesian statisticians the problems that enter into these "model distributions" are personal problems, no different in kind from the problems to which prior distributions pertain. However, in usual statistical problems most observers will, as a first approximation, adopt the same distribution for the experimental model. This does not mean that there are no differences in opinion about the stochastic mechanism of the experiment but rather that the differences are negligible relative to other variations. The generality of personal probability may be seen in this context. One's personal probability for an event may be based upon a long-run frequency which is common to the experiences of his associates, and hence his personal probability will be "public" in the sense that it is held by everyone involved in the problem. Similarly, symmetry may make a personal probability public.

The posterior distribution is readily defined by Bayes' theorem and has the same personalistic interpretation as the prior distribution. Some experiments may be overwhelming in their effects on the divergence of prior opinions and thus the posterior distribution will be a "public" distribution. This is discussed in detail in Edwards, Lindman, and Savage [6].

Analysis of the posterior distribution is guided by the probability interpretation of the distribution. If a point estimate is desired for a parameter, the mean, median, or mode of the distribution would be appropriate. If an interval estimate is desired, it may be calculated directly from the posterior distribution. In general, the way one uses the posterior distributions is not unlike the use of a probability distribution in games of chance. (The reader can find information in [6], [9], [15], and [18].)

The rest of this section will give Bayesian analyses to Examples CI and E of the previous section.

Example CI (see above, Sec. III)

1. *Prior distribution.* This example will give an opportunity to illustrate the extent of the dependence of the posterior distribution on the

prior distribution. For this purpose let $p_\theta(t)$ be an arbitrary prior distribution for θ.

2. *Model for the experiment.* The experiment in this example is to obtain n independent observations on the population. Let $X = (X_1, X_2, \ldots, X_n)$ be the vector of independent observations and let $x = (x_1, x_2, \ldots, x_n)$ be a generic outcome. The density of probability for X at x, given $\theta = t$, is the product of n copies of the function graphed in Figure 1; if all the coordinates of x are between $t - 1$ and $t + 1$, the density is $(\frac{1}{2})^n$, and, if at least one coordinate is not between $t - 1$ and $t + 1$, the density is 0. Thus,

$$p_{X|\theta}(x|t) = (\tfrac{1}{2})^n \qquad t - 1 < x_i < t + 1 \qquad i = 1, 2, \ldots, n, \tag{7}$$
$$= 0 \qquad \text{elsewhere}.$$

3. *Posterior distribution.* Application of Bayes' theorem to the above components will yield the density function of the posterior distribution for θ, given the data $X = x$. In the calculation x is fixed at the observed outcome, and $p_{X|\theta}(x|t)$ enters as a function of t. The inequalities of (7) are more helpful if rewritten. The set of inequalities $t - 1 < x_i < t + 1$, $i = 1, 2, \ldots, n$ is equivalent to the set $x_i - 1 < t < x_i + 1$, $i = 1, 2, \ldots, n$, which is equivalent to the one inequality $R(x) - 1 < t < L(x) + 1$, where $L(x)$ and $R(x)$ denote the smallest and the largest observations among $x_1, x_2, \ldots, x_n$, respectively. Therefore

$$p_{X|\theta}(x|t)\, p_\theta(t) = p_\theta(t)(\tfrac{1}{2})^n \qquad R(x) - 1 < t < L(x) + 1 \tag{8}$$
$$= 0 \qquad \text{for other values of } t.$$

Bayes' theorem says that the conditional density function for θ, given $X = x$, is proportional to (8); therefore,

$$p_{\theta|X}(t|x) \propto p_\theta(t) \qquad R(x) - 1 < t < L(x) + 1$$
$$= 0 \qquad \text{for other } t \text{ values.}$$

In words, an individual's posterior density function is obtained by truncating his prior density function to the interval of values of θ that are possible given the data, and then normalizing it so its integral over this interval is unity. For the first set of observations given in the classical discussion of this example (see above, p. 45), $L(x) = 1.5$ and $R(x) = 3.0$, so $p_\theta(t)$ is truncated to the interval [2.0, 2.5]. If an individual's prior density is represented by the dotted line in Figure 2, then his posterior

density function would be proportional to the solid portion. For the second set of observations given, $L(x) = 1.4$ and $R(x) = 1.5$, so $p_\theta(t)$ is truncated to the interval [0.5, 2.4]. The effect of these data on the $p_\theta(t)$ of Figure 2 is shown in Figure 3.

