TRANSACTIONS OF SOCIETY OF ACTUARIES
1967 VOL. 19 PT. 1 NO. 54

BAYESIAN GRADUATION

GEORGE S. KIMELDORF* AND DONALD A. JONES†

INTRODUCTION

A GRADUATION process is commonly justified on the basis of a statistical theory of random errors whereby each observed rate is the sum of the true rate and a random error ($U_x = V_x + e_x$). On the other hand, there are very few ideas common to the statistics section and the graduation section of the syllabus for the actuarial examinations. We believe that this isolation of the graduation problem from general statistical theory stems from the limited nature of statistical theories which are based upon a "relative frequency" concept of probability. In this paper we shall show how a graduation method may be developed as straightforward statistical estimation within a statistical theory based upon personal probability, that is, so-called Bayesian statistics.

Familiar examples of statistical estimation are the estimation of the mean, the standard deviation, or the coefficients of the regression line, given a specific set of observations. Graduation as statistical estimation differs from these examples in two important respects.

First, rather than estimating just one quantity or pair of quantities, graduation simultaneously estimates a large set of quantities such as a set of mortality rates for many different ages. The development of this multivariate estimation procedure requires certain more powerful techniques, such as the use of vectors, matrices, and the multivariate normal distribution, which are summarized in the appendixes to this paper.

Second, graduation differs from some other statistical estimation problems in its dependence upon information which is not contained in the observed data. Elphinstone, in discussing the logic of graduation, remarks,

* George S. Kimeldorf, not a member of the Society, is an assistant professor in the Department of Mathematics, California State College, Hayward, California. This paper is based on his dissertation [8] submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in The University of Michigan (1965) and prepared under the supervision of Donald A. Jones. This dissertation was written while Dr. Kimeldorf was an Actuarial Science Fellow at the University.

† The research of Donald A. Jones was supported in part by the National Science Foundation Grant No. GP-6008.

Computer time for this research was provided by the Computing Center at The University of Michigan.

"If we ignore relevant knowledge . . . , we get the wrong answer—we make a mistake" [3]. If the only information available to us were the observed data, then our estimate of the "true" rates, that is, the graduated rates, would equal the crude rates and we would not graduate at all.

One element of information which actuaries have been using for many years is the fact (or perhaps, belief) that the "true" rates form a smooth sequence. Thus, graduation has traditionally been associated with smoothing. But there are other properties of the "true" rates which should be included in "relevant knowledge."

For example, imagine the result of a graduation of mortality rates which yielded the smooth sequence

$$q_x = 1 - .001x \qquad \text{for} \qquad x = 10, 11, \ldots, 90 .$$

You would probably reject such a result because it contradicts your belief that mortality rates, except at the juvenile ages, increase with age. Similarly, suppose a graduation yielded a set of mortality rates double those found in a similar population just a short time previously. Again, you might doubt such results and look for an error in your calculation.

Our point is this: In order for the result of a graduation to be acceptable, it must be consistent not only with the ungraduated data and a concept of smoothness but also with other relevant knowledge possessed before observing the data. This totality of relevant knowledge, judgment, and belief possessed prior to observing the data is called "prior opinion" in Bayesian statistics.

A good graduation method, therefore, is one which makes maximal use, in some sense, of all aspects of prior opinion as well as the observed data. We shall not be able to present such an ideal method in this paper, but we shall present one approach to the graduation process which allows the use of the graduator's prior opinion in a more objective way than existing methods and yet involves only tractable calculations. While the method discussed is applicable to a broader range of graduation situations, we shall, when necessary, assume the more specific properties of the graduator's prior opinion regarding mortality rates.

In Section I the theoretical foundation for the Bayesian method of graduation is developed. The graduation problem is stated in the context of multivariate statistical estimation and analyzed according to Bayesian procedures. The core of our Bayesian graduation is shown to be the adoption of an appropriate prior covariance matrix. The prior covariance matrix is discussed in Section II, where we present necessary conditions for a matrix to be admissible as a prior covariance matrix in a graduation problem and examine some simple classes of admissible matrices. In Sec-

tion III we suggest procedures for selecting a prior covariance matrix from the class of admissible matrices and for selecting the other parameters of the graduator's prior distribution. Section IV contains an example of a Bayesian graduation and some Whittaker graduations of the same data. Finally, we conclude with remarks on some extensions of the Bayesian graduation method.

Much of the background material which is needed for a full understanding of this paper is contained in the appendixes. Appendix I contains an introduction to the algebra of vectors and matrices. In Appendix II this algebra is generalized to include a study of random vectors and random matrices and their application to probability. Appendix III defines the multivariate normal distribution and discusses some of its properties.

I. A BAYESIAN GRADUATION PROCESS

We will consider the following general graduation problem. We are given n observed rates $u_1, u_2, \ldots, u_n$, one for each of n values of an indexing variable such as age, duration, and so forth, and want to derive a set $v_1, v_2, \ldots, v_n$ of graduated rates.

We shall use the vector notation of Appendix I to denote the sequences of observed rates and graduated rates by the column vectors

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_n \end{bmatrix} \quad \text{and} \quad v = \begin{bmatrix} v_1 \\ v_2 \\ \cdot \\ \cdot \\ \cdot \\ v_n \end{bmatrix},$$

respectively. In order to construct the graduation process, that is, the means of deriving the vector v from the vector u, we shall have to introduce some random variables and vectors. In particular, let the random variable U_i be the observed rate for the ith of the n index values and let U be the column vector whose transpose is $[U_1, U_2, \ldots, U_n]$, or more succinctly $U = [U_1, U_2, \ldots, U_n]'$. Thus, U is a random vector of observed rates and u is a particular observation of U.

For $i = 1, 2, \ldots, n$, let W_i be the "true" rate which prevails for the ith index value. In classical statistics W_i is not a random variable at all but rather a fixed parameter whose value is unknown, while in Bayesian statistics W_i is a random variable. We define the random vector W of "true" rates by $W = [W_1, W_2, \ldots, W_n]'$.

In this context graduation is the estimation of the random vector W, given that the random vector U has value u. We will define v, the vector of graduated rates, as the value of this estimate. Having defined the

graduation problem as a statistical estimation problem, we can divide the Bayesian analysis into the four steps suggested by Jones [7]: formulation of the prior distribution, determination of a model for the experiment, derivation of the posterior distribution, and analysis of this distribution in accordance with the objectives of the problem.

The graduator's first step in this procedure is the formulation of his prior distribution for W, the vector of "true" rates. His prior opinion prescribes that certain sets of vectors are more probable than others, and, in theory, he can translate all the factors which constitute his prior opinion into a unique multivariate probability density function for the random vector W. The practical problem of determining the probability distribution which best expresses his prior opinion is a formidable one. To facilitate the graduator's selection of a prior distribution and to make ensuing calculations tractable, we shall limit his choice to the class of multivariate normal distributions. We believe that this class of distributions is sufficiently broad and robust to accommodate the graduator's prior opinion.

Once this restriction to the class of multivariate normal distributions has been accepted, the graduator's probability density function for the random vector W can be expressed, according to Appendix III, in the form

$$p_W(w) = k_1 \cdot \exp\left[-\tfrac{1}{2}(w - m)'A^{-1}(w - m)\right], \qquad (1)$$

where m, the mean vector, and A, the positive definite covariance matrix, are the parameters of the family and $k_1 = [(2\pi)^n |A|]^{-1/2}$. This first of the four steps of the Bayesian analysis is then completed by the graduator's assignment of values to these parameters, a process to be discussed in Section III.

The second phase of the Bayesian analysis is the determination of a model for the experiment, that is, adoption of a conditional probability distribution for U given a certain value for W. Often the "experiment" in graduation consists of obtaining rates from a certain sample by means of some process such as a mortality study. We are then asking: Given that the vector of true rates is equal to some fixed vector, say, $w = [w_1, w_2, \ldots, w_n]'$, what would be the graduator's conditional distribution of U, the vector of observed rates? In mortality studies and similar experiments most graduators assume that the U_i's are independent and that each U_i is "binomially" distributed with mean w_i, the "true" ith rate. As an approximation to such a multivariate conditional distribution of U, given $W = w$, we shall adopt a multivariate normal distribution with mean w and covariance matrix, say, B, for the graduator's model of the experiment. The independence of the U_i's, given $W = w$, is ex-

pressed by taking B to be a diagonal matrix. Thus, the model for the experiment prescribes the density function of U, given $W = w$, to be

$$p_{U|W}(u|w) = k_2 \exp\left[-\tfrac{1}{2}(u - w)'B^{-1}(u - w)\right], \tag{2}$$

where B is some positive definite diagonal matrix as yet undetermined and $k_2 = [(2\pi)^n |B|]^{-1/2}$.

In other applications of graduation where the observed variables are not rates, such as smoothing average weights against the independent variable height, the argument for an approximation by a normal distribution may or may not be convincing. For the cases where a normal distribution does not adequately describe the graduator's conditional distribution for the observed data, the specific method of this paper is not applicable.

The third phase of the Bayesian analysis is the computation of the graduator's posterior distribution by means of Bayes's theorem, that is, the derivation of the conditional density function for the "true" rates W given the observed rates U.

For continuous random variables, Bayes's theorem takes the form

$$p_{W|U}(w|u) = k_3(u)p_W(w)p_{U|W}(u|w), \tag{3}$$

where $k_3(u)$ is some function of u. If we view (3) as a function of w for some constant u and substitute into equation (3) the formulas for $p_W(w)$ and $p_{U|W}(u|w)$ as given by equations (1) and (2), respectively, we obtain

$$\begin{aligned} p_{W|U}(w|u) = k_4 \exp\{-\tfrac{1}{2}[&(w - m)'A^{-1}(w - m) \\ &+ (u - w)'B^{-1}(u - w)]\}, \end{aligned} \tag{4}$$

where $k_4 = k_1 \cdot k_2 \cdot k_3$ is a constant, that is, depends only on u. Using the results of matrix algebra contained in Appendix I, we can rewrite the right side of (4) as the density of a multivariate normal distribution as follows:

$$\begin{aligned} &k_4 \exp\{-\tfrac{1}{2}[w'A^{-1}w + w'B^{-1}w - w'B^{-1}u - w'A^{-1}m \\ &\quad - u'B^{-1}w - m'A^{-1}w]\} \exp\{-\tfrac{1}{2}[u'B^{-1}u + m'A^{-1}m]\} \\ &= k_5 \exp\{-\tfrac{1}{2}[w'(A^{-1} + B^{-1})w - w'(B^{-1}u + A^{-1}m) \\ &\quad - (u'B^{-1} + m'A^{-1})w]\}. \end{aligned} \tag{5}$$

Now let $v = (A^{-1} + B^{-1})^{-1}(B^{-1}u + A^{-1}m)$ and $C = (A^{-1} + B^{-1})^{-1}$ and "complete the square" in the exponent to write (5) as

$$k_6 \exp\{-\tfrac{1}{2}[w'C^{-1}w - w'C^{-1}v - v'C^{-1}w]\} \exp\{-\tfrac{1}{2}v'C^{-1}v\}$$
$$= k_6 \exp\{-\tfrac{1}{2}(w-v)'C^{-1}(w-v)\},$$

where k_6 is not a function of w.

Thus, the distribution of the vector W of "true" rates, given that the vector U of observed rates is equal to some fixed vector u, is a multivariate normal distribution with mean $v = (A^{-1} + B^{-1})^{-1}(B^{-1}u + A^{-1}m)$ and covariance matrix $C = (A^{-1} + B^{-1})^{-1}$.

Having obtained the posterior distribution, we are now in a position to perform the fourth step of the Bayesian analysis: the analysis of the posterior distribution in accordance with the objectives of the problem. If we wanted the "most probable" rates, as argued by G. King [9], we would take the mode of the posterior distribution. As usual in statistical theory good arguments could be put forth for taking the median or the mean of the graduator's posterior distribution as the graduated rates. In our present model, which uses normal distributions, the mean, median, and mode coincide. Hence, it is natural to define this common value, namely,

$$v = (A^{-1} + B^{-1})^{-1}(B^{-1}u + A^{-1}m) \tag{6}$$

to be the vector of graduated rates.

Equation (6) is a formula which defines the graduated rates v in terms of the ungraduated rates u; thus equation (6) defines a graduation method, which we shall call the Bayesian method. But before we can use this method, we must assign values to the n components m_i of m, the n positive elements b_{ii} of the diagonal matrix B, and the entries which determine the symmetric positive definite matrix A. The symmetry of A implies that A is determined by specifying the $n(n+1)/2$ elements on and above (or on and below) the diagonal. The assignment of values to these $n + n + (n)(n+1)/2 = (n^2 + 5n)/2$ parameters of the Bayesian method will be discussed below. Before doing so, however, we shall give some interpretations of equation (6) to provide insight into the roles of these parameters.

Equation (6) has a natural interpretation in terms of statistical estimation. Recalling the formula for the weighted average of two quantities, say, x, y with respective weights a, b to be $(xa + yb)/(a + b)$, we can analogously say that the vector v of graduated rates defined by formula (6) is a generalized weighted average of the vectors of prior means and observed rates, each weighted by the inverse of the appropriate covariance matrix. This weighted-average interpretation emphasizes the symmetric

roles of the prior means and the observed values in determining the graduated rates, thus reinforcing the concept of blending two inputs.

Since B is a diagonal matrix, its inverse is easily calculated; hence the solution for v by means of equation (6) involves two nontrivial matrix inversions. To derive a formula for v equivalent to equation (6) but simpler computationally, we can rewrite equation (6) in the form

$$\begin{aligned} v &= u + (A^{-1} + B^{-1})^{-1}[(B^{-1}u + A^{-1}m) - (A^{-1} + B^{-1})u], \\ v &= u + (A^{-1} + B^{-1})^{-1}[A^{-1}(m - u)], \\ v &= u + [(A^{-1} + B^{-1})^{-1}A^{-1}](m - u), \\ v &= u + [A(A^{-1} + B^{-1})]^{-1}(m - u), \\ v &= u + (I + AB^{-1})^{-1}(m - u), \end{aligned} \tag{7}$$

where I denotes the identity matrix of order n. Because formula (7) involves only one nontrivial matrix inversion, the solution for v by means of (7) is simpler computationally than by use of (6). Formula (7) also possesses an interesting interpretation in that it expresses the traditional view of graduation as a modification of observed data, whereby to the observed vector u is added the adjusted difference vector $(I + AB^{-1})^{-1}(m - u)$ to yield the graduated vector v.

A third way of writing formula (6) naturally suggests itself. If we follow the procedure used in deriving formula (7) from (6) but interchange the roles of A and B and of u and m, we derive the formula

$$v = m + (I + BA^{-1})^{-1}(u - m). \tag{8}$$

While not as practical for computational purposes, equation (8) stresses the Bayesian view of graduation as a systematic revision of the graduator's opinion in the light of new data. Prior to observing the data his estimate of the "true" rates is m, the mean of his prior distribution, while, after viewing the observed rates, his opinion is modified by the quantity $(I + BA^{-1})^{-1}(u - m)$.

Let us now compare the general Bayesian method presented herein with the difference-equation method originated by Whittaker [14, p. 303] which could be considered a first Bayesian approach to graduation. We noted previously that the vector v as defined by formula (6), (7), or (8) is, at the same time, the mean, median, and mode of the posterior distribution. The definition of v as the mode implies that v is the unique value which maximizes the posterior density given by (4), or equivalently, which minimizes the quadratic form

$$(u - w)'B^{-1}(u - w) + (w - m)'A^{-1}(w - m). \tag{9}$$

If we denote the (i, j)th element of A^{-1} by z_{ij} so that $A^{-1} = [z_{ij}]$ and denote b_{ii}^{-1} by e_i, equation (9) takes the form

$$\sum_{i=1}^{n} e_i(u_i - w_i)^2 + \sum_{j=1}^{n}\sum_{i=1}^{n}(w_i - m_i)z_{ij}(w_j - m_j). \qquad (10)$$

Now, in the mixed difference type B method of graduation the graduated rates are those which minimize the quadratic form

$$\sum_{i=1}^{n} e_i(w_i - u_i)^2 + \left[h_1\sum_{i=1}^{n-1}(\Delta w_i)^2 + h_2\sum_{i=1}^{n-2}(\Delta^2 w_i)^2 + \ldots + h_k\sum_{i=1}^{n-k}(\Delta^k w_i)^2\right], \qquad (11)$$

where e_i is the weight ascribed to the ith observed value and h_j expresses the emphasis to be placed on the measure of roughness[1] defined by jth differences.

The striking similarity between expressions (10) and (11) is further justification for considering Whittaker's method a special case of the more general Bayesian method. The first summation of (10)

$$\sum_{i=1}^{n} e_i(w_i - u_i)^2 \qquad (12)$$

is identical in appearance to the first summation, the measure of departure, of the corresponding Whittaker formula (11). The only difference between the two is perhaps in the assignment of values to the set of e_i's. Of course, no general statement can be made about this assignment as it depends on the nature of the data; however, we can make the following comparison for the case of mortality data.

