TRANSACTIONS OF SOCIETY OF ACTUARIES
1991 VOL. 43

INFORMATION THEORETIC APPROACH TO ACTUARIAL SCIENCE: A UNIFICATION AND EXTENSION OF RELEVANT THEORY AND APPLICATIONS

PATRICK L. BROCKETT*

ABSTRACT

This paper presents a unifying non-Bayesian statistical method for incorporating prior information into the determination of probability distributions (and other finite measures) and develops the statistical properties of this method together with actuarial applications. This method is shown to be useful in several areas in actuarial modeling, including:

- How to choose between competing models for a stochastic phenomenon under investigation.
- How to use a standard distribution (mortality, loss, duration, and so on) and client data to obtain an appropriate distribution that is tailored to characteristics of the client.
- How to adjust mortality tables in a statistically valid manner to obtain exactly certain known or assumed individual characteristics, while simultaneously developing a table that is as close as possible to a given standard mortality table.
- How to graduate or smooth observed insurance data to obtain smoothed estimates that are as close as possible to the observed data, subject to convexity and smoothness constraints.
- How to incorporate monotonicity constraints into a life table graduation. This graduation technique can encompass either univariate or multivariate mortality tables with equal facility.

These goals are accomplished by means of constrained information theoretic techniques. They are shown to provide a unified philosophical approach with a firm statistical foundation capable of being applied to many problems currently addressed separately or in an ad-hoc fashion in actuarial education. Moreover, several other probabilistic models of relevance to actuarial analysis (logit or logistic regression techniques, loglinear and multiplicative models, certain Bayesian techniques, and so on) are shown to arise

* Dr. Brockett, not a member of the Society, is the Joseph H. Blades Professor of Insurance at the University of Texas at Austin.

naturally as a *consequence* of this information theoretic formulation, and common mortality table assumptions, such as constant force of mortality and uniform distribution of deaths assumptions, are given new interpretation and justification in terms of information theoretic concepts.

1. INTRODUCTION

A substantial portion of actuarial training is designed to assist in the process of decision-making under uncertainty. For example, in modeling individual and insurance company choice behavior, the outcomes are often evaluated in terms of the expected utility of the alternative outcomes (for example, Bowers et al. [12], Hurlimann [43], Borch [9]–[11], or Briys [13]). Even the determination of net single premiums involves the calculation of the expectations of loss for the phenomenon under investigation. Usually in the actuarial literature such problems are approached by separately determining (or postulating) the pertinent probability distribution, and then performing the requisite calculation of expected values, statistical assessment of multivariate probabilistic structure, or testing of the appropriate hypotheses. Sometimes stochastic models are developed in an ad-hoc manner (examples are certain adjustments to a standard mortality table, visual smoothing of graduated series, or certain finite difference or polynomial regression techniques used in ratemaking), and the approaches used can vary substantially from one application to another depending upon the particular training of the actuary.

A fundamental principle or approach can be used to show the actuarial science researcher how to: (1) integrate the distribution determination with the analysis method, (2) draw the modeling formulation *as a consequence* of this fundamental principle, while (3) providing a unified framework for statistically testing the significance of the obtained relationships. The approach should be sufficiently flexible to incorporate (or generalize) the frequently encountered quantitative actuarial science methods as special cases and, hopefully, should even allow the decision-making process to proceed when (as often happens in actual actuarial applications) there is only limited, partial, or perhaps even conflicting, information about the random variables involved. This partial information may, for example, consist of a sample of empirical data that may, however, differ in qualitative ways (for example, non-monotonicity of observed mortality rates, or non-unimodality of loss

distribution data, and so on) from what is known about the stochastic structure. Or it could consist of having forecast or exogenously determined expected values of certain functions of the random variables (for example, interest rate forecasts, known or published percentiles of a mortality or loss distribution, and so on). Moreover, this desired fundamental philosophical approach should be simple and intuitive so that it is easily understood by both researchers and practitioners. We show that statistical information theory provides such a fundamental actuarial principle that addresses these goals.

We proceed as follows: First, we define and review the mathematical foundations of information theory as needed in this paper. Next we show how the information theoretic method can be used to guide the actuarial science researcher in selecting statistical models for analysis when the true underlying distributions are unknown. The resulting criterion, known as Akaike's Information Criterion (AIC), presents a unified approach to the solution of numerous different actuarial problems (for example, insurance models for ratemaking, interest and bond rate process modeling, modeling changing demographic characteristics, and so on).

We next relate the information theoretic method to Bayes' Theorem. Of course the Bayesian approach has been widely presented in actuarial applications (for example, credibility theory, graduation of mortality tables, economic forecasting, and so on), so this connection unifies the current research on Bayesian methods with that of the classical (frequentist) approach commonly studied by actuarial students. We then derive the canonical Minimum Discrimination Information (MDI) model and, from this stochastic model, develop each of several important actuarial science models (logit or logistic regression models, multiplicative models, loglinear models, and so on). These are shown to arise naturally in the information theoretic context, and in addition, we indicate how the MDI hypothesis-testing capability makes these models more useful for actuarial decision-making and assessment of statistical significance of relationships. For brevity, the mathematical demonstrations are confined to the simplest case of each subsidiary model (for instance, the univariate dichotomous logit model), but extensions are indicated, as well as new variants of the basic models that may be suggested by the MDI framework. We then give several concrete applications of these techniques to actuarial problems such as univariate and multivariate graduation incorporating constraints and construction of loss distributions using partial information.

2. INFORMATION THEORY

The concept of "information" has a long and prominent history in the development of statistics. Clearly the amount of "uncertainty" and the amount of "information" are inversely related: the information in an experiment is the amount of uncertainty that would be eliminated by performing the experiment. The common measure of "information" utilized