4. *Analysis of the posterior distribution.* If the Bayesian statistician were now asked for an interval for θ upon which he would place $100\,p$ per cent confidence, he could choose some l and r such that

$$\int_r^l p_{\theta|x}(t \mid x)\,dt = p\,.$$

His interpretation of this interval is: "After seeing the data of the experiment, p is the probability, for me, that θ lies between l and r." Savage [17]

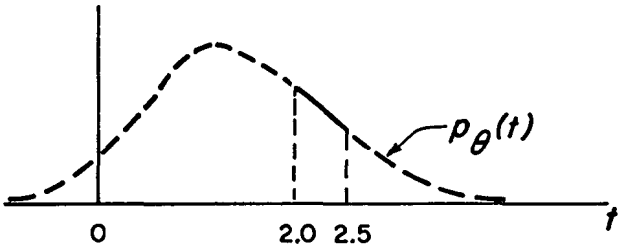

FIG. 2

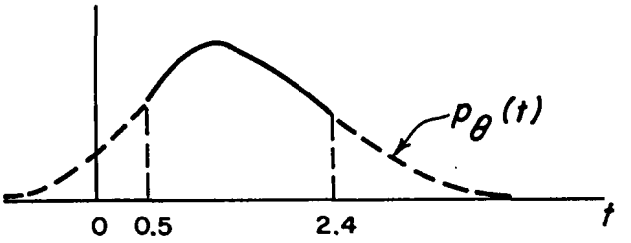

FIG. 3

calls this a credible interval to distinguish it from the classical confidence intervals.

The anomalous behavior of the confidence interval in this example is vivid. As the span of the n observations $R(x) - L(x)$ increases, the length of the confidence interval $R(x) - L(x)$ increases, but the length of the credible interval [for continuous $p_\theta(t)$] decreases as it should.

Example E (see above, Sec. III)

The conditional probability function for the number of survivals in a sequence of independent trials is perhaps one of the best examples of a

"public" probability (i.e., one agreed upon by all concerned). In this spirit let us assume that actuaries A and B will continue to use equations (5) and (6), respectively, for their probabilities, given the survival rate t. With that assumption, then the Bayesian analysis of actuaries A and B will coincide by virtue of the likelihood principle. Precisely, recall that, if the nth patient is the yth death, then

$$p_{X|\theta}(x \mid t) = \frac{n}{n-x}\, p_{Z|\theta}(x \mid t),$$

and the factor $n/(n-x)$ will be absorbed into the constant of proportionality when Bayes' theorem is applied. In other words, it will suffice to know only that the model distribution for the experiment is proportional to $t^x(1-t)^y$.

Step 1. Sometimes it is possible to choose a prior distribution from a family of so-called conjugate distributions which makes the calculation of the posterior distribution by Bayes' theorem very tractable. Precisely, a set of probability distributions is *conjugate* for a given experiment with the parameter θ; if, when the prior distribution $p_\theta(t)$ is in the set, then, for every outcome x of the experiment, the posterior distribution $p_{\theta|X}(t|x)$ is in the set. For this example a conjugate set is the set of all probability distributions with a density function of the form

$$\begin{aligned} f(t) &= K(p, q)\, t^p\, (1-t)^q \qquad 0 < t < 1 \\ &= 0 \qquad\qquad\qquad\quad \text{elsewhere}, \end{aligned} \tag{9}$$