In the Bayesian formula, each $(e_i)^{-1}$ is the variance of the graduator's conditional distribution for the observed rate U_i, given the true rates $W_1 = w_1$, $W_2 = w_2$, . . . , $W_n = w_n$. For a mortality study we think of U_i as Θ_i/E_i, where Θ_i, the (random) number of deaths, and E_i, the exposure, are measured in lives or amounts (here E_i is capitalized to conform with the usual exposure notation of mortality studies and not to conform with our use of capital letters for random variables). If the unit of measurement is lives and if there is no migration then, for most graduators, Θ_i would have a binomial distribution with parameters w_i and E_i. Hence the variance of Θ_i would be $E_i w_i(1 - w_i)$, and the variance

[1] Since $\Sigma e_i(q_i - q_i'')^2$ and $\Sigma(\Delta^k q)^2$ *decrease* with *increases* in fit and smoothness, respectively, we prefer to call such expressions "measures of departure" and "measures of roughness," as suggested to us by T. N. E. Greville.

of $U_i = \Theta_i/E_i$ would be $w_i(1 - w_i)/E_i$. If the unit of measurement is amounts and if there is no migration, then the graduator's variance of $U_i = \Theta_i/E_i$ would be

$$\sum_{j=1}^{n_i} s_{ij}^2 w_i(1 - w_i) \Big/ \Big(\sum_j s_{ij}\Big)^2 = w_i(1 - w_i) \sum_1^{n_i} s_{ij}^2/E_i^2, \quad (13)$$

where s_{ij} is the amount on the jth life of the n_i insured lives in the ith age group. Since the normal distribution model which we have adopted requires that the variance of the graduator's conditional distribution (13) be free of w_i, we suggest using $m_i(1 - m_i)/E_i$ or $m_i(1 - m_i)/(E_i/\bar{s}_i)$, where m_i is the mean of the graduator's prior distribution for W_i—to be discussed in Section III—and $\bar{s}_i$ is an approximation to

$$\sum_j s_{ij}^2 \Big/ \Big(\sum_j s_{ij}\Big) .$$

For other discussions of work requiring similar approximations, see D. Cody [2] and I. Rosenthal [11]. We shall not suggest corrections for the case of nonnegligible migration.

In Whittaker's method the e_i's have traditionally been taken equal to 1 (in the type A graduation) or equal to the exposure E_i's (in the type B graduation). In 1955, Camp [1] suggested assigning to e_i the value

$$\frac{E_i}{v_i^*(1 - v_i^*)},$$

where v_i^* is the ith rate obtained from a preliminary type A graduation. For a history of similar suggestions made in the context of other graduation methods see reference [15], p. 96.

The significant difference between expressions (10) and (11) lies in their second summations,

$$\sum_{j=1}^{n} \sum_{i=1}^{n} (w_i - m_i) z_{ij} (w_j - m_j) \quad (14)$$

and

$$\sum_{j=1}^{k} \sum_{i=1}^{n-j} h_j (\Delta^j w_i)^2, \quad (15)$$

respectively, both of which are quadratic functions of the w_i.

In the following section we shall investigate various covariance matrices A leading to corresponding quadratic forms (14). It will be shown that (14) has a meaningful interpretation as a measure of the degree of disconformity of w with the graduator's prior opinion in much the same

sense as (15) is used as his measure of roughness in the Whittaker formula. We shall consider, in particular, a class of covariance matrices whose inverses generate quadratic forms similar to those of Whittaker's method.

II. THE PRIOR COVARIANCE MATRIX

Section I presented the theoretical foundations for a statistical method of graduation. The vector v of graduated rates was defined as $v = (A^{-1} + B^{-1})^{-1}(B^{-1}u + A^{-1}m)$, where u is the vector of observed rates. The elements of the vector m and of the matrices A and B serve as parameters of the graduation. As in other graduation methods the graduator must assign values to the parameters in order to derive a set of graduated rates. But, unlike other graduation methods, Bayesian graduation has been placed within the framework of a statistical theory which guides the selection of the parameter values to the extent that the selection of any statistical procedure is guided. The interpretation of B as the covariance matrix of the conditional distribution of the vector U of observed rates given the vector W of true rates guided the selection of values for the n nonzero elements of the diagonal matrix B. In this section we shall study the graduator's prior distribution of W in order to get insight into a procedure for selecting values for the elements of the covariance matrix A of his prior normal distribution for W.

The fundamental basis for graduation is the existence of a strong relationship among the rates. R. Henderson thus requires that data suitable for graduation must constitute a "connected series in which each . . . bears a special relation to the groups immediately preceding or following it" [5, p. 2]. M. D. W. Elphinstone [3, p. 18] expresses this same idea as follows:

> Unless we postulate that there are relations between neighboring rates, we are wrong to graduate—wrong in the sense that we make a mistake, for the crude rates are the only right answer. The theory of graduation is then the theory of relations between neighboring rates, and it is in the power we have to choose between different relations that our minds have room to differ, to give effect to individual judgments.

The corresponding statement in the context of the Bayesian statistical view of graduation is the assertion of a prior opinion in which the random variables $W_1, W_2, \ldots, W_n$ are not independent or—what is equivalent for normally distributed random variables—are not uncorrelated. Thus, the "special relation" to which Henderson refers and the "relations between neighboring rates" to which Elphinstone refers are summarized by the correlation coefficients in the graduator's prior distribution for W, called his prior correlations.

Intuitively, the correlation coefficient c_{ij} between the normally dis-

tributed random variables W_i and W_j has two significant properties which make it an important concept in graduation. First, the correlation is a measure of the degree and "direction" of stochastic dependence between two normally distributed random variables, a positive correlation meaning that large values of one random variable are generally associated with large values of the second random variable. For example, when we say that our prior opinion prescribes a positive and large (i.e., only slightly less than 1) correlation coefficient between the true mortality rates (as distinguished from the observed mortality rates) at ages 30 and 31, we mean that for any population in which the true mortality rate prevailing at age 30 is large we should expect the true mortality rate at age 31 also to be large, and conversely.

Second, the correlation coefficient—or, more precisely, the absolute value of the correlation coefficient—is a measure of the extent to which knowledge of one normally distributed random variable improves the ability to estimate the second. In particular, if c_{ij} is the correlation coefficient between normal random variables W_i and W_j and the (marginal) standard deviation of W_j is p, then the standard deviation of the conditional distribution of W_j given W_i is $p\sqrt{1 - c_{ij}^2}$ (see Appendix III). For example, on the basis of existing knowledge and belief, a graduator is able to make an educated guess about the true mortality rate prevailing at age 30 in a certain population. If, however, he actually knew the true rate prevailing at age 31, he could predict the true rate at age 30 with more certainty.

These two intuitive properties of the correlation coefficient—stochastic dependence and reduction of uncertainty—will be made more precise in the next section when we discuss some procedures for determining the prior correlation coefficients defined by a person's prior opinion.

A functional relationship exists between the elements of the prior covariance matrix A and the prior correlation coefficients. The correlation c_{ij} between the random variables W_i and W_j is defined by

$$c_{ij} = c_{ji} = \frac{\operatorname{cov}(W_i, W_j)}{\sigma_i \sigma_j},$$

where σ_i is the standard deviation of W_i. Since the covariance matrix $A = [a_{ij}]$ is defined by $a_{ij} = \operatorname{cov}(W_i, W_j)$ for $i \neq j$ and $a_{ii} = \sigma_i^2$, we have the formula

$$c_{ij} = \frac{a_{ij}}{\sqrt{a_{ii} \cdot a_{jj}}}, \tag{16}$$

which relates the prior correlation coefficients with the elements of the matrix A.

An important goal of this section and the next is the development of a procedure whereby a graduator, on the basis of his own prior opinion, can select a matrix A to use in the graduation procedure. A first step in this direction would be to abstract certain features common to every graduator's prior opinion and then to restrict our attention to classes of matrices which at least reflect these common features.

First, any prior opinion in a graduation problem would prescribe that nearby rates, that is, rates for nearby values of the independent variable, are more highly correlated than are more distant rates. Stated more precisely, if j is between i and k in the sense that $i \leq j \leq k$ or $i \geq j \geq k$, then $c_{ij} \geq c_{ik}$. Using formula (16), we can restate this condition in terms of the elements of the matrix A as follows:

$$i \leq j \leq k \quad \text{or} \quad i \geq j \geq k \quad \text{implies} \quad a_{ij}/\sqrt{a_{jj}} \geq a_{ik}/\sqrt{a_{kk}}\,. \tag{17}$$

Second, we would probably agree that in most graduation problems we would want c_{ij} to be positive (or at least nonnegative) for all i and j. By formula (16) it is clear that the condition $c_{ij} \geq 0$ for all i and j is equivalent to the condition

$$a_{ij} \geq 0 \qquad \text{for all } i \text{ and } j \tag{18}$$

for the entries a_{ij} of the matrix A.

Properties (17) and (18) are two conditions which a matrix $A = [a_{ij}]$ must satisfy in order to be admissible as a covariance matrix of one's prior distribution of the "true" rates in a graduation problem. In addition, any covariance matrix A has the property:

$$A \text{ is symmetric} \tag{19}$$

and, by the definition of the multivariate normal distribution,

$$A \text{ is positive definite}\,. \tag{20}$$

Hereafter, any nth order matrix (where n is the number of values of the independent variable in our graduation problem) which has all of the properties (17), (18), (19), and (20) will be called *admissible*, and we shall restrict our attention to admissible matrices.

Let us now study a simple class of admissible matrices which possess certain interesting properties. Consider the class, which we shall call $\mathcal{A}_1$, of all nth order matrices of the form $A = [a_{ij}]$, where $a_{ij} = p^2 r^{|i-j|}$ for $p > 0$ and $0 \leq r < 1$. The general matrix in class $\mathcal{A}_1$ for, say, $n = 4$, is

$$A = \begin{bmatrix} p^2 & p^2 r & p^2 r^2 & p^2 r^3 \\ p^2 r & p^2 & p^2 r & p^2 r^2 \\ p^2 r^2 & p^2 r & p^2 & p^2 r \\ p^2 r^3 & p^2 r^2 & p^2 r & p^2 \end{bmatrix}.$$

The numbers p and r serve as parameters which index the class $\mathcal{A}_1$ in the sense that each pair of p and r determines one, and only one, matrix of the class. It follows from formula (16) that if a graduator selects a matrix $A = [a_{ij}] = [p^2 r^{|i-j|}]$ as a covariance matrix for his prior distribution of W, then his prior correlation $c_{i,i+1}$ between the rate W_i and its immediate neighbor W_{i+1} is r, while his prior standard deviation of any rate is p. Thus, the two parameters p and r have meaning in terms of familiar statistical concepts to guide their evaluation.

We now prove that every matrix in class $\mathcal{A}_1$ is admissible. Let $A = [a_{ij}] = [p^2 r^{|i-j|}]$ be any matrix in class $\mathcal{A}_1$. To verify condition (17), it is sufficient to note that if j is between i and k, then $|i-j| \leq |i-k|$; hence $r^{|i-j|} \geq r^{|i-k|}$, since $0 \leq r < 1$. It is obvious that $a_{ij} \geq 0$ for all i and j, since r is nonnegative. A is symmetric since $a_{ij} = p^2 r^{|i-j|} = p^2 r^{|j-i|} = a_{ji}$. To prove that A is positive definite, it is sufficient to prove that A^{-1} is positive definite. Consider the matrix $Z = [z_{ij}]$, where

$$p^2(1-r^2)z_{ij} = \begin{cases} 1 & \text{for} \quad i = j = 1 \\ 1 + r^2 & \text{for} \quad 1 < i = j < n \\ 1 & \text{for} \quad i = j = n \\ -r & \text{for} \quad |i-j| = 1 \\ 0 & \text{otherwise;} \end{cases}$$

that is, the elements on the principal diagonal of Z are $(1 + r^2)/p^2(1 - r^2)$ except for the elements in the upper-left and lower-right corner, which are $1/p^2(1 - r^2)$; the elements on the diagonals immediately above and below the principal diagonal are all $-r/p^2(1 - r^2)$; and all the other elements of Z are zero. If we multiply Z by A, we get I_n, the nth order identity matrix; hence $Z = A^{-1}$. Now, if $y = [y_1, y_2, \ldots, y_n]'$ is any vector, then $y'Zy =$

$$\sum_{j=1}^{n} \sum_{i=1}^{n} y_i z_{ij} y_j$$

$$= \frac{1}{p^2(1-r^2)} \left[y_1^2 + y_n^2 - 2r \sum_{i=1}^{n-1} y_i y_{i+1} + (1 + r^2) \sum_{i=2}^{n-1} y_i^2 \right] \quad (21)$$

$$= \frac{1}{p^2} \left[\frac{1}{1-r^2} \sum_{i=1}^{n-1} (y_i - r y_{i+1})^2 + y_n^2 \right].$$

But (21), a sum of squares, is positive for all $y \neq 0$. Therefore Z and hence A are positive definite.

We showed previously that the vector v of graduated rates defined by

the Bayesian method is that particular w which minimizes the quadratic form

$$(u - w)'B^{-1}(u - w) + (w - m)'A^{-1}(w - m)\,. \tag{22}$$

The first term of (22) was shown to correspond to the measure of departure of Whittaker's method. In order to gain some insight into the significance of adopting a matrix A in the class a_1 as an approximation to one's true prior covariance matrix, let us now examine the second quadratic form,

$$(w - m)'A^{-1}(w - m)\,, \tag{23}$$

as a generalization of the roughness term of Whittaker's method.

According to formula (21), quadratic form (23) is

$$\frac{1}{p^2(1-r^2)}\Big[(w_1-m_1)^2+(w_n-m_n)^2 - 2r\sum_{i=1}^{n-1}(w_i-m_i)(w_{i+1}-m_{i+1})+(1+r^2)\sum_{i=2}^{n-1}(w_i-m_i)^2\Big],$$

and this can be written in the form

$$\frac{1}{p^2}\Big\{\frac{r}{1-r^2}\sum_{i=1}^{n-1}[\Delta(w_i-m_i)]^2+\frac{1-r}{1+r}\sum_{i=1}^{n}(w_i-m_i)^2 + \frac{r}{1+r}[(w_1-m_1)^2+(w_n-m_n)^2]\Big\}\,. \tag{24}$$

The first summation of (24),

$$\frac{r}{p^2(1-r^2)}\sum_{i=1}^{n-1}[\Delta(w_i-m_i)]^2 \tag{25}$$

is never negative; and is zero for $r \neq 0$ if, and only if, there exists a constant c such that $w_i = m_i + c$; that is, if, and only if, the distance between the graphs of the sequences m_i and w_i is constant. Thus (25) can be interpreted as a measure of the disparity between the shapes in the graphs of the w_i and the graduator's prior means m_i. The remainder

$$\frac{1-r}{p^2(1+r)}\sum_{i=1}^{n}(w_i-m_i)^2+\frac{r}{p^2(1+r)}[(w_1-m_1)^2+(w_n-m_n)^2] \tag{26}$$

of (24) is a weighted sum of squares of the deviations of w_i from m_i and can be interpreted as a measure of the departure of w from the vector m of prior means. Thus expression (24), which is the sum of (25) and (26), serves as a measure of the disconformity of a vector w to the graduator's

prior opinion, the measure consisting of two components: a measure of the incongruity of w and m and a measure of the departure of w from m. Hence the graduated rates minimize a sum of a measure of the departure of the rates from the observed rates and a measure of the disconformity of the rates with prior opinion. This is a direct generalization of Whittaker's method, in which the only measure of disconformity with prior opinion is roughness.

As a further illustration of the interpretation, let us regard (24) as a function of r, the prior correlation coefficient between neighboring rates. If, as is often the case, the graduator's prior opinion about the level of the rates is vague, although his opinion with regard to the "shape" of the rates is relatively strong, then his prior standard deviation p for W_i may be large, although his conditional standard deviation $p\sqrt{1-r^2}$ for W_i, given W_{i+1}, would be small. Hence his prior opinion would dictate a value of r close to 1. As r approaches 1, expression (25) increases relative to (26), thus reflecting this emphasis on the shape rather than the general level of the rates in the graduator's prior opinion.

Conversely, for r small, the measure (26) of the departure rather than the measure of incongruity dominates (24), the total measure of disconformity with the prior opinion. If r were actually zero corresponding to independence of the rates, then (25) would vanish and the ith graduated rate v_i would merely be a weighted average of the ith prior mean m_i and the ith observed rate u_i.

If the prior means m_i were equal (which they normally would not be) and if r were increased toward 1 while p increases without bound in such a manner that $r/p^2(1-r^2)$ approached some positive constant, say, h, then expression (24), the expansion of $(w-m)'A^{-1}(w-m)$ would approach

$$h\sum_{i=1}^{n-1}(\Delta w_i)^2 .$$

Hence, expression (22), the quantity minimized by the graduated rates in the Bayesian method, would approach

$$\sum_{i=1}^{n} e_i(u_i-w_i)^2+h\sum_{i=1}^{n-1}(\Delta w_i)^2 ,$$

which is the quadratic form minimized by the graduated rates in a type B Whittaker graduation which uses a first-difference measure of roughness. If, in addition, the set of e_i for both methods was assigned the same set of values, then the two methods would define the same set of graduated

rates. Thus, class a_1 produces a graduation method which is a generalization of a type B first-difference Whittaker method.