where p and q are non-negative integers and $K(p, q)$ is the constant such that the integral of $f(t)$ is 1. If $p_\theta(t)$ is in this set, then

$$\begin{aligned} p_{\theta|X}(t|x) &\propto t^x\, (1-t)^y\, t^p\, (1-t)^q \qquad 0 < t < 1 \\ &= 0 \qquad\qquad\qquad\qquad\quad \text{elsewhere}, \end{aligned} \tag{10}$$

that is,

$$p_{\theta|X}(t|x) = K(x+p, y+q)\, t^{x+p}\, (1-t)^{y+q} \qquad 0 < t < 1,$$

and is in the set. The larger set of distributions with density functions of the form (9), for *any* numbers $p > -1$ and $q > -1$, is also conjugate for the experiment of this example and allows more choice for the prior distribution.

In formulating his prior distribution for the survival rate, the actuary will look to his experience, information, and opinions. As an illustration he may find his opinion about the survival rate for *non*surgically treated cases, say, θ', is firm; for example, suppose that, for him, $P(0.55 < \theta' <$

0.65) = 0.99. From a meeting with medical people he may learn that the surgical treatment cannot be detrimental to survival, but the treatment is new and its positive effects unknown. The actuary might summarize to himself thusly: "I'm virtually certain θ lies between 0.55 and 1.0, and I doubt that the treatment has no effect or that it's perfect—but it's even money between these extremes for me." A graph of his prior opinion might be like Figure 4. If the number of observed cases is large (e.g., 400) and if the sample survival rate falls within the actuary's "indifference interval,"

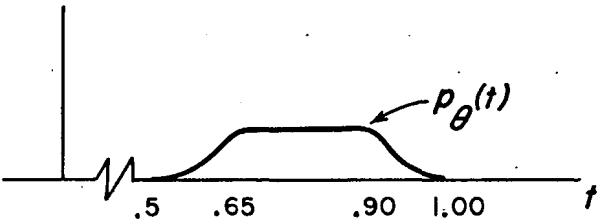

FIG. 4

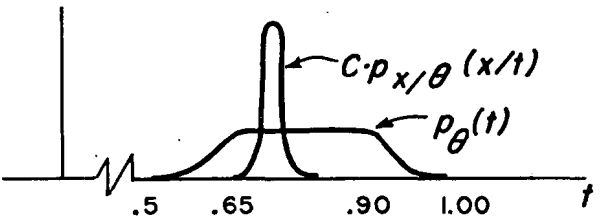

FIG. 5

his posterior distribution will not be unlike that derived by use of a uniform prior distribution (that member of the conjugate set with $p = q = 0$). This interaction of a sharp likelihood and a gentle prior distribution may be seen in Figure 5, where one can visualize the product of the two functions defining the graphs.

Bayesians conjecture that, when there is quantitative concurrence of opinion following any statistical analysis, it is due to this phenomenon of adequate data producing a sharp likelihood which washes out rather gentle differences in prior opinions. Sufficient conditions for a prior distribution to produce a posterior distribution reasonably close to the posterior produced by a uniform prior distribution are given in Edwards, Lindman, and Savage [6].

The actuary's experience, information, and opinion may not lead to the generally flat prior density of the situation above. If this surgical treat-

ment has been applied to cure a similar disease in the past, he may expect similar results to obtain in this application. If a smooth unimodal density function is appropriate for his prior distribution, then he may search in the conjugate set of distributions for one that "fits" his cumulative probabilities at three or four points. (The distributions of the conjugate set are beta distributions, which were tabled by Karl Pearson [14].) As an example, for $q = 7.5$ and $p = 28$,

t	.50	.60	.70	.80	.90	.95
$P(\theta < t)$	.000	.011	.142	.628	.985	1.000

and the density function has its mode at 0.789. As $p/(p + q)$ increases with q fixed, the distribution shifts to the right and becomes more skewed. As p increases with $p/(p + q)$ fixed, the distribution is more peaked and symmetric. Within this two-parameter family the actuary may find a distribution suitably close to his prior opinion.

Step 3. If the prior distribution was adequately described by a beta function, the posterior distribution is the one given in (10). If it were necessary to go outside the conjugate distributions to find a realistic prior, then this step would be accomplished by numerical procedures.

Step 4. The object of the experiment of this example was to estimate the survival rate. The actuary may choose the mean, median, or mode of his posterior density, whichever seems appropriate. In an economic situation with a loss function available, he would choose his estimate to minimize his expected loss (calculated with respect to the posterior distribution).

V. IMPLICATIONS FOR ACTUARIES

In summary I would liken statistical analysis to the black box of systems engineering (i.e., one feeds certain inputs into a black box in return for an output). The classical statistician has many black boxes, each bearing some of the labels "unbiased," "sufficient," "consistent," "efficient," "stringent," "uniformly most powerful," "asymptotically unbiased," ad infinitum. Each of his boxes has an intake hole at the top for the experimental data and an output hole at the right end. There is a warning on each box which says: "Use seriously only with k units or more of input." He would then offer you any of his boxes, counseling you that your choice should be based on the labels. The Bayesian has only one black box—but it has three holes: an additional intake hole on the left for experience, information, and prior opinion.

I think actuaries should be among the most ardent welcomers and users of the Bayesians' one black box. In all their work and training actuaries

emphasize that decisions are based on the data and judgment, but classical theories of statistics have never offered a place for, or required, a well-defined judgment factor. This has led actuaries to develop other methods of analysis, allowing input of judgment, in several areas of their work.

Credibility theory, including experience rating, is the actuaries' solution to a statistical estimation problem which was not amenable to classical methods. These rate-making problems are characterized by the existence of a small amount of data which has direct bearing on the rate class in question, a large amount of data for many similar rate classes combined, and the actuary's judgment. Arthur Bailey's search for a theory to explain the casualty actuaries' empirical methods of combining these sources of information for rate-making led him to Bayes' theorem and statistical procedures like those of the Bayesians. Although he was not a Bayesian statistician in the sense of expounding personalistic probability, his 1950 paper [1] is an excellent introduction to the application of Bayesian statistics to actuarial problems. Allen Mayerson recast much of Bailey's credibility work in a Bayesian statistical setting in a paper recently presented to the Casualty Actuarial Society [12].

The theory of graduation is a good example of special procedures developed by actuaries to solve a statistics problem that would not fit into the classical statisticians' mold where there is no room for judgment and experience. Graduation problems arise when the actuary wishes to estimate several values of an unknown function, say, $g(x)$, on the basis of observed data. The first step of classical procedures is to define a family of admissible functions indexed by a few parameters; for example, the two-parameter family of lines $g(x) = A + xB$. With the assumptions that no information, except the observed data, about the parameters is at hand and that any function outside the defined family is *not* admissible, then classical procedures define estimators of the parameters based only on the observed data. These two assumptions are not realistic for the actuary when he uses graduation theory. On the other hand, E. T. Whittaker's justification of difference-equation method of graduation as we discussed it in Section I is close to the actuary's state of mind. If we now compare Whittaker's justification with the four-step Bayesian analysis outlined in Section IV, we can see that Whittaker-Henderson graduation is a Bayesian solution to the graduation problem.

In addition to the rather poetic developments which may obtain when actuaries are converted to Bayesian statistics, there is an awkward one for the Society to face. We are in the early years of a ponderous change in statistical theory and practice. In the next five to ten years the diversity of training received by actuarial students preparing for the Society exam-

inations will increase considerably. These will be difficult years to give an "achievement examination . . . based on the material usually covered in undergraduate mathematics courses in probability and statistics" (Syllabus, Fall, 1963).