By restricting his search for a covariance matrix to matrices belonging to class a_1, a graduator achieves the economy of having to assign values to only two parameters, r and p, but perhaps sacrifices fidelity to his true prior opinion, for it is possible that no matrix in this class is an adequate approximation to his true prior covariance matrix for W. Matrices of class a_1 would be appropriate only if he felt that the correlation coefficient between any rate W_i and its immediate successor W_{i+1} were the same for all i. While this condition may be a good approximation in the case of equally spaced data, it might not be appropriate for unequally spaced data. For example, suppose it were required to graduate a vector u of 5 observed mortality rates, say $u = [q_{10}, q_{20}, q_{25}, q_{30}, q_{50}]'$. A graduator would then define random variables W_1, W_2, W_3, W_4, W_5 to be the true mortality rates at ages 10, 20, 25, 30, and 50, respectively. In his prior distribution for the random vector $W = [W_1, \ldots, W_5]'$, it would be unrealistic to have the correlation between W_2 and W_3, the true rates at ages 20 and 25, equal to the correlation between W_4 and W_5, the true mortality rates at ages 30 and 50.

For graduation of unequally spaced data or any data in which the sequence of prior correlation coefficients between W_i and W_{i+1} is not constant for different values of i, a matrix from the class a_1 would not be appropriate for expressing prior opinion. Let us therefore consider a larger class a_2 of nth order matrices of the form $A = [a_{ij}]$ defined by the rule

$$a_{ij} = \begin{cases} p^2 \prod_{k=i}^{j-1} r_k & \text{for} \quad i < j \\ p^2 & \text{for} \quad i = j \\ p^2 \prod_{k=j}^{i-1} r_k & \text{for} \quad i > j \,, \end{cases} \tag{27}$$

where p is any positive number and $r_1, r_2, \ldots, r_{n-1}$ are each nonnegative numbers less than 1. For example, the general matrix in class a_2 for $n = 4$ is

$$A = \begin{bmatrix} p^2 & p^2 r_1 & p^2 r_1 r_2 & p^2 r_1 r_2 r_3 \\ p^2 r_1 & p^2 & p^2 r_2 & p^2 r_2 r_3 \\ p^2 r_1 r_2 & p^2 r_2 & p^2 & p^2 r_3 \\ p^2 r_1 r_2 r_3 & p^2 r_2 r_3 & p^2 r_3 & p^2 \end{bmatrix}.$$

Clearly, any matrix which is a member of class a_1 is also a member of class a_2.

If a matrix A as defined by (27) is adopted as the covariance matrix in a prior distribution of W, then by formula (16), the prior standard deviation of W_i is p for all i, and the correlation coefficient between W_i and W_{i+1} is r_i. Hence an adoption of the common standard deviation of each of the rates and the correlation between every pair of adjacent rates in the prior distribution determine a unique matrix from the class a_2. It can be proved that every matrix in class a_2 is admissible. We omit the proof but instead refer the interested reader to reference [8].

All the elements a_{ii} on the principal diagonal of any matrix A in class a_1 or a_2 are equal to p^2. Therefore, if A is the graduator's prior covariance matrix for a random vector W, his prior standard deviations of all the W_i are equal to p. But this may not be a good approximation to his true prior opinion. For example, for him it may be more probable that q_{30} lies in a certain interval centered at m_{30} than that q_{70} lies in an interval of equal length centered at m_{70}. Hence his prior standard deviation for the true rate at age 30 would be smaller than his prior standard deviation for the true rate at age 70, and no matrix in class a_2 would be an adequate approximation to his covariance matrix.

Fortunately, we can further enlarge our classes of admissible matrices to include some matrices with varying prior standard deviations. We omit the proof of the following theorem.

Theorem.—If an nth order matrix of the form $A = [c_{ij}]$ is admissible and if $p_1, p_2, \ldots, p_n$ are any set of positive constants, then the matrix $A^* = [p_i p_j c_{ij}]$ is also admissible.

The import of this theorem is the ability to construct an admissible matrix from a matrix of correlation coefficients and a set of standard deviations. In particular, a graduator may be able to find a matrix among those in the class a_1 with $p = 1$ which expresses his prior correlation coefficients adequately. Next, he may determine the set of n standard deviations $p_1, p_2, \ldots, p_n$ of his prior distribution. Then, by (16), his covariance matrix is $A = [a_{ij}] = [p_i p_j r^{|i-j|}]$, which is admissible by the theorem. We shall denote by a the class of all matrices generated in this manner.

III. ELICITING THE PRIOR DISTRIBUTION

Section I presented the theoretical foundations for a graduation method in which the graduated rates depend not only on the observed data but also on prior information available to the graduator. This prior information or "prior opinion" is represented by the graduator's prior probability distribution for the random vector W of true rates. Because of the adoption of a multivariate normal model, his prior distribution can be uniquely specified by the mean vector m and the covariance matrix A. Section II

presented various classes of matrices from which a graduator might choose an approximation to his covariance matrix. In this section we discuss procedures by which a graduator can elicit (approximately) m and A from his prior opinion.

The Vector m of Prior Means

The ith element m_i of m is the mean of the graduator's prior distribution for the ith true rate W_i. In a normal distribution, the mean is also the median and the mode; hence m_i is the value which has probability .5 above it and .5 below it and which is the "most probable" value in the graduator's opinion for the true rate W_i.

Suppose there were no observed data at all. Then the prior and posterior distributions would be identical, and m_i would be the ith "graduated" rate. Therefore, m_i can be further characterized as being that value which, based on his prior information, the graduator would use for the ith rate if there were no observed data.

The Diagonal Entries of A

The graduator's prior distribution of the ith rate W_i by itself (i.e., his prior marginal distribution of W_i) is a univariate normal distribution with mean m_i and variance $\sigma_i^2 = a_{ii}$. Having elicited m_i, the graduator can determine $\sigma_i^2 = a_{ii}$ by specifying the length of the interval symmetric about m_i, in which a given fraction α of his probability lies. Then, after determining by introspection the value of b for which

$$Pr\{m_i - b < W_i < m_i + b\} = \alpha\,, \tag{28}$$

he can use a table of the standardized normal distribution to solve for σ_i. In general, if he chooses a different α, determines the corresponding b which satisfies equation (28), and again solves for σ_i, he would obtain the same result only if his prior distribution were precisely a normal distribution. We therefore suggest that in practice the results for several pairs of α and b be averaged.

As an example, let us suppose that for $\alpha = .50$, $\alpha = .90$, and $\alpha = .999$, the graduator's values of b which satisfy (28) are $b = .0007$, $b = .0015$, and $b = .0025$, respectively. According to the table for the normal distribution,

$$Pr\{\mu - .674\sigma < X < \mu + .674\sigma\} = .5\,;$$

hence $.674\sigma_i = .0007$. Similarly, $1.645\sigma_i = .0015$, and $2.574\sigma_i = .0025$. The average of the three solutions for σ_i is $\sigma_i = .00097$.

A calculation in this detail for each σ_i in the graduation of a large table would probably not be justified. In such a case, the graduator might

adopt a functional form for σ_i; for example, for all i, $\sigma_i = c$, $\sigma_i = c\sqrt{m_i}$, or $\sigma_i = c\sqrt{m_i(1 - m_i)}$. Probability statements such as (28) may then be made for several values of i and α and solved for c. These solutions would then be averaged.

The Remaining Elements of A

Having determined each m_i and a_{ii}, the graduator has specified his prior marginal distribution for each W_i. But, in order to specify completely his prior joint distribution for the random vector W, he must, in addition, specify how the W_i's are stochastically related; that is, he must specify the set of covariances a_{ij} between each pair W_i and W_j for $i \neq j$. Rather than work with covariances directly, it is often simpler to work with the set of correlations c_{ij} and then to compute the covariances by the relation $a_{ij} = c_{ij}\sqrt{a_{ii}a_{jj}}$. Thus the determination of the graduator's covariance matrix is reduced to the selection of an appropriate set of correlation coefficients.

To aid the graduator in eliciting his correlations, we choose to interpret them as his expression of smoothness. A graduator has opinions regarding the value of the true rate at each value of the independent variable, which are summarized by his prior marginal distribution for each rate. Thus for each i his prior density for W_i will be distributed along a line segment at i as in Figure A. If he were given the *true* values of the neighboring rates

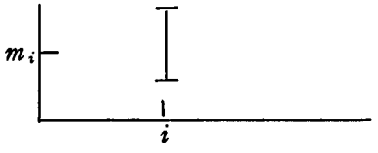

FIG. A

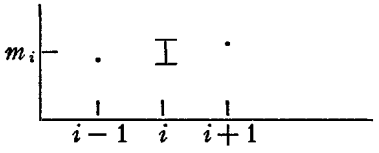

FIG. B

at $i - 1$ and at $i + 1$, then he would sharply contract the dispersion of his opinion for the value of the unknown rate at i, as indicated in Figure B. In concepts of traditional graduation this contraction is an expression of his requirement that the rates W_{i-1}, W_i, and W_{i+1} be part of a smooth sequence. In the concepts of Bayesian statistics it is the characteristic of W_i's being highly correlated with W_{i-1} and W_{i+1}, that is, the greater the contraction the greater is the correlation.

Now we want a procedure which translates the magnitude of this contraction of his prior distribution to his conditional distribution into values (or equations) for his correlations. In general, the contraction may be measured by using equation (28) to compare the lengths of intervals sym-

metric about his prior mean m_i and containing prior probability α, that is, $2t(\alpha)$ in (29),

$$Pr\{m_i - t(\alpha) < W_i < m_i + t(\alpha)\} = \alpha\,, \tag{29}$$

and the lengths of intervals symmetric about his conditional mean m_i^* and containing conditional probability α, that is, $2t^*(\alpha)$ in (30),

$$Pr\{m_i^* - t^*(\alpha) < W_i < m_i^* + t^*(\alpha) \,|\, W_{i-1}, W_{i+1}\} = \alpha\,, \tag{30}$$

for a set of selected values for α. In particular, with our restriction to normal distributions, this reduces to comparing the standard deviation of the graduator's prior distribution for W_i and the standard deviation of his conditional distribution for W_i, given W_{i-1} and W_{i+1}, which we will denote by σ_i^*. Now

$$(\sigma_i^*)^2 = \sigma_i^2(1 - d'\,D^{-1}\,d)\,, \tag{31}$$

where

$$d = \begin{bmatrix} c_{i,\,i-1} \\ c_{i,\,i+1} \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} 1 & c_{i-1,\,i+1} \\ c_{i-1,\,i+1} & 1 \end{bmatrix}; \tag{32}$$

so, after determining σ_i and σ_i^* by the method described earlier using equation (28), the graduator has an equation for the three correlation coefficients.

Repetition of this procedure for $\binom{n}{2}$ suitably chosen sets of three rates (some would necessarily not be adjacent) would guide an idealized graduator to a determination of his matrix of correlation coefficients from the class of positive definite matrices with ones on the principal diagonal. But due to the unwieldy system of $\binom{n}{2}$ nonlinear equations which result and, more especially, to the imprecise structure of a real graduator's prior opinion, use of the procedure without some previous restriction of the covariance matrix to a smaller class of matrices is not practical. In Section II we discussed some classes of positive definite matrices which may serve as covariance matrices. We will now use one of these classes in an example.

Let us assume that the graduator will select a matrix from the class $\mathcal{A}$ (defined on p. 82) as the covariance matrix of his prior distribution. Then his correlation matrix will be in the class of nth order matrices $[c_{ij}] = [r^{|i-j|}]$ for $0 \leq r < 1$. Selection of his correlation matrix from this class is equivalent to the selection of a number r between 0 and 1 (0 included) such that r is his prior correlation between every pair of rates W_i and W_{i+1} with subscripts one unit apart, r^2 is his prior correlation with subscripts two units apart, and so forth. Thus, substituting into (32) and then (31), we have

$$d = \begin{bmatrix} r \\ r \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} 1 & r^2 \\ r^2 & 1 \end{bmatrix},$$

and

$$(\sigma_i^*)^2 = \sigma_i^2\left(\frac{1-r^2}{1+r^2}\right).$$

Now, to determine r, the graduator can determine his conditional standard deviation $\sigma_i\sqrt{(1-r^2)/(1+r^2)}$ by the method used to determine his marginal standard deviation σ_i and then solve the resulting equations for r. In particular, suppose for $\alpha = .50$, $\alpha = .90$, and $\alpha = .999$, his values of b which satisfy (28), given the true values of W_{i-1} and W_{i+1}, are $b = .00008$, $b = .0002$, and $b = .0003$, respectively. Hence $.674\sigma_i\sqrt{(1-r^2)/(1+r^2)} = .00008$, $1.645\sigma_i\sqrt{(1-r^2)/(1+r^2)} = .0002$, and $2.574\sigma_i\sqrt{(1-r^2)/(1+r^2)} = .0003$. The average of the solutions for $\sigma_i\sqrt{(1-r^2)/(1+r^2)}$ in the above equations is $\sigma_i\sqrt{(1-r^2)/(1+r^2)} = .00012$; if $\sigma_i = .00097$, then $\sqrt{(1-r^2)/(1+r^2)} = .00012/.00097 = .1237$, and hence $r = .985$.

Conclusions

The methods that we have suggested for eliciting the graduator's prior distribution may appear vague and imprecise. However, we do not feel that this vagueness or imprecision invalidates the use of the Bayesian method. I. J. Good, in his book *The Estimation of Probabilities* [4, p. 11], remarks: "Consider any statistical technique with which you have some sympathy. Find out whether it is equivalent to the use of an initial [prior] probability distribution by using Bayes' theorem in reverse. *Then replace this initial distribution by a better one*" (italics his). Whittaker's graduation method is a technique with which actuaries have sympathy. From Whittaker's exposition [14, p. 303] we know that the method is equivalent to the use of an incompletely specified prior probability distribution. Here, while our model and suggestions have been toward approximations to the graduator's prior distribution, we are confident that the replacement will be a "better prior" than the incomplete one first suggested by Whittaker. We subscribe to the comparison, attributed to the probabilist and actuary B. de Finetti, of those who are deterred from using Bayesian methods because of the availability of only vague determinations of prior distributions to the man who refuses to build his house on sand and attempts to build it instead on a void.

IV. EXAMPLES

In this section examples will be presented of graduations effected by the Whittaker and Bayesian methods. The same data will be graduated by each method in order to facilitate a comparative analysis of the results.

The data to be graduated for this example consist of a sequence of

thirteen weighted observed mortality rates. The rates summarize the experience between the 1961 and 1962 anniversaries of policies issued in 1955 on standard medically examined female lives as published in the *1963 Reports* [13, p. 29]. Each of the first twelve rates is the mortality rate for a five-year age group from ages 10–14 through 65–69, while the thirteenth is the rate for ages 70 and over.

The results of graduating the data by second-difference Whittaker type B formulas are presented in Table 1. The set v_i of graduated rates is that sequence w_i which minimizes the function

$$\sum_{i=1}^{13} e_i(w_i - u_i)^2 + h \sum_{i=1}^{11} (\Delta^2 w_i)^2,$$

where u_i is the ith observed rate, e_i is the total of the face amounts in units of \$1,000,000 for the ith age group, and h is an arbitrary parameter which

TABLE 1

WHITTAKER GRADUATIONS FOR FIVE VALUES OF h

i	AGE GROUP	e_i	$u_i \times 10^3$	GRADUATED RATES $v_i \times 10^3$				
				$h = .1$	$h = 1$	$h = 10$	$h = 100$	$h = 1{,}000$
1.....	10–14	11.64	0.00	0.00	−0.01	−0.10	−0.40	−1.28
2.....	15–19	13.19	0.00	0.00	−0.02	0.00	−0.13	−0.64
3.....	20–24	23.80	0.04	0.05	0.09	0.22	0.17	0.02
4.....	25–29	34.94	0.80	0.80	0.78	0.69	0.59	0.71
5.....	30–34	51.62	1.32	1.31	1.26	1.09	1.16	1.47
6.....	35–39	65.83	1.11	1.12	1.20	1.52	1.96	2.32
7.....	40–44	73.22	3.41	3.40	3.36	3.22	3.21	3.28
8.....	45–49	60.67	4.70	4.70	4.72	4.75	4.51	4.28
9.....	50–54	33.60	6.01	6.03	6.09	6.04	5.64	5.29
10.....	55–59	18.12	7.72	7.64	7.24	6.70	6.51	6.26
11.....	60–64	6.98	4.15	4.33	5.18	6.24	7.11	7.21
12.....	65–69	1.85	5.93	5.86	5.57	6.00	7.69	8.15
13.....	70 and over	0.31	9.74	9.16	6.85	5.88	8.28	9.09

represents the relative emphasis to be placed on the measures of roughness and departure. For h small the graduated rates are close to the observed rates, while for h large the graduated rates tend to lie on a straight line. Results for five values of h between these two extremes are presented in Table 1.

For none of the five values does the Whittaker method produce an acceptable (to us) sequence of rates. For $h = .1$, 1, or 10, the sequence is not always increasing, as our opinion demands. For $h = 100$ or 1,000,

each sequence is always increasing, but each possesses significant negative values which are, of course, impossible for mortality rates. Furthermore, none of the five sequences displays the increasing rate of increment which is characteristic of human mortality at the higher age groups.

The most plausible explanation for such unsatisfactory results is the unusually small observed rates at the three highest age groups. For application of the Whittaker method to such scant data, an artful user of the method would make certain adjustments either to the observed rates or to the graduated rates to bring them into conformity with his opinion. We submit that a more objective and theoretically correct insertion of his opinion may be effected by use of personal probability and Bayesian statistics, as outlined in the previous sections and as now illustrated for this example.

Table 2 presents a Bayesian graduation of the same data that were graduated by the Whittaker method. The Bayesian method defines the vector v of graduated rates to be the mean of the graduator's posterior distribution, which is

$$v = (A^{-1} + B^{-1})^{-1}(B^{-1}u + A^{-1}m),$$

where u is the vector of observed rates, m the vector of his prior means, A the covariance matrix for his prior distribution of the rates, and B the covariance matrix for his conditional distribution of the observed rates given the true rates. The parameters of the method are the n prior means m_i, the n positive elements b_{ii} of the diagonal matrix B, and the $n(n+1)/2$ elements a_{ij} which determine the positive definite, symmetric matrix A.

For this example the vector m of our prior means consists of a sequence of thirteen weighted graduated mortality rates taken from the 1955–60 Select Basic Tables [12, p. 46]. These rates were derived from crude mortality rates in the seventh policy year among female lives on which standard ordinary life insurance policies had been issued in the years 1949–54. The policies had been issued on the basis of a medical examination, except that the data for lives aged below 25 years at the issue date included policies issued without a medical examination. The tabulated rates were the result of graduating the crude rates as part of a larger collection of data. These graduated rates from the 1955–60 Select Basic Tables, which we take for our prior means, represent the same age groups as do the observed data of our example and are similarly weighted by the face amount of the policies.

Our prior standard deviations p_i of the random variables W_i were derived by the method of Section III; that is, if m_i is our prior mean of

W_i, then our prior probability of the event $|W_i - m_i| \leq p_i$ is approximately .683. Hence, once we determine the symmetric interval about m_i in which 68.3 per cent of our prior probability lies, we can then determine p_i. Rather than perform this procedure for each i, it was decided to take our prior standard deviations approximately proportional to $\sqrt{m_i}$. The values for p_i given in Table 2 imply, for example, that to each of the events $0.07 \leq 1{,}000\, W_1 \leq 0.59$ and $29.84 \leq 1{,}000\, W_{13} \leq 34.94$ we assign prior probability .683.

In this example the correlation coefficients of our prior distribution of W were derived by a somewhat different method from that suggested in Section III. Rather than examine the reduction in uncertainty about a rate W_i caused by knowledge of both neighboring rates W_{i-1} and W_{i+1}, we considered the reduction of uncertainty about W_i caused by knowledge only of W_{i+1}. Then, if p_i is our prior standard deviation of W_i, our prior standard deviation of W_i, given the value of W_{i+1}, is $p_i\sqrt{1 - r_i^2}$, where r_i is our correlation between W_i and W_{i+1}. Thus the ratio

$$\frac{p_i}{p_i\sqrt{1 - r_i^2}}$$

of our prior standard deviation of W_i to our standard deviation of W_i, given W_{i+1}, is a measure of the extent to which knowledge of W_{i+1} decreases our uncertainty regarding W_i. In this example the above ratio is taken to be 3 for all i; that is, our standard deviation for W_i, given W_{i+1}, is one-third of our prior standard deviation for W_i. Thus, we have $r_i = r = 2\sqrt{2}/3 = .942809$. Our prior correlation coefficient between W_i and W_{i+k} for $k > 1$ is taken to be r^k, so that our prior covariance matrix is $A = [a_{ij}] = [p_i p_j r^{|i-j|}]$, where the sequence p_i is given by Table 2 and $r = .942809$. The matrix A is a member of the class a defined at the end of Section II.

The diagonal elements $b_{ii} = e_i^{-1}$ of the covariance matrix for our conditional distribution of the observed rates given the true rates were derived according to the method discussed in Section II. We assumed an "equivalent average amount" of \$7,500 and derived each b_{ii} by the formula $b_{ii} = 7{,}500\, m_i(1 - m_i)/N_i$, where N_i is the exposure (in dollars) for the ith age group.

The results of the Bayesian graduation are presented in Table 2. The shape of the sequence of our graduated rates generally follows that of our prior means, although the influence of the observed rates is apparent. For the larger values of i, where the conditional variance of U_i is large because of small exposures, our graduated rates are more strongly influenced by our prior means.

The actual calculation of the graduated rates (including calculation of the b_{ij} from the exposures) was executed on an IBM model 7090 digital computer in well under one minute. The calculation was performed on the basis of formula (7) by solving thirteen linear equations in thirteen unknowns. We shall be happy to supply upon request a FORTRAN II version of the program, which can easily be adapted to run on many medium- and large-scale computers.

TABLE 2

A BAYESIAN GRADUATION

i	Age Group	$N_i \times 10^{-4}$	$u_i \times 10^{3}$	$m_i \times 10^{3}$	$p_i \times 10^{4}$	$v_i \times 10^{3}$
1.....	10–14	11.64	0.00	0.33	2.57	0.22
2.....	15–19	13.19	0.00	0.41	2.86	0.29
3.....	20–24	23.80	0.04	0.58	3.41	0.46
4.....	25–29	34.94	0.80	0.67	3.66	0.59
5.....	30–34	51.62	1.32	1.18	4.86	1.10
6.....	35–39	65.83	1.11	1.91	6.18	1.81
7.....	40–44	73.22	3.41	2.81	7.50	2.87
8.....	45–49	60.67	4.70	3.95	8.88	4.08
9.....	50–54	33.60	6.01	5.33	10.32	5.41
10.....	55–59	18.12	7.72	7.27	12.06	7.21
11.....	60–64	6.98	4.15	12.80	15.99	12.49
12.....	65–69	1.85	5.93	20.50	20.25	20.03
13.....	70 and over	0.31	9.74	32.39	25.50	31.81

V. FURTHER DEVELOPMENTS

We conclude with a brief report on some further developments of the Bayesian graduation method and some indications of other problems to which an extension of the method may be applicable.

In Section II we demonstrated that matrices in class a, that is, of the form $A = [p^2 r^{|i-j|}]$, produce generalizations of Whittaker graduations with first-difference measures of roughness. Similarly, there exist classes of admissible matrices which generalize measures of roughness involving higher-order differences (see reference [8]). As actuaries gain experience in using the Bayesian method, we anticipate the discovery of still other interesting classes of matrices which will prove useful for approximating the true covariance matrix of one's prior distribution of the rates.

The Bayesian method is also applicable to the problem of combining graduation with interpolation. For example, we may have crude rates available at intervals of five years of age and want to produce graduated rates for every year. To apply the Bayesian method to this problem, we take n equal to the number of graduated rates we want to produce; for

each i for which we have no observation, we take the corresponding reciprocals of the variances to be zero. Furthermore, if for each value of i for which there is an observation, the conditional variance b_{ii} of U_i is small compared with the prior variance p_i^2 of W_i, then the graduated sequence v_i will yield interpolated values for the remaining values of i. Thus, the Bayesian method may also produce a near nonmodified interpolation formula. The ideas of this paragraph apply equally well to extrapolation or projection.

A common actuarial problem for which no generally acceptable solution yet exists is the graduation of select data and multiple decrement data. The method as herein presented is not directly applicable, because our definition of admissibility assumed the rates arranged in a natural sequence. Some research has been conducted on expanding the notion of admissibility and on generating classes of matrices necessary for Bayesian graduation of select data, but there is need for more research in this area (see reference [8]).

APPENDIX I

The purpose of these appendixes is to make the paper self-contained without including explanations of unfamiliar ideas which are not germane to the graduation theory being developed. This first appendix contains an introduction to the algebra of matrices of real numbers and functions. The second appendix contains the applications of matrices to probability.

Definition 1.—An $m \times n$ *matrix* is a rectangular array of mn scalars, the array having m rows and n columns. The mn scalars, which are called the *elements* of the matrix, are members of a set of things for which we have addition, subtraction, multiplication, and division. In our applications the scalars will be real numbers, functions, or random variables. For example,

$$\begin{bmatrix} 2 & 4 & 0 \\ 3 & 0 & 1 \end{bmatrix} \qquad \begin{bmatrix} e^x & 3 & x \\ x^2 & 1/x & 0 \\ x+5 & \sin x & 2x \end{bmatrix}.$$

Definition 2.—The *dimension* (we shall abbreviate it "dim") of an $m \times n$ matrix is the ordered pair $m \times n$; m is the *row dimension* and n is the *column dimension.*

The dims of the two matrices above are 2×3 and 3×3, respectively.

Definition 3.—A *square matrix* is an $n \times n$ matrix and is said to be of *order* n.

Definition 4.—A *row vector* (or row matrix) is a $1 \times n$ matrix.

Definition 5.—A *column vector* (or column matrix) is an $n \times 1$ matrix.

In our discussion of the algebra of matrices some abbreviations will be helpful. An $m \times n$ matrix of arbitrary scalars is written

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

or, more compactly, as

$$[a_{ij}]_{m\times n} .$$

If the dim is clear from the context, we may suppress the $m \times n$. For a vector we may write

$$[a_i]_{1\times n} \quad \text{or} \quad [a_i]_{n\times 1} .$$

The sentence "A is the $m \times n$ matrix whose element in the ith row and jth column will be denoted by a_{ij}" is abbreviated to

$$A = [a_{ij}]_{m\times n} .$$

Definition 6.—Two matrices are equal, written $[b_{ij}] = [a_{ij}]$, if they have the same dim, and $b_{ij} = a_{ij}$, for all i and j.

$$\begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & \frac{2}{4} \end{bmatrix},$$

$$\begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & \frac{1}{2} \end{bmatrix} \neq \begin{bmatrix} 1 & 2 & 3 \\ 3 & 4 & \frac{1}{2} \end{bmatrix},$$

$$\begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & \frac{1}{2} \end{bmatrix} \neq \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 3 & \frac{1}{2} \end{bmatrix}.$$

Just as the operations of addition and multiplication for numbers and functions give rise to "ordinary" algebra, there are operations performed on matrices that give rise to the algebra of matrices.

Definition 7.—The *transpose* of the $m \times n$ matrix A is the $n \times m$ matrix, denoted by A', whose ith column is the ith row of A, $i = 1, 2, \ldots, m$ (and hence the ith row of A' is the ith column of A, $i = 1, 2, \ldots, n$).

$$\begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & \frac{1}{2} \end{bmatrix}' = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 3 & \frac{1}{2} \end{bmatrix},$$

$$\begin{bmatrix} e^x & 3 & x \\ x^2 & 1/x & 0 \\ x+5 & \sin x & 2x \end{bmatrix}' = \begin{bmatrix} e^x & x^2 & x+5 \\ 3 & 1/x & \sin x \\ x & 0 & 2x \end{bmatrix},$$

$$[1 \quad x \quad x^2 \quad x^3 \quad x^4]' = \begin{bmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \end{bmatrix},$$

Rule 1. $(A')' = A$.

Definition 8: Scalar multiplication.—Let b be a scalar and $A = [a_{ij}]_{m\times n}$. Then $bA = [c_{ij}]_{m\times n}$ and $Ab = [d_{ij}]_{m\times n}$, where $c_{ij} = ba_{ij}$ and $d_{ij} = a_{ij}b$, for all i and j.

$$[x \quad x^2 \quad x^3] = x[1 \quad x \quad x^2],$$

$$\begin{bmatrix} 2 & 1 & 3 \\ 4 & 3 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} 4 & 2 & 6 \\ 8 & 6 & 1 \end{bmatrix}\frac{1}{2}.$$

Rule 2. For all scalars b and c and matrices A,

(i) $cA = Ac$,

(ii) $(bc)A = b(cA) = (bcA)$,

(iii) $(cA)' = cA'$.

Definition 9: Matrix addition.—

$$[a_{ij}]_{m\times n} + [b_{ij}]_{m\times n} = [c_{ij}]_{m\times n},$$

where $c_{ij} = a_{ij} + b_{ij}$, $i = 1, 2, \ldots m$, and $j = 1, 2, \ldots n$. Note that addition of matrices is defined only for matrices of the same dim.

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix},$$

$$\begin{bmatrix} 3 & 2 & 5 \\ 4 & 1 & 1 \end{bmatrix} + \begin{bmatrix} 3 & 6 & 7 \\ 8 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 8 & 12 \\ 12 & 2 & 2 \end{bmatrix},$$

$$\begin{bmatrix} 3 & 1 \\ -5 & 4 \end{bmatrix} + \begin{bmatrix} -3 & -1 \\ 5 & -4 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Rule 3. For all matrices A, B, and C (of equal dim) and scalars b and c,

(i) $(A + B) + C = A + (B + C)$,

(ii) $A + B = B + A$,

(iii) $c(A + B) = cA + cB$,

(iv) $(c + b)A = cA + bA$,

(v) $(A + B)' = A' + B'$.

Definition 10: Matrix subtraction.—

$$A - B = A + (-1)B .$$

Definition 11.—The *inner product* $x \cdot y$ of the vectors $x = [x_1 x_2 \ldots x_n]$ and $y = [y_1 y_2 \ldots y_n]$ is the scalar

$$\sum_1^n x_i y_i .$$

For example,

$$[1\ 2\ 3\ 4] \cdot [5\ 7\ 9\ 11] = 5 + 14 + 27 + 44 = 90 ,$$

$$[x_1 x_2 x_3 \ldots x_n] \cdot [x_1 x_2 x_3 \ldots x_n] = \sum_1^n x_i^2 .$$

Definition 12: Matrix multiplication.—The product AB of $A = [a_{ij}]_{m \times n}$ and $B = [b_{ij}]_{n \times p}$ is the $m \times p$ matrix C whose element c_{ij} in the ith row and jth column is the inner product of the ith row vector of A and the jth column vector of B; compactly,

$$[a_{ij}]_{m \times n} [b_{ij}]_{n \times p} = [c_{ij}]_{m \times p} ,$$

where

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj} .$$

Observe that, for AB to be defined, the column dimension of A must equal the row dimension of B, and then the row dimension of AB equals the row dimension of A, and the column dimension of AB equals the column dimension of B. For example,

$$\begin{bmatrix} 3 & 2 & 5 \\ 4 & 1 & 1 \end{bmatrix} \quad \begin{bmatrix} 3 & 8 \\ 6 & 1 \\ 7 & 1 \end{bmatrix} = \begin{bmatrix} 56 & 31 \\ 25 & 34 \end{bmatrix} ,$$

$$2 \times 3 \qquad 3 \times 2 \qquad 2 \times 2$$

$$c_{11} = (3)(3) + (2)(6) + (5)(7) = 56 ,$$

$$c_{12} = (3)(8) + (2)(1) + (5)(1) = 31 ,$$

$$c_{21} = (4)(3) + (1)(6) + (1)(7) = 25 ,$$

$$c_{22} = (4)(8) + (1)(1) + (1)(1) = 34 .$$

Rule 4. For all matrices A, B, and C (of dims such that the indicated products are defined) and scalars a,

$$\text{(i)}\quad a(BC) = (aB)C\,,$$

$$\text{(ii)}\quad A(BC) = (AB)C\,,$$

$$\text{(iii)}\quad A(B+C) = AB + AC\,,$$

$$\text{(iv)}\quad (B+C)A = BA + CA\,,$$

$$\text{(v)}\quad (AB)' = B'A'\,.$$

But, in general, $\boxed{AB \neq BA}$.

The algebra of matrices differs from real number algebra in two important ways. The first of these is the failure of the commutative law for matrix multiplication, which we illustrate.

a)
$$\begin{bmatrix}4 & 2\\ 3 & 4\\ 1 & 1\end{bmatrix}\begin{bmatrix}4 & 0\\ 1 & 1\end{bmatrix} = \begin{bmatrix}18 & 2\\ 16 & 4\\ 5 & 1\end{bmatrix},$$

but

$$\begin{bmatrix}4 & 0\\ 1 & 1\end{bmatrix}\begin{bmatrix}4 & 2\\ 3 & 4\\ 1 & 1\end{bmatrix}$$

is not defined.

b)
$$\begin{bmatrix}3 & 2 & 5\\ 4 & 1 & 1\end{bmatrix}\begin{bmatrix}3 & 8\\ 6 & 1\\ 7 & 1\end{bmatrix} = \begin{bmatrix}56 & 31\\ 25 & 34\end{bmatrix},$$

but

$$\begin{bmatrix}3 & 8\\ 6 & 1\\ 7 & 1\end{bmatrix}\begin{bmatrix}3 & 2 & 5\\ 4 & 1 & 1\end{bmatrix} = \begin{bmatrix}41 & 14 & 23\\ 22 & 13 & 31\\ 25 & 15 & 36\end{bmatrix}.$$

c)
$$\begin{bmatrix}4 & 0\\ 1 & 1\end{bmatrix}\begin{bmatrix}0 & 2\\ 3 & 0\end{bmatrix} = \begin{bmatrix}0 & 8\\ 3 & 2\end{bmatrix},$$

but

$$\begin{bmatrix}0 & 2\\ 3 & 0\end{bmatrix}\begin{bmatrix}4 & 0\\ 1 & 1\end{bmatrix} = \begin{bmatrix}2 & 2\\ 12 & 0\end{bmatrix}.$$

These examples illustrate the three ways for $AB = BA$ to fail. In (*a*) only one product was defined, in (*b*) both products were defined but the dim of AB does not equal the dim of BA, and in (*c*) both products were defined and had equal dims but the products were not equal.

Definition 13.—The *zero matrix* of dim $m \times n$, denoted $[0]_{m\times n}$ or just $[0]$, is the $m \times n$ matrix with zero for every element.

Rule 5.

$$[a_{ij}] + [0] = [a_{ij}]\,,$$

$$[a_{ij}] - [a_{ij}] = [0],$$

$$[a_{ij}]_{m\times n}\,[0]_{n\times p} = [0]_{m\times p}.$$

In the algebra of real numbers, if $ab = 0$, then $a = 0$ or $b = 0$. The failure of this implication in matrix algebra is the other important difference between the algebras.

For example,

$$\begin{bmatrix}3 & 1\\ 6 & 2\end{bmatrix}\begin{bmatrix}-1 & 4\\ 3 & -12\end{bmatrix} = \begin{bmatrix}0 & 0\\ 0 & 0\end{bmatrix},$$

$$\begin{bmatrix}3 & 1 & 3\\ 1 & 2 & 2\end{bmatrix}\begin{bmatrix}4\\ 3\\ -5\end{bmatrix} = \begin{bmatrix}0\\ 0\end{bmatrix}.$$

We will restrict our consideration of the corresponding matrix implication to the product of two square matrices and the product of a square matrix and a vector. In fact, all the rules and definitions henceforth will be for square matrices and vectors.

Definition 14.—A square matrix $[a_{ij}]_{n\times n}$ is *nonsingular* if

$$[a_{ij}]_{n\times n}\,[b_{j1}]_{n\times 1} = [0]_{n\times 1}$$

implies

$$[b_{j1}]_{n\times 1} = [0]_{n\times 1}.$$

If a matrix fails to be nonsingular, it is called singular.

Thus, from above

$$\begin{bmatrix}3 & 1\\ 6 & 2\end{bmatrix}$$

is singular.

Rule 6. If $[a_{ij}]_{n\times n}$ is nonsingular and $[a_{ij}]_{n\times n}\,[b_{ij}]_{n\times n} = [0]_{n\times n}$, then $[b_{ij}]_{n\times n} = [0]_{n\times n}$. (Look at the n products $[a_{ij}]_{n\times n}\,[b_{ij}]_{n\times 1}$, $j = 1, 2, \ldots, n$ one at a time.)

There are two more formulations of nonsingularity for square matrices, each of which has its direct application. The first that we give will be the analogue of the reciprocal $1/a$, and the second will be in terms of determinants to serve as a computational check for nonsingularity. First we need a unit, that is, a multiplicative identity.

Definition 15.—The *principal diagonal* of the square matrix $[a_{ij}]$ is the set of elements a_{ii}, $i = 1, 2, \ldots n$, which are called *diagonal elements*. A *diagonal matrix* is a square matrix with nonzero elements appearing only on the principal diagonal. For example,

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

is a diagonal matrix.

Definition 16.—The *identity matrix* of order n, denoted by I_n or just I if the order is clear from the context, is the diagonal matrix whose principal diagonal consists only of 1's.

$$I_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Rule 7.

$$I_n[a_{ij}]_{n\times p} = [a_{ij}]_{n\times p}$$

and

$$[b_{ij}]_{m\times n} I_n = [b_{ij}]_{m\times n}.$$

From this rule we can see that I_n is nonsingular and that it commutes with every $n \times n$ matrix.

Rule 8. For each nonsingular matrix A there exists a unique matrix B such that $AB = BA = I$.

Definition 17.—The *inverse* of the nonsingular matrix A is the unique matrix B of Rule 8. B is denoted by A^{-1}. For example,

$$A = \begin{bmatrix} 4 & 0 \\ 1 & 1 \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} \frac{1}{4} & 0 \\ -\frac{1}{4} & 1 \end{bmatrix};$$

$$I_n^{-1} = I_n;$$

$$A = \begin{bmatrix} d_1 & 0 & \dots & 0 \\ 0 & d_2 & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ 0 & & & d_n \end{bmatrix}, \qquad A^{-1} = \begin{bmatrix} 1/d_1 & 0 & \dots & 0 \\ 0 & 1/d_2 & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ 0 & & & 1/d_n \end{bmatrix}, d_i \neq 0;$$

$$A = \begin{bmatrix} 4 & 0 & 1 \\ 1 & 1 & 1 \\ 7 & -1 & 1 \end{bmatrix}, \qquad A^{-1} \text{ does not exist.}$$

Rule 9.

$$(A^{-1})' = (A')^{-1},$$

$$(AB)^{-1} = B^{-1}A^{-1}.$$

A system of linear equations can be expressed as a matrix equation whose solution involves finding the inverse of the matrix whose elements

are the coefficients in the system of equations. For example, let

$$A = \begin{bmatrix} 4 & 0 & 1 \\ 1 & 1 & 0 \\ 7 & -1 & 1 \end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix};$$

then

$$Aw = \begin{bmatrix} 4w_1 + w_3 \\ w_1 + w_2 \\ 7w_1 - w_2 + w_3 \end{bmatrix}.$$

The matrix equation

$$Aw = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

is equivalent to the system

$$4w_1 + 0w_2 + 1w_3 = 3,$$

$$1w_1 + 1w_2 + 0w_3 = 3,$$

$$7w_1 - 1w_2 + 1w_3 = 3.$$

If we solve the system by substitution, we have $w_1 = \frac{3}{4}$, $w_2 = 2\frac{1}{4}$, and $w_3 = 0$. For the matrix equation we may multiply both sides by A^{-1}:

$$A^{-1}(Aw) = A^{-1}\begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix},$$

where

$$A^{-1} = \begin{bmatrix} -\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ 2 & -1 & -1 \end{bmatrix}.$$

Now

$$A^{-1}(Aw) = (A^{-1}A)w = I_3 w = \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$$

and

$$A^{-1}\begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = \tfrac{1}{4}\begin{bmatrix} -1 & 1 & 1 \\ 1 & 3 & -1 \\ 8 & -4 & -4 \end{bmatrix}\begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} = \tfrac{1}{4}\begin{bmatrix} 3 \\ 9 \\ 0 \end{bmatrix},$$

so

$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} = \tfrac{1}{4}\begin{bmatrix} 3 \\ 9 \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{3}{4} \\ 2\frac{1}{4} \\ 0 \end{bmatrix},$$

as before.

For the other formulation of nonsingularity we need the following definition.

Definition 18.—The *determinant* of $A = [a_{ij}]_{n\times n}$, denoted by $|A|$, is the scalar

$$|A| = a_{11} \qquad \text{if } n = 1,$$

$$= \sum_{j=1}^{n} a_{1j} |A_{1j}| (-1)^{j+1} \qquad \text{if } n > 1,$$

where A_{1j} is the $n-1 \times n-1$ matrix obtained by deleting the first row and the jth column of A.

For example, $|I_n| = 1$,

$$\begin{vmatrix} 3 & 1 \\ 6 & 2 \end{vmatrix} = 3\cdot 2 + 1\cdot 6(-1) = 0,$$

$$\begin{vmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & \\ \cdot & & \cdot & \\ \cdot & & & \cdot \\ 0 & & & d_n \end{vmatrix} = d_1 d_2 \cdots d_n,$$

$$\begin{vmatrix} 4 & 0 & 1 \\ 1 & 1 & 0 \\ 7 & -1 & 1 \end{vmatrix} = 4 \cdot \begin{vmatrix} 1 & 0 \\ -1 & 1 \end{vmatrix} + 0 \begin{vmatrix} 1 & 0 \\ 7 & 1 \end{vmatrix} (-1) + 1 \begin{vmatrix} 1 & 1 \\ 7 & -1 \end{vmatrix} (-1)^2$$

$$= 4 \cdot 1 + 0 + 1(-8) = -4.$$

Rule 10. The determinant as herein defined is the same as the "determinant" studied for the General Mathematics Examination, so you may recall your own rules of calculation here. For example,

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}.$$

Rule 11. A square matrix A is nonsingular if and only if $|A| \neq 0$.

Summary of nonsingularity: For a square matrix A the following are equivalent:

1. A is nonsingular.
2. $A[b_j]_{n\times 1} = [0]_{n\times 1}$ implies $[b_j] = [0]$.
3. $|A| \neq 0$.
4. There exists a matrix A^{-1} such that $A^{-1}A = AA^{-1} = I_n$.

A principal application of matrices was indicated above in the relationship between matrix equations and systems of linear equations. The other important application of matrices in this paper is the manipulation of sums of squares or, more generally, quadratic forms (sums of squares with cross-product terms). Following Definition 11, we saw that the sum of the

squares of n quantities may be represented as the inner product of a vector and itself where the n quantities are the elements of the vector; for example,

$$[x_1x_2x_3] \cdot [x_1x_2x_3] = x_1^2 + x_2^2 + x_3^2 .$$

Formally we may also view the inner product of two vectors as the matrix product of a row matrix and a column matrix by treating 1×1 matrices as scalars. (1×1 matrices act like scalars among themselves, and we have no other need for 1×1 matrices.) For example,

$$[x_1x_2x_3]\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = [x_1y_1 + x_2y_2 + x_3y_3] ,$$

which we shall identify as

$$x_1y_1 + x_2y_2 + x_3y_3 .$$

For example, if we introduce vectors as column matrices, that is, $x' = [x_1x_2 \ldots x_n]$, then we may write

$$x'x = [\, x_1x_2 \ldots x_n]\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \sum_1^n x_i^2 .$$

As a preliminary to Definitions 19 and 20, let us examine the general "quadratic form" in three variables, that is, the sum of a weighted sum of squares and second-degree cross-products. Let $x' = [x_1, x_2, x_3]$ and $A = [a_{ij}]_{3\times 3}$, and consider the product $x'Ax$; that is,

$$[x_1x_2x_3]\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

$$= a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2$$

$$+ (a_{12} + a_{21})x_1x_2 + (a_{31} + a_{13})x_1x_3 + (a_{23} + a_{32})x_2x_3 .$$

We should observe that (1) if A is a diagonal matrix, the product is the weighted sum of squares $a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2$; (2) if $A = I_3$, the product is the "unweighted" sum of squares $x_1^2 + x_2^2 + x_3^2$; and (3) if A is replaced by

$$B = \begin{bmatrix} a_{11} & 0 & 0 \\ (a_{21} + a_{12}) & a_{22} & 0 \\ (a_{31} + a_{13}) & (a_{23} + a_{32}) & a_{33} \end{bmatrix} ,$$

the same quadratic form is obtained; that is,

$$x'Ax = x'Bx$$

for all x.

Part (3) of this example raises questions like the following one: Given the quadratic form $x_1^2 + 3x_2^2 + 7x_3^2 + 8x_1x_2 + x_2x_3 - 2x_1x_3$ to be written in the form $x'Ax$, what matrix should be chosen for A? The natural choice is to require that A be "symmetric."

$$A = \begin{bmatrix} 1 & 4 & -1 \\ 4 & 3 & \frac{1}{2} \\ -1 & \frac{1}{2} & 7 \end{bmatrix}.$$

The reason for calling this the natural choice will be clear after some examples.

Definition 19.—A *symmetric matrix* is a square matrix with the property $a_{ij} = a_{ji}$ for all i and j, that is, symmetry about the principal diagonal.

Rule 12.

(i) Diagonal matrices are symmetric.
(ii) A is symmetric if and only if $A' = A$.
(iii) For any matrix B, $B'B$ is symmetric.
(iv) A nonsingular matrix B is symmetric if and only if B^{-1} is symmetric.

For example, part (iii) of Rule 12 is related to the reason for choosing symmetric matrices to support quadratic forms. To emphasize the rule, we shall use a nonsquare matrix which will serve us again in another example. Let

$$B = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix};$$

then $B'B$ is

$$\begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ -2 & 5 & -4 & 1 & 0 \\ 1 & -4 & 6 & -4 & 1 \\ 0 & 1 & -4 & 5 & -2 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix}.$$

Definition 20.—A *quadratic form* (Q.F.) in the n variables $x_1, x_2, \ldots, x_n$ is $x'Ax$, where $x' = [x_1x_2 \ldots x_n]$ and A is a symmetric matrix. For example,

a) $[x_1x_2]\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = x_1^2 + 2x_1x_2 + x_2^2 = (x_1 + x_2)^2$;

b) $[x_1x_2]\begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = (x_1 - x_2)^2 + x_2^2$;

c) the sum of the squares of the second differences of x_1, x_2, x_3, x_4, and x_5 can be written as $v'v$, where v is the vector of second differences, that is, $v' = [\Delta^2 x_1, \Delta^2 x_2, \Delta^2 x_3]$.

$$v'v = \sum_1^3 (\Delta^2 x_i)^2 .$$

Now the three second differences are three linear combinations of the x_i's, so we can construct a 3×5 matrix to multiply by $x' = [x_1 x_2 x_3 x_4 x_5]$ to obtain these second differences. The matrix B of the example preceding Definition 20 is the necessary 3×5 matrix. Thus,

$$Bx = \begin{bmatrix} 1 & -2 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} \Delta^2 x_1 \\ \Delta^2 x_2 \\ \Delta^2 x_3 \end{bmatrix} = v .$$

Substituting for v, we have

$$\begin{aligned} v'v &= (Bx)'(Bx) \\ &= (x'B')(Bx) \\ &= x'(B'B)x . \end{aligned} \qquad \text{(by Rules 1 and 4)}$$

In words, the sum of the squares of linear combinations of several variables is a quadratic form (sum of squares with cross-product terms) in the several variables. The matrix of the quadratic form is the product of the transpose of the matrix of the linear combinations and itself, that is, a symmetric matrix.

Quadratic forms are classified by the range of their values. Thus the Q.F. $x_1^2 + x_2^2$, which is positive for all $[x_1 x_2] \neq [0]$, is said to be positive definite. The Q.F. $x_1^2 + 2x_1x_2 + x_2^2 = (x_1 + x_2)^2$, which is nonnegative for all $[x_1 x_2] \neq [0]$, is said to be positive semidefinite. The ideas of negative definite and negative semidefinite are defined as one would expect. There are Q.F.'s which are neither positive semidefinite nor negative semidefinite; for example, $x_1^2 + 2x_1x_2 = (x_1 + x_2)^2 - x_2^2$ may assume positive values and negative values.

A symmetric matrix is classified according to the classification of the Q.F. defined by the matrix.

Definition 21.—$x'Ax$ (or the symmetric matrix A) is *positive definite* in case $x'Ax > 0$ for all $x \neq [0]$. If we change ">0" to "≥ 0," then $x'Ax$ (or A) is *positive semidefinite.*

Rule 13.

(i) A positive definite Q.F. (or symmetric matrix) is positive semidefinite.

(ii) The sum of positive semidefinite Q.F.'s (or symmetric matrices) is positive semidefinite. If at least one of the Q.F.'s (or symmetric matrices) is positive definite, then the sum is positive definite.

(iii) A positive definite symmetric matrix is nonsingular. (If A were singular, there would exist a nonzero vector $x' = [x_1 x_2 \ldots x_n]$ such that $x'A = [0]$. It follows that $x'Ax = 0$, a contradiction to A's being positive definite.)

(iv) $[a_{ij}]_{n\times n}$ is positive definite if and only if all of the n determinants

$$|a_{11}|, \quad \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}, \quad \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}, \text{ etc.},$$

are positive.

APPENDIX II

Appendix I describes some of the elements of matrix algebra which may be used to facilitate the manipulation of several variables—indeed, we shall use them to manipulate the many variables of graduation. In this appendix we shall summarize the definitions and rules for the use of matrices in the manipulation of several random variables.

Our random variables will be real valued—the usual kind. In general, a random variable is a mathematical construct which enables one to make statements about the outcome of a random situation. To use real valued random variables means to describe the outcome by a numerical characteristic. Instead of asking, "What is the probability that this scale will indicate a number which exceeds 200 when I step on it?" we just write $P(X > 200) = ?$, where the X stands for the verbal description of the outcome. Since the outcomes are described by real numbers, we may treat random variables as real valued functions with respect to addition, subtraction, multiplication, and division.

Definition 22.—A *random vector* is a vector whose elements (i.e., scalars) are random variables.

At the risk of laboring this definition, let us illustrate it with some simple, familiar, finite situations. If a ball is to be placed in one of three numbered urns, we may define three random variables, Y_1, Y_2, Y_3, where $Y_1 = 1$ or 0 according to whether the ball is in urn 1 or not, and Y_2 and Y_3 are defined correspondingly for urns 2 and 3. Thus $Y_1 = 0$, $Y_2 = 1$,

$Y_3 = 0$ is the outcome with the ball placed in the number 2 urn. Now we may define the random vector

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix},$$

which will have three possible outcomes:

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

If the urns are "equally likely," we would write

$$P\left(\boldsymbol{Y} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}\right) = \tfrac{1}{3}, \text{ etc.}$$

When a red die and a green die are rolled, we may describe the outcome with a two-dimensional random vector $\boldsymbol{X}$, whose first element is the random variable for the red die and the second element is for the green die. The sample space for $\boldsymbol{X}$ is the familiar set of 36 outcome pairs.

In a mortality study we have the random vector $\boldsymbol{\Theta}$, whose ith element is the number of deaths to be observed in the ith age group. There is also the vector $\boldsymbol{E}$ of the corresponding exposures, which is usually considered to be nonrandom. There is the random vector of observed death rates, $\boldsymbol{Q}' = [\Theta_1/E_1\ \Theta_2/E_2 \ldots \Theta_n/E_n]$. (Here we display $\boldsymbol{Q}'$ only to avoid printing the long vertical vector, $\boldsymbol{Q}$.) In our analysis by personal probability there is also the random vector of "true" death rates.

Definition 23.—A *random matrix* is a matrix whose elements (i.e., the scalars) are random variables. Since vectors are special matrices, this definition includes Definition 22, but we shall only make technical use of random matrices which are not vectors, so we set these definitions apart.

For example, if $\boldsymbol{X}' = [X_1X_2X_3]$ is a random vector, then $\boldsymbol{X}'\boldsymbol{X}$ is the random variable $X_1^2 + X_2^2 + X_3^2$, and $\boldsymbol{X}\boldsymbol{X}'$ is the 3×3 random matrix

$$\begin{bmatrix} X_1^2 & X_1X_2 & X_1X_3 \\ X_2X_1 & X_2^2 & X_2X_3 \\ X_3X_1 & X_3X_2 & X_3^2 \end{bmatrix}.$$

Definition 24.—For an $m \times n$ random matrix, say,

$$\boldsymbol{X} = \begin{bmatrix} X_{11} & X_{12} & \ldots & X_{1n} \\ X_{21} & X_{22} & \ldots & X_{2n} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ X_{m1} & X_{m2} & \ldots & X_{mn} \end{bmatrix},$$

the *expected value* of X, written $E[X]$, is the $m \times n$ matrix of real numbers (if such exist)

$$E[X] = \begin{bmatrix} E[X_{11}] & E[X_{12}] & \cdots & E[X_{1n}] \\ E[X_{21}] & E[X_{22}] & \cdots & E[X_{2n}] \\ \vdots & & & \\ E[X_{m1}] & E[X_{m2}] & \cdots & E[X_{mn}] \end{bmatrix},$$

where E is the expectation operator,

$$E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$$

or $\Sigma x P(X = x)$ (see reference [6], p. 133).

For example, referring to the random vector Y of the ball and urns example and assuming equal chances for each urn, we have

$$E[Y] = \begin{bmatrix} E[Y_1] \\ E[Y_2] \\ E[Y_3] \end{bmatrix} = \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}.$$

For example, in graduation theory it has usually been assumed that $E[\Theta_x] = q_x E_x$ (see reference [10], p. 26). Thus, for the random vectors Θ' and Q', we have

$$E[\Theta'] = [q_1 E_1, \ldots, q_n E_n]$$

and

$$E[Q'] = [q_1, q_2, \ldots, q_n].$$

Rule 14. Let X and Y be random matrices, and let A and B be real matrices such that the following sums and products are defined. Then

$$\text{(i)}\quad E[X + Y] = E[X] + E[Y],$$

$$\text{(ii)}\quad E[AXB] = AE[X]B.$$

(Any element of AXB is of the form $\Sigma\Sigma c_{ij} X_{ij}$ and $E[\Sigma\Sigma c_{ij} X_{ij}] = \Sigma\Sigma c_{ij} E[X_{ij}]$. Thus we may obtain $E[AXB]$ by replacing X in AXB by $E[X]$.)

Definition 25.—The *covariance matrix* of an n-dimensional random vector X is the $n \times n$ matrix $[\sigma_{ij}]$, where $\sigma_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])]$, the covariance of X_i and X_j, if $i \neq j$, and the variance of X_i when $i = j$.

Rule 15. $[\sigma_{ij}]$ is symmetric.

$$\sigma_{ij} = E[X_i X_j] - E[X_i]\, E[X_j].$$

For example, consider the random vector Y in the example of the ball and (equally likely) urns.

$$E[\boldsymbol{Y}'] = [\tfrac{1}{3}\ \tfrac{1}{3}\ \tfrac{1}{3}]\,;$$

$$E[Y_iY_j] = 0\,,$$

if $i \neq j$, since at most one Y_i can be different from zero;

$$\therefore \sigma_{ij} = 0 - (\tfrac{1}{3})(\tfrac{1}{3}) = -\tfrac{1}{9}\,, \qquad i \neq j\,.$$

Since Y_i is 0 or 1,

$$Y_i^2 = Y_i \quad \text{and} \quad E[Y_i^2] = E[Y_i] = \tfrac{1}{3}\,;$$

$$\therefore \sigma_{ii} = E[Y_i^2] - E[Y_i]^2 = \tfrac{1}{3} - \tfrac{1}{9} = \tfrac{2}{9}\,.$$

Thus, the covariance matrix for $\boldsymbol{Y}$ is

$$\tfrac{1}{9}\begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}.$$

Rule 16. The covariance matrix of any random vector $\boldsymbol{X}$ is the matrix

$$E[(\boldsymbol{X} - E[\boldsymbol{X}])(\boldsymbol{X} - E[\boldsymbol{X}])']\,.$$

For example, consider this rule for $n = 2$:

$$\begin{bmatrix} X_1 - E[X_1] \\ X_2 - E[X_2] \end{bmatrix} \quad [X_1 - E[X_1],\ X_2 - E[X_2]]$$

$$= \begin{bmatrix} (X_1 - E[X_1])^2 & (X_1 - E[X_1])\ (X_2 - E[X_2]) \\ (X_2 - E[X_2])\ (X_1 - E[X_1]) & (X_2 - E[X_2])^2 \end{bmatrix}.$$

The expected value of this matrix is the covariance matrix of $\boldsymbol{X}$.

Rule 17. If $\boldsymbol{Y} = A\boldsymbol{X}$, C_X and C_Y are, respectively, the covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, then $C_Y = AC_XA'$. (This is the n-dimensional generalization of the following one-dimension rule: If $Y = aX$, then $\sigma_Y^2 = a\sigma_X^2 \cdot a = a^2\sigma_X^2$.) To prove the rule:

$$\begin{aligned} C_Y &= E[(\boldsymbol{Y} - E[\boldsymbol{Y}])\ (\boldsymbol{Y} - E[\boldsymbol{Y}])'] \\ &= E[(A\boldsymbol{X} - E[A\boldsymbol{X}])\ (A\boldsymbol{X} - E[A\boldsymbol{X}])'] \\ &= E[(A\boldsymbol{X} - AE[\boldsymbol{X}])\ (A\boldsymbol{X} - AE[\boldsymbol{X}])'] \\ &= E[A(\boldsymbol{X} - E[\boldsymbol{X}])\ \{A(\boldsymbol{X} - E[\boldsymbol{X}])\}'] \\ &= E[A(\boldsymbol{X} - E[\boldsymbol{X}])\ (\boldsymbol{X} - E[\boldsymbol{X}])'A'] \\ &= AE[(\boldsymbol{X} - E[\boldsymbol{X}])\ (\boldsymbol{X} - E[\boldsymbol{X}])]A' \\ &= AC_XA'\,. \end{aligned}$$

For example, let $X' = [X_1, X_2]$,

$$C_X = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

and

$$A = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

Then

$$Y = AX = \begin{bmatrix} X_1 - X_2 \\ X_1 + X_2 \end{bmatrix}$$

and

$$C_Y = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}.$$

The covariance of a pair of random variables X and Y is readily calculated, but it is not as informative as the scale-free correlation (coefficient) of X and Y; that is,

$$\rho_{XY} = \frac{\operatorname{cov}(X, Y)}{\sigma_X \sigma_Y} = \operatorname{cov}\left(\frac{X}{\sigma_X}, \frac{Y}{\sigma_Y}\right).$$

If, in the covariance matrix of $\boldsymbol{X}$, we divide every element in the ith row by $\sqrt{\operatorname{var} X_i} = \sigma_i$ and every element in the jth column by $\sqrt{\operatorname{var} X_j} = \sigma_j$, the resulting matrix will have a principal diagonal of ones and correlations as the off-diagonal elements.

Definition 26.—The *correlation matrix* of the n-dimensional random vector $\boldsymbol{X}$, denoted by R_X, is

$$R_X = \begin{bmatrix} \sigma_1^{-1} & 0 & \dots & 0 \\ 0 & \sigma_2^{-1} & \dots & 0 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ 0 & 0 & \dots & \sigma_n^{-1} \end{bmatrix} \cdot C_X \cdot \begin{bmatrix} \sigma_1^{-1} & 0 & \dots & 0 \\ 0 & \sigma_2^{-1} & \dots & 0 \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & & \cdot & \cdot \\ 0 & 0 & \dots & \sigma_n^{-1} \end{bmatrix}.$$

Note that left (right) multiplication of a given matrix by a diagonal matrix multiplies each element of the ith row (col.) of the given matrix by the scalar in the ith row (col.) of the diagonal matrix.

APPENDIX III

In this appendix we shall discuss the multivariate normal distribution.

Definition 27.—The n-dimensional random vector $\boldsymbol{X}$ has a *(multivariate) normal distribution* if there is a positive definite (hence nonsingular), symmetric, $n \times n$ matrix C and an n-dimensional vector m such that the probability density for $\boldsymbol{X}$ at the outcome vector x is given by

$$(2\pi)^{-n/2}|C|^{-1/2} \exp\left[-\tfrac{1}{2}(x - m)'C^{-1}(x - m)\right]$$

for all n-dimensional vectors x.

For example, if this is a good definition of the normal distribution for an n-dimensional random vector, then for $n = 1$ and 2 the definition should yield the familiar univariate and bivariate normal distributions, respectively (see reference [6], pp. 99, 198).

For $n = 1$, $(x - m) = (x - m)$, and the positive definite matrix C is a positive number, say, c. Thus, the density is

$$(2\pi)^{-1/2} c^{-1/2} \exp\left[-\tfrac{1}{2}(x-m)\frac{1}{c}(x-m) \right]$$

$$= \frac{1}{\sqrt{2\pi c}} \exp\left[-\tfrac{1}{2}\frac{(x-m)^2}{c} \right].$$

For $n = 1$, we see that m is the expected value of X and c is the variance of X. For n dimensions this generalizes as shown in Rule 18.

Rule 18. In Definition 27, $m = E[X]$ and $C = E[(X - E[X])\,(X - E[X])']$, the covariance matrix of X.

For example, for $n = 2$ in Definition 27, we have

$$(x -)' = m(x_1 - m_1, x_2 - m_2),$$

$$C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix},$$

where $c_{12} = c_{21}$, so that C is symmetric, and $c_{11} > 0$ and $c_{11}c_{22} - c_{12}c_{21} > 0$, so that C is positive definite.

$$|C| = c_{11}c_{22} - c_{12}c_{21}$$

and

$$C^{-1} = \frac{1}{|C|}\begin{bmatrix} c_{22} & -c_{21} \\ -c_{12} & c_{11} \end{bmatrix}.$$

Thus, the density at x is

$$(2\pi)^{-1}(c_{11}c_{22} - c_{12}^2)^{-1/2} \exp\left\{ -\tfrac{1}{2}(x -)'m\frac{1}{|C|}\begin{bmatrix} c_{22} & -c_{21} \\ -c_{12} & c_{11} \end{bmatrix}(x-)m \right\}$$

$$= \frac{1}{2\pi\sqrt{c_{11}c_{22} - c_{12}^2}} \exp\left\{ \left[\frac{-1}{2|C|} \right] [\, c_{22}(x_1 - m_1)^2 \right.$$

$$\left. - 2c_{12}(x_1 - m_1)(x_2 - m_2) + c_{11}(x_2 - m_2)^2] \right\}.$$

According to Rule 18, we may write

$$\sigma_1 = \sqrt{c_{11}}, \qquad \sigma_2 = \sqrt{c_{22}}, \qquad \rho = \frac{c_{12}}{\sqrt{c_{11}c_{22}}},$$

and the density as

$$\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-m_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-m_1}{\sigma_1}\right)\left(\frac{x_2-m_2}{\sigma_2}\right)+\left(\frac{x_2-m_2}{\sigma_2}\right)^2\right]\right\},$$

which is the formula in reference [6] on page 198. Thus the one-dimensional and two-dimensional normal distributions are the univariate and bivariate normal distributions (familiar from the Part II Syllabus).

The symmetry of the bivariate normal density implies that the marginal distributions and conditional distributions are univariate normal distributions (see reference [6], p. 199). This "hereditary" property is also possessed by the n-dimensional normal distribution; that is, the marginal distribution of any k of n normally distributed random variables is a k-dimensional normal distribution, and the conditional distribution of any k of n normally distributed random variables, given the other $n-k$ variables, is a k-dimensional normal distribution. The formulas to make this statement more precise are given in the next paragraph.

Let Y be an n-dimensional random vector with a normal distribution, and let m and C denote the expected value vector and covariance matrix for this distribution. Now, let Y, m, and C be partitioned; that is,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_k \\ \hline Y_{k+1} \\ \vdots \\ Y_n \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \} = {}_1Y \\ \\ \\ \\ \} = {}_2Y \end{matrix} \qquad m = \begin{bmatrix} m_1 \\ m_2 \\ \vdots \\ m_k \\ \hline m_{k+1} \\ \vdots \\ m_n \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \} = {}_1m \\ \\ \\ \\ \} = {}_2m \end{matrix}$$

$$C = \underbrace{\left[\begin{array}{c:c} C_{11} & C_{12} \\ \hdashline C_{21} & C_{22} \end{array}\right]}_{\begin{matrix} k \text{ cols.} \end{matrix} \quad \underbrace{\phantom{C_{12}}}_{n-k \text{ cols.}}} \begin{matrix} \} k \text{ rows} \\ \} n-k \text{ rows} \end{matrix}.$$

In this notation, ${}_1m$ and C_{11} (${}_2m$ and C_{22}) are the expected value and the covariance matrix, respectively, of ${}_1Y({}_2Y)$.

Rule 19. The marginal distribution of ${}_1Y$ is the k-dimensional normal distribution with expected value vector ${}_1m$ and covariance matrix C_{11}.

Rule 20. The conditional distribution of ${}_1Y$, given that ${}_2Y = {}_2y$, is the k-dimensional normal distribution with expected value vector

$$ {}_1m + C_{12}C_{22}^{-1}({}_2y - {}_2m) $$

and covariance matrix

$$ C_{11} - C_{12}C_{22}^{-1}C_{21} . $$

For example, to apply Rule 20 to the bivariate distribution, we may set $n = 2$, $k = 1$, and

$$ C = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}. $$

Then $Y_1(= {}_1Y)$, given $Y_2 = y_2$, has a univariate normal distribution with expected value

$$ m_1 + (\rho\sigma_1\sigma_2)(\sigma_2^2)^{-1}(y_2 - m_2) = m_1 + \rho\frac{\sigma_1}{\sigma_2}(y_2 - m_2), $$

the familiar linear regression equation, and variance (1-dimensional covariance matrix)

$$ \sigma_1^2 - (\rho\sigma_1\sigma_2)(\sigma_2^2)^{-1}(\rho\sigma_1\sigma_2) = \sigma_1^2(1 - \rho^2) . $$

For example, let Z_1, Z_2, Z_3, Z_4, and Z_5 be five random variables with expected values μ_1, μ_2, μ_3, μ_4, and μ_5; variances 1, 1, 1, 1, and 1; covariances $\sigma_{12} = \sigma_{23} = \sigma_{34} = \sigma_{45} = \rho$, $\sigma_{13} = \sigma_{24} = \sigma_{35} = \rho^2$, $\sigma_{14} = \sigma_{25} = \rho^3$, and $\sigma_{15} = \rho^4$; and a 5-dimensional normal distribution. Let us calculate the conditional distribution of Z_3, given $Z_1 = z_1$, $Z_2 = z_2$, $Z_4 = z_4$, and $Z_5 = z_5$.

To put our problem into the formulas of Rule 20, we let $n = 5$, $k = 1$,

$$ Y = \begin{bmatrix} Z_3 \\ \hline Z_1 \\ Z_2 \\ Z_4 \\ Z_5 \end{bmatrix} \qquad m = \begin{bmatrix} \mu_3 \\ \hline \mu_1 \\ \mu_2 \\ \mu_4 \\ \mu_5 \end{bmatrix} \qquad C = \left[\begin{array}{c|cccc} 1 & \rho^2 & \rho & \rho & \rho^2 \\ \hline \rho^2 & 1 & \rho & \rho^3 & \rho^4 \\ \rho & \rho & 1 & \rho^2 & \rho^3 \\ \rho & \rho^3 & \rho^2 & 1 & \rho \\ \rho^2 & \rho^4 & \rho^3 & \rho & 1 \end{array}\right]. $$

We may compute as follows:

$$ C_{22}^{-1} = \frac{1}{1-\rho^4}\begin{bmatrix} 1+\rho^2 & -\rho(1+\rho^2) & 0 & 0 \\ -\rho(1+\rho^2) & 1+\rho^2+\rho^4 & -\rho^2 & 0 \\ 0 & -\rho^2 & 1+\rho^2+\rho^4 & -\rho(1+\rho^2) \\ 0 & 0 & -\rho(1+\rho^2) & 1+\rho^2 \end{bmatrix}; $$

$$C_{12}C_{22}^{-1} = \rho[\rho\ 1\ 1\ \rho]C_{22}^{-1} = \frac{\rho}{1+\rho^2}[0\ 1\ 1\ 0];$$

$$C_{12}C_{22}^{-1}C_{21} = \frac{\rho}{1+\rho^2}[0\ 1\ 1\ 0]\begin{bmatrix}\rho^2\\ \rho\\ \rho\\ \rho^2\end{bmatrix} = \frac{2\rho^2}{1+\rho^2}.$$

Thus, the conditional distribution of Z_3, given $Z_i = z_i$, $i = 1, 2, 4, 5$, is the univariate normal distribution with expected value

$$\mu_3 + \frac{\rho}{1+\rho^2}[0\ 1\ 1\ 0]\begin{bmatrix}z_1-\mu_1\\ z_2-\mu_2\\ z_4-\mu_4\\ z_5-\mu_5\end{bmatrix} = \mu_3 + \frac{\rho}{1+\rho^2}(z_2-\mu_2+z_4-\mu_4)$$

and variance

$$1 - \frac{2\rho^2}{1+\rho^2} = \frac{1-\rho^2}{1+\rho^2}.$$

REFERENCES

1. CAMP, K. "New Possibilities in Graduation," *TSA*, VII (1955), 6.
2. CODY, D. D. "Actuarial Note: The Standard Deviation in the Rate of Mortality by Amounts," *TASA*, XLII (1941), 69.
3. ELPHINSTONE, M. D. W. "Summation and Some Other Methods of Graduation: The Foundations of Theory," *TFA*, XX (1951), 15.
4. GOOD, I. J. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods.* ("Research Monograph," No. 30.) Cambridge, Mass.: M.I.T. Press, 1965.
5. HENDERSON, R. *Mathematical Theory of Graduation.* ("Actuarial Studies," No. 4.) New York: Actuarial Society of America, 1938.
6. HOEL, P. G. *Introduction to Mathematical Statistics.* 3d ed. New York: John Wiley & Sons, 1962.
7. JONES, D. A. "Bayesian Statistics," *TSA*, XVII (1965), 33.
8. KIMELDORF, G. S. *Applications of Bayesian Statistics to Actuarial Graduation.* (*Dissertation Abstracts*, Vol. XXVII.) Ann Arbor: University of Michigan, 1966. (University Microfilms; available through University of Michigan Inter-Library Loan.)
9. KING, G. Discussion of "The Graphic Method of Adjusting Mortality Tables," by T. B. SPRAGUE, *JIA*, XXVI (1887), 114.
10. MILLER, M. D. *Elements of Graduation.* ("Actuarial Monographs," No. 1.) New York: Actuarial Society of America and American Institute of Actuaries, 1946.
11. ROSENTHAL, I. "Limits of Retention for Ordinary Life Insurance," *RAIA*, XXXVI (1947), 6.

12. *TSA 1962 Reports of Mortality and Morbidity Experience*, 1963.
13. *TSA 1963 Reports of Mortality and Morbidity Experience*, 1964.
14. WHITTAKER, E. T., and ROBINSON, G. *The Calculus of Observations*. 4th ed. London and Glasgow: Blackie and Son, Ltd., 1944.
15. WOLFENDEN, H. H. *The Fundamental Principles of Mathematical Statistics*. Toronto: Macmillan, 1942.

DISCUSSION OF PRECEDING PAPER

FRANK S. IRISH:

The work of these authors, in bringing the newer statistical methods to bear on the problem of dealing with "prior data," is, of course, closely related to the credibility methods that have been developed over the years, particularly by casualty actuaries. I note in particular that equation (8) of the paper sets forth the graduation equation in the form of a credibility relationship.

The similarity becomes even clearer if we modify equation (8) by eliminating the nondiagonal elements of the matrix A; that is to say, eliminating, for the moment, the factor of smoothness from consideration. The equation then reduces to a form that is reminiscent of the most frequently used credibility formula. There is, however, one important difference—the formula derived in this way has the exposure in the position that is usually occupied by expected deaths:

$$v = m + (u - m)\left(1 + \frac{b}{a}\right)^{-1},$$

where $a = p^2 = 0.0002\ m$; $b = m(1 - m)/n$; n = lives exposed; therefore

$$v \doteq m + (u - m)\left(\frac{n}{5{,}000 + n}\right).$$

This suggests that the definition of the matrix elements might possibly be changed to create a greater consistency with credibility methods. The difficulty can be seen better by examining the results of the actual graduation example in the paper, where the methods used give much greater credibility to Groups 1 and 2 (although there is less than one expected death in each group) than to Group 12 (where there are five expected deaths). I think that the results could be improved if p were made proportional to m rather than $\sqrt{m}$; this would also make equation (8) consistent with the traditional credibility formula.

This sort of modification of equation (8) also suggests that the general methods of the authors might well be used to combine the credibility approach with a less complex method of measuring smoothness (a simple example would be combining the credibility approach with a linear compound formula), the goal being a method that would be as suitable for a desk calculator as for a large-scale computer.

JAMES C. HICKMAN:

This is an extraordinary paper. The potential applicability of Bayesian statistics to actuarial science was established by Professor Jones's earlier paper. What had been a stimulating possibility has now become a reality, for the authors have provided us with a delightful blend of a new and novel approach to a classical actuarial problem, a primer on the mathematics needed to use the new approach, and a convincing numerical example.

In equation (2) the authors exhibit a density function for the vector of observed rates, $\boldsymbol{u}$. They have assumed that the U_i's ($i = 1, 2, \ldots, n$) are each normally, but not identically, distributed and that they are mutually statistically independent. This assumption is certainly in accordance with actuarial tradition. For example, Miller's (*Elements of Graduation*) exposition on adjusted average graduation methods makes use of an independence assumption. This model would certainly be appropriate for the usual snapshot look at mortality experience used in estimating survival functions. However, it is interesting to determine the impact on the new method of a longer period of study in which a cohort of lives might be used to estimate several successive conditional mortality probabilities. In this case, it appears that the matrix B would no longer be a diagonal matrix. The elements of B in a belt along the principal diagonal would no longer be zero but would reflect the correlation between the observed rates for successive ages. Fortunately, the development leading to equation (6), the equation that defines the vector of graduated values, does not depend on B's being diagonal. Therefore, although the computation would be more laborious since an additional nontrivial matrix inversion would be involved, equation (6) stands as a satisfactory definition of the vector of graduated values even for the more complicated model, where observed values are not assumed to be statistically independent.

It seems of interest to carry the notion of dependence among the components of the vector $\boldsymbol{u}$ still further. For example, suppose that the experimental model involves observing a cohort of lives from birth to death. The objective of the study is to produce a generation-type mortality table. For such an experiment, a multinomial distribution is the appropriate model for the distribution of the number of deaths at each age. We let $\boldsymbol{D}' = (D_1, D_2, \ldots, D_n)$ be a random vector, where each D_i, $i = 1, 2, \ldots, n$, is a random variable associated with the number of deaths, from an initial group of k lives, that takes place in age group i. Note that

$D_n = k - D_1 - \ldots - D_{n-1}$ and thus is in fact determined by the other $n - 1$ variables.

For this type of "marching through life" experiment, the multinomial distribution, appropriate for the distribution of the vector D, has a mass function given by

$$p_{D|W}(d|w) = \left[k!\Big/\Big(\prod_1^n d_i!\Big)\right]\prod_{i=1}^n w_i^{d_i}, \quad 0 \le d_i, \quad i = 1, 2, \ldots, n,$$

$$\sum_{i=1}^n d_i = k, \qquad 0 < w_i, \qquad \text{and} \qquad \sum_1^n w_i = 1.$$

We now quantify our prior knowledge about the distribution of W, the vector of mortality probabilities, in the form of a Dirichlet distribution. The mass function for the prior distribution is given by

$$p_W(w) = \left[\Gamma\Big(\sum_1^n a_i\Big)\Big/\prod_1^n \Gamma(a_i)\right]\prod_1^n w_i^{a_i-1},$$

$$0 < w_i, \qquad \sum_1^n w_i = 1, \qquad 0 < a_i.$$

This distribution was chosen because it is conjugate to the multinomial in the sense that the posterior distribution of W is once again a Dirichlet distribution. In fixing the prior distribution, we make use of the fact that

$$E_W(W_i) = a_i\Big/\sum_1^n a_i, \qquad i = 1, 2, \ldots, n.$$

Thus a graduator might initially fix T, where

$$T = \sum_1^n a_i.$$

The parameter T measures the strength of the prior knowledge about the distribution of W. This interpretation of T will become more apparent later in the development. Large values of T are associated with compact prior distributions, and small values of T are associated with diffused distributions. Then the graduator may determine the parameters of the prior distribution by setting $m_i = a_i/T$, $i = 1, 2, \ldots, n$. The m's are, as in the paper, the prior means of the w's.

The posterior distribution of W, after the k lives have been followed

from their birth until they enter age group n, will have a mass function given by

$$p_{W|d}(w \mid d) = \left[\Gamma\left(\sum_1^n a_i + k\right) \Big/ \prod_1^n \Gamma(a_i + d_i)\right] \prod_1^n w_i^{d_i + a_i},$$

$$\sum_1^n d_i = k, \qquad 0 < w_i, \qquad \text{and} \qquad \sum_1^n w_i = 1.$$

If we follow the authors and use the vector of mean of the posterior distribution as our vector of graduated values, we have

$$v_i = E_{W|d}(W_i \mid d) = (a_i + d_i)/(T + k)$$
$$= (d_i/k)[k/(T+k)] + (a_i/T)[T/(T+k)].$$

Note that d_i/k is the estimate of mortality probability in age group i obtained from the experiment and that a_i/T is the estimate (in the expected value sense) obtained from the prior distribution. The graduated value is a weighted average of these experimental and prior values with the weight shifting to the experimental value as k increases. The blending of prior opinion and experimental results is rather obvious. This result, when $n = 2$, appeared in the estimation example (Example E) in Jones's earlier paper.

In their discussion of equations (11), (14), and (15), the authors observe that the Whittaker method of graduation involves finding a vector v that minimizes a particular quadratic form. Table 1 of their paper exhibits several such solutions for a particular set of data and various values of the parameter h.

It occurred to Mr. Brian Harvey, a graduate student in the Department of Statistics, University of Iowa, that the problem of minimizing a positive definite quadratic form is a common problem in operations research. In fact, he was acquainted with a computer program that employed the Shetty algorithm and which would find a vector v such that the positive definite quadratic form $u'Bu$ is a minimum subject to the restriction that $Av \geq b$, where B is a positive definite n by n matrix, u and v are column vectors with n components, A is an m by n matrix, and b is an m row vector. Harvey took B to be the matrix of the quadratic form associated with the Whittaker Type B graduation formula found in Section IV of the paper, $A = I$, where I is the 13 by 13 identity matrix, and $b = 0$, the 13 row zero vector. In a word, he sought to perform a traditional Whittaker graduation, with the additional restriction that the graduated values be nonnegative, by using a quadratic programming

computer routine that permitted the imposition of linear restrictions. The parameter h was taken as 100, and it was hoped to overcome the distressing negative values found for the same h in Table 1. The numerical results are shown in the accompanying tabulation.

In this example prior opinion influenced the choice of the Whittaker type quadratic loss function that was to be minimized and the linear restrictions imposed on the solution vector. Other plausible linear restrictions could have been imposed equally well.

THE QUADRATIC PROGRAMMING, WITH LINEAR RESTRICTION, SOLUTION OF A GRADUATION EXAMPLE FOUND IN THE KIMELDORF-JONES PAPER

($v_i \times 10^3$ Is Recorded)

i	Quadratic Program ($h=100$)	Whittaker Table 1 ($h=100$)	Bayesian Table 2
1.....	0.00	−0.40	0.22
2.....	0.09	−0.13	0.29
3.....	0.27	0.17	0.46
4.....	0.61	0.59	0.59
5.....	1.15	1.16	1.10
6.....	1.95	1.96	1.81
7.....	3.20	3.21	2.87
8.....	4.50	4.51	4.08
9.....	5.64	5.64	5.41
10.....	6.51	6.51	7.21
11.....	7.11	7.11	12.49
12.....	7.70	7.69	20.03
13.....	8.28	8.28	31.81

T. N. E. GREVILLE:

I regard this paper as a most important contribution to the study of graduation. It has long and often been asserted by statistically minded people that graduation ought to be based on probabilistic considerations; here at last is a paper that does this, not in a half-hearted way, as in the original derivation of the Whittaker method, but in a thoroughgoing fashion. The proposed method deserves careful study and analysis, as well as testing in practical situations.

My chief criticism of the paper is that the authors, in their understandable zeal to promote their own method, seem to sell the Whittaker method very short. When they say, "For application of the Whittaker method to such scant data, an artful user of the method would make certain adjustments either to the observed rates or to the graduated

rates to bring them into conformity with his opinion," it seems to me that they greatly underestimate the remarkable versatility of the Whittaker method.

Camp showed in 1950, in his privately printed manual, *The Whittaker-Henderson Graduation Processes*, how to constrain a specified order of differences of the graduated values toward a geometric trend. Using the observed data of the numerical example in the paper, I have tried using a Whittaker method that constrains the values themselves toward a geometric trend, as this leads to a second-order difference equation, like the Whittaker method applied in the paper. This is accomplished by taking as the measure of roughness

$$\sum_{i=1}^{n-1} (v_{i+1} - r\,v_i)^2,$$

where r is a suitably chosen constant. The resulting graduated values, taking $r = 1.5$ and $h = 10$ and 100, are shown in the accompanying tabulation. Except for the fact that the total expected mortality is somewhat below the actual (which could easily be remedied by adding an appropriate constant to the rates of mortality at all ages), I think that most actuaries would regard either of these as a better graduation of the observed data than the Bayesian graduation in the paper. Taking into

VARIOUS GRADUATIONS OF ILLUSTRATIVE DATA

i	Age Group	$u_i \times 10^3$	Graduated Rates $v_i \times 10^3$			Prior Means
			Whittaker Method*		Bayesian Method	
			$h=10$	$h=100$		
1...	10–14	0.00	0.04	0.16	0.22	0.33
2...	15–19	0.00	0.09	0.26	0.29	0.41
3...	20–24	0.04	0.24	0.42	0.46	0.58
4...	25–29	0.80	0.73	0.71	0.59	0.67
5...	30–34	1.32	1.19	1.07	1.10	1.18
6...	35–39	1.11	1.41	1.53	1.81	1.91
7...	40–44	3.41	3.20	2.42	2.87	2.81
8...	45–49	4.70	4.46	3.24	4.08	3.95
9...	50–54	6.01	5.51	4.00	5.41	5.33
10...	55–59	7.72	6.36	4.97	7.21	7.27
11...	60–64	4.15	6.64	6.45	12.49	12.80
12...	65–69	5.93	9.17	9.58	20.03	20.50
13...	70 and over	9.74	13.64	14.36	31.81	32.39

* Constrained toward geometric trend with $r = 1.5$.

account the offhand manner in which I chose the value of r and the values of h, I think that both graduations are remarkably successful. In both cases, the values are always positive, and the sequence is always increasing.

Though they do not actually say so, the authors leave the impression with the reader that their method is, in some sense, a generalization of the Whittaker method. It is my opinion that, from one point of view, the reverse is true. In 1957 (*Journal of the Society for Industrial and Applied Mathematics*, V, 150), I showed that a fairly general form of the Whittaker graduation method is expressed by the matrix equation

$$v = (I + W^{-1}G)^{-1}u ,$$

where W is a positive definite matrix and G a positive semidefinite matrix. This equation was obtained by minimizing the quadratic form

$$(v - u)'W(v - u) + v'Gv .$$

If a Whittaker graduation were performed, not on the observed values themselves but on their deviations from those of a standard table (represented by the elements of the vector m), and if W were replaced by B^{-1} and G by A^{-1}, the above equation would become identical with equation (8) of the paper. In fact, W is usually a diagonal matrix whose diagonal elements are weights assigned to the respective deviations of graduated from observed values and bears a strong resemblance to the matrix B^{-1} of the paper. G, however, is permitted to be singular, and, when it is, it could not be written as A^{-1}.

As the class of positive semidefinite matrices includes the positive definite matrices, the matrix G of the above equation can be chosen from a broader class of matrices than the matrix A^{-1} of the paper. In this sense the Whittaker method is more general than the Bayesian method. Every Bayesian graduation could be regarded as a Whittaker graduation (for some choice of the above matrices W and G), but the converse is not true.

The option of choosing G as a singular matrix (not available in the Bayesian method) is an important property of the Whittaker method, as it permits a class of vectors u to be left unchanged by the graduation. The vectors left unchanged are precisely the vectors annihilated by G. In the usual (third-difference) formulation, these are the vectors whose elements are the corresponding ordinates of some second-degree polynomial. However, as Camp has shown, they could also be, for example, vectors such that the second differences of successive elements are in geometric progression with common ratio c, or, in other words, vectors whose ith element is $A + Hi + Bc^i$ for some constants A, H, and B. In

the Bayesian method the vector u of observed values is left unchanged by the graduation only if it is identical with the vector m of prior means.

It would seem offhand that in the Bayesian method the user has more difficulty in exercising control over the graduation process than in the usual Whittaker method. Neither term of the quadratic form (9) that is minimized by the Bayesian graduation directly measures the roughness of the graduated values, and it would appear from examination of this quadratic form that the only way to ensure smoothness is to take w close to m. This is borne out by the numerical example in the paper, in which the results of the Bayesian graduation look more like a graduation of the prior means than of the observed values.

One wonders if a similar method could be devised that would permit the user to express his prior opinion as to the general form and shape of the curve representing the data, without having to commit himself to a specific set of numerical values. As a practical matter, the Whittaker method (in its broadest interpretation) seems to do this very successfully indeed, and perhaps all that is needed is to furnish it with a probabilistic justification.

HARWOOD ROSSER:

A number of years ago the Education and Examination Committee, aided by a syllabus change and by the T.N.E.C. investigation, gave me a lifelong inferiority complex on the subject of statistics. Not long thereafter, probability and statistics became so intermingled in the syllabus that whenever an actuarial student had a question on a Part 2 problem, I always ducked, although I regarded probability as an easy subject. Now come Messrs. Kimeldorf and Jones to extend my self-doubt to mathematical graduation—a subject about which I had regarded myself as reasonably knowledgeable.

Despite the damage to my ego and although, at first reading, there is much that I do not completely follow, I find this a very interesting paper. The main thesis is that whatever "prior opinion" the graduator has should be systematically applied. They suggest a fairly elaborate scheme for doing this, which would have been completely impractical in precomputer days. It is a very sophisticated cousin of graduation by reference.

This reviewer suspects that it will be quite some time before Bayesian graduation is widely used by actuaries, regardless of its theoretical justification and in spite of the offer of a Fortran II program. This is not a detraction of the expository powers of the authors, although there are some of us who would have preferred to see some of the intermediate figures in the numerical example. Rather, this suspicion is based on the fact

that, although nearly twenty years have passed since the appearance of Greville's excellent papers, with tables, in *RAIA*, Volumes XXXVI and XXXVII, giving a method of graduation using the linear-compound, or multiplier, approach, I have yet to see it utilized by anyone other than me. Yet an understanding of Tchebycheff polynomials is not required in order to use Greville's tables; the computations can be done easily on either desk or electronic computer.

If his paper were required reading, the method he gives would be better known. But the questions that could be asked thereon would be either too easy or too difficult for examination purposes; that is, it would be almost impossible for the Part 5 committee to ascertain whether or not a candidate had sufficient knowledge of this method. To a lesser extent, this is true of the paper under discussion; but there would be fewer questions that were too easy. This is so, despite segregation into three appendixes of some of the more technical aspects of the subject.

For skimmers, I would recommend reading first Section IV, entitled "Examples." One could wish, however, that more examples had been given, especially in view of the statement that less than one minute of computer time was required for the single example shown. The Whittaker-Henderson graduations in Table 1, with five different values of h, are all deemed unsatisfactory. The implication is then strong that the one Bayesian graduation in Table 2 is the answer to an actuary's prayer. But surely there must be occasions when a graduator would have a "subsequent opinion" as well as a "prior opinion," that is, when he would wish to revise the latter, after testing the results of the graduation. It is standard advice to the student not to accept a first graduation blindly. Is Bayesian graduation to be exempted?

Nonetheless, this is an excellent addition to the lengthening series, from the University of Michigan, of adaptations of doctoral dissertations to an actuarial audience. (When I was of college age, it seemed inevitable that candidates for advanced degrees would run out of worthwhile thesis subjects. This contingency now seems more remote.) If, even after adaptation, the results impress us as highly technical, we may be looking in a mirror that shows us how we appear to the average citizen.

DONALD C. BAILLIE:

I have been very much impressed with the work that Dr. Jones has been doing, because it has long been my conviction that there is something about the so-called classical statistics of this century that just does not seem to fit with actuaries. I think that we are all familiar with that, and

I think that Jones has been very polite about it in his reference, in the early part of his paper, to the curriculum.

I did try to make up a little story that might be of some value to some of us here as to what the distinctions are all about—why the actuaries often do not like statistics as ordinarily taught. I think that the main reason is that actuaries are all basically Bayesian, whether they know it or not.

Now I imagined a situation the other evening with which I think most of us are familiar. An actuary has hired a bright young man who has been taught a good deal of statistics in a conventional statistics course. One of the first tasks he assigns to this young man is to have him make a mortality investigation; let us assume that something on the order of twenty thousand lives of mixed ages are involved over perhaps a matter of five years. The young man finally comes in with the data, and the actuary asks, "Well, how is the mortality?" The young statistician says, "Sir, the mortality is zero." The actuary asks, "How is that?" to which the student replies, "There are no death cards."

Now, what is the actuary's reaction? The statistician has taken his observations and he did not find any deaths. The actuary then replies, "Don't be silly; go back and find them." And that, I believe, is the Bayesian notion.

The actuary explains: "Look, you have 100,000 years of exposure, and, even at the rates that people are experiencing in some of our well-known insurance companies, you must have at least ten deaths, probably even two or three hundred or more, somewhere." He then adds that one of two things has happened—either the deaths have been left out of the program or, if those old-fashioned "square" packets of two or three hundred cards in a bunch were being used, one of them may have fallen off the table.

That is the contrast between Bayesian and the classically trained twentieth-century statistician who says, "I have observed 20,000 variates—X_1, X_2, and X_3—I have added them all up, and it comes to zero."

It may very well be that this is the first set of observations from the real world that the young man has actually made. It is even possible that he has studied statistics in the university without observing anything, not even an artificial experiment.

Well, some deaths are eventually located, and a graduation is to be performed. The young man does not know much about graduation, but he is given certain ideas associated with Gompertz and Makeham and goes off and graduates his data. When he comes back, the actuary asks, "Well, what have you got this time?" to which the young man replies,

"Sir, I've got a dandy graduation." The actuary says, "Good, what do you get for c?" When the young man replies, "1.55," the actuary says, "That's crazy. Are you sure these are humans? You have found mortality rates that are going up 50 per cent a year." The actuary then sends him back to reconsider the situation. As he leaves, the actuary adds, "Don't look at any c that is less than 1; in fact, don't let the machine waste any time on c's that are larger than 1.12 or less than 1.06." Let me say that the actuary is expressing the Bayesian point of view right there.

The weakness in connection with any ordinary mathematical statistics course is that you may be formally taught to put up blinders and forget almost everything known about the material being studied.

In conclusion, I would like to say that the sort of approach presented by Jones and Kimeldorf is long overdue, especially if actuaries are seriously going to make use of multivariate statistical techniques. It is also long overdue for some statisticians to realize that everything does not happen in a controlled laboratory.

Again, I extend my heartiest congratulations to the authors of this paper.

(AUTHORS' REVIEW OF DISCUSSION)

GEORGE S. KIMELDORF AND DONALD A. JONES:

We would like to thank Messrs. Irish, Hickman, Greville, Rosser, and Baillie for their thoughtful discussions of our paper. While Baillie's parable is somewhat extreme, it does focus attention on the interaction between prior knowledge and new data, which is the essence of statistical inference.

Irish outlines the close relationship provided by Bayesian statistics between graduation theory and credibility methods. While we have looked forward to utilizing this relationship in order to further the development of credibility theory, Irish uses it to apply credibility ideas to graduation. In particular, he shows that greater consistency between existing credibility ideas and Bayesian graduation would be achieved if the graduator used the relation $p_i = cm_i$ to approximate his actual prior standard deviations rather than the relation $p_i = c\sqrt{m_i}$ which we used in our example in Section IV. In theory, of course, there exists no functional relationship whatever between the prior means m_i and the prior standard deviations p_i, and each prior standard deviation should be elicited by honest introspection (perhaps with the aid of equation [28] of our paper). Because this procedure is tedious, it is often worthwhile to seek some formula for the p_i which would serve as a good approximation.

In this context, Irish's suggestion for the use of the formula $p_i = cm_i$ is based on the observation that many actuaries using credibility methods have found it to be an adequate approximation in the sense that they have been willing to accept its consequences. Column 8 of the accompanying tabulation presents the result of graduating our data of Section IV using Irish's suggestion with $c = 0.6$.

Hickman's discussion presents two valuable ideas. The first, an extension of the Bayesian method in which the observed rates are not assumed to be independent, is applied to graduating mortality rates among

SEVERAL GRADUATIONS OF ILLUSTRATIVE DATA

i	Age Groups	Expected Deaths	Exposures $N_i \times 10^{-6}$	Crude Rates $u_i \times 10^3$	Graduated Rates $v_i \times 10^3$				Basic Table $m_i \times 10^3$
					Greville		Irish	Kimeldorf-Jones	
					$h = 10$	$h = 100$			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1.....	10–14	3.84	11.64	0.00	0.04	0.16	0.25	0.22	0.33
2.....	15–19	5.41	13.19	0.00	0.09	0.26	0.32	0.29	0.41
3.....	20–24	13.80	23.80	0.04	0.24	0.41	0.46	0.46	0.58
4.....	25–29	23.41	34.94	0.80	0.73	0.71	0.59	0.59	0.67
5.....	30–34	60.91	51.62	1.32	1.19	1.07	1.08	1.10	1.18
6.....	35–39	125.74	65.83	1.11	1.41	1.53	1.70	1.81	1.91
7.....	40–44	205.75	73.22	3.41	3.20	2.42	3.04	2.87	2.81
8.....	45–49	239.65	60.67	4.70	4.46	3.24	4.41	4.08	3.95
9.....	50–54	179.09	33.60	6.01	5.51	4.00	5.62	5.41	5.33
10.....	55–59	131.73	18.12	7.72	6.36	4.97	6.68	7.21	7.27
11.....	60–64	89.34	6.98	4.15	6.64	6.45	8.65	12.49	12.80
12.....	65–69	37.92	1.85	5.93	9.17	9.58	12.54	20.03	20.50
13.....	70 and over	10.04	0.31	9.74	13.64	14.36	19.97	31.81	32.39

a cohort of lives observed from birth to death. His second idea generalizes Whittaker's method by minimizing a positive definite quadratic form subject to certain linear inequality constraints. Both of Hickman's ideas offer exciting possibilities for further development and application to actuarial problems.

An interesting and very practical question is the one raised by Rosser: How does the Bayesian theory provide for testing a graduation? In reply, we would advise the graduator to test his prior opinion before graduating the crude data. A graduation can be unacceptable only when the results conflict with some aspect of the graduator's opinion. Hence, if all aspects of opinion were actually expressed as input to the graduation process, the result would in fact represent the graduator's true posterior opinion,

and no subsequent testing would be necessary. As a practical matter, however, an inexperienced graduator or one who is using a new method may experience some difficulty in quantifying his prior opinion precisely. In such cases, we would suggest that the graduator, prior to graduating the given data, test his prior opinion by examining the results of graduating several sets of hypothetical data.

Greville's chief criticism, that we "sell the Whittaker method very short," and his quarrel with our "impression of Bayesian graduation as a generalization of Whittaker's method" are based upon a difference between his definition of the Whittaker method and ours. Greville defines the Whittaker method as one which derives the graduated rates v from the ungraduated rates u by minimizing the quadratic form

$$(v - u)'W(v - u) + (v - m)'G(v - m) ,$$

where W is a positive definite matrix and G is a positive semidefinite matrix, while in our paper we are thinking of Whittaker's method in terms of the general mixed difference Type B method. We should have pointed out in our paper the similarity between our equation (8) and an equation previously published by Greville.

We did not intend to leave the reader with the impression that "in the Bayesian method the user has more difficulty in exercising control over the graduation process than in the usual Whittaker method [in that] the only way to ensure smoothness is to take w close to m." The graduated rates will be smooth but not necessarily close to the prior means m if the following conditions hold: (1) the prior means are smooth; (2) the prior correlation coefficients are large (i.e., close to 1); and (3) the prior standard deviations are large as compared with the standard deviations of the observations. In graduations of extensive data, these conditions will usually prevail.

Greville's assertion that "the option of choosing G as a singular matrix . . . is an important property of the Whittaker method" demonstrates one essential difference between Bayesian and classical procedures. He bases the importance of G's singularity on the existence of a nontrivial set of vectors which are left unchanged by such a graduation process. We recognize this as a mathematically elegant property but not one particularly relevant for a graduation process. From the Bayesian viewpoint the singularity of G reflects a prior distribution which is uniform on this set of vectors and on each of its translates. We believe that in the majority of applications of graduation the graduator's prior density function is unimodal rather than constant on such sets and hence the only vectors which should be left invariant under a graduation process are

those sufficiently close (in the sense of numerical accuracy) to his modal vector.

As an example Greville chooses as his measure of "roughness" the expression

$$\sum_{i=1}^{12} (v_{i+1} - r v_i)^2,$$

which is zero if and only if the rates v_i satisfy a relation of the form $v_i = ar^i$ for $i = 1, 2, \ldots, 13$. This measure is derived from his prior belief that mortality rates exhibit a geometric trend. Moreover, he is willing to commit himself to the value $r = 1.5$ to describe the "shape" of the mortality curve. With this measure of "roughness," which corresponds to a matrix G which is singular, *every* vector of the form $v = [v_1, v_2, \ldots, v_{13}]'$, where $v_i = a(1.5)^i$ is left invariant by Greville's graduation *regardless of the value of* a. From the Bayesian point of view the adoption of this singular "roughness" matrix, which corresponds to a degenerate prior probability distribution, would imply that *every* such vector was equally probable in his prior opinion *regardless of the value of* a, thus ignoring the facts that a must certainly be restricted to the narrow range $0 < a < (1.5)^{-13} \approx 0.0051$ (so that the mortality rates are between 0 and 1), that a almost certainly satisfies $0 < a < (0.1)(1.5)^{-13} \approx 0.00051$, and that a probably (based on previous mortality studies for similar lives) satisfies $0.0001 < a < 0.0003$.

It seems unrealistic to us to graduate under the assumption that the value of r can be stated precisely while denying any knowledge whatever about the value of a. Hence we do not share Greville's interest in graduation methods in which prior opinion determines the shape of the curve but is completely silent about the level of the curve, as would be implied by a singular "roughness" matrix.

The tabulation shown on page 124 displays several graduations of the data presented in Section IV. The crude rates in column 5 are the seventh policy-year experience of policies issued in the single year 1955 on standard medically examined female lives, while the prior means given in column 10 are the graduated rates for the seventh policy-year experience for like issues for the years 1949 through 1954. We view graduation not merely as smoothing but as the more general process of estimating the true rates which actually prevail in the population, these estimates to be consistent both with the observed data and the graduator's experience, judgment, and knowledge. Hence we believe that a graduation should be judged not only on the basis of columns 4 and 5 but also (especially for such scant data as these) on the basis of column 10, as

well as all other relevant information. In this context we cannot agree that either of Greville's graduations is preferable to our Bayesian graduation. For $h = 10$, his initial values are too low; for $h = 100$, his graduated rates are significantly below both the crude rates and the Basic Table rates for age groups 7–10, which constitute 67 per cent of the expected deaths (Basic Table) and 47 per cent of the exposures. The ability of the Bayesian method to make explicit and maximal use of prior information is one of its chief advantages over the Whittaker method as defined by Greville.