Professors J. C. Hickman, A. L. Mayerson, C. J. Nesbitt, and L. J. Savage each read at least one draft of this paper in detail and gave me many helpful comments. I would take this opportunity to thank them.

BIBLIOGRAPHY

[1] BAILEY, A. L. "Credibility Procedures," *Proceedings of the Casualty Actuarial Society*, XXXVII (1950), 7–23. (See also the discussion, p. 94.)

[2] BARNARD, G. A. "Statistical Inference," *Journal of the Royal Statistical Society*, Ser. B, XI (1949), 115–49.

[3] BAYES, THOMAS. "Essay toward Solving a Problem in the Doctrine of Chances," *The Philosophical Transactions*, LIII (1763), 370–418. Reprinted in *Biometrika*, XLV (1958), 293–315.

[4] BIRNBAUM, ALLAN. "On the Foundations of Statistical Inference," *Journal of the American Statistical Association*, LVII (1962), 269–306.

[5] CRAMÉR, HARALD. *The Elements of Probability Theory and Some of Its Applications*. New York: John Wiley & Sons, 1955.

[6] EDWARDS, W.; LINDMAN, H. R.; and SAVAGE, L. J. "Bayesian Statistical Inference for Psychological Research," *Psychological Review*, LXX (1963), 193–242.

[7] FISHER, R. A. *Statistical Methods for Research Workers*. Edinburgh and London: Oliver & Boyd, 1925. 12th ed., 1954.

[8] FRASER, D. A. S. *Statistics: An Introduction*. New York: John Wiley & Sons, 1958.

[9] GRAYSON, C. J., JR. *Decisions under Uncertainty: Drilling Decisions by Oil and Gas Operators*. Boston: Harvard University Press, 1960.

[10] HOEL, P. G. *Introduction to Mathematical Statistics*. 3d ed. New York: John Wiley & Sons, 1962.

[11] KING, G. Discussion on T. B. Sprague's paper, *Journal of Institute of Actuaries*, XXVI (1886), 113–15.

[12] MAYERSON, A. L. "A Bayesian View of Credibility," *Proceedings of the Casualty Actuarial Society*, Vol. LI (1964).

[13] MILLER, M. D. *Elements of Graduation*. Chicago: ASA and AIA, 1946.

[14] PEARSON, KARL. *Tables of the Incomplete Beta-Function*. Cambridge: Cambridge University Press, 1934.

[15] RAIFFA, HOWARD, and SCHLAIFER, ROBERT. *Applied Statistical Decision Theory*. Boston: Harvard University, Graduate School of Business Administration, Division of Research, 1961.

[16] SAVAGE, L. J. *The Foundations of Statistics.* New York: John Wiley & Sons, 1954.

[17] ———. *The Subjective Basis of Statistical Practice.* Ann Arbor: University of Michigan, 1961.

[18] SCHLAIFER, ROBERT. *Probability and Statistics for Business Decisions.* New York: McGraw-Hill Book Co., 1959. (*Introduction to Statistics for Business Decisions* [New York: McGraw-Hill Book Co., 1961] is an abridged version of [18].)

[19] WHITTAKER, E. T. "On Some Disputed Questions of Probability," *Transactions of the Faculty of Actuaries,* VIII (1920), 163–206.

[20] ———. "On a New Method of Graduation," *Proceedings of the Edinburgh Mathematical Society,* XLI (1923), 63–75.

[21] WHITTAKER, E. T., and LIDSTONE, G. J. Correspondence in *Transactions of the Faculty of Actuaries,* XI (1926–27), 233–37.

[22] WHITTAKER, E. T., and ROBINSON, G. *The Calculus of Observations.* 4th ed. London and Glasgow: Blackie & Son, Ltd., 1944.

[23] KYBURG, H. E., JR., and SMOKLER, H. E. *Studies in Subjective Probability.* New York: John Wiley & Sons, 1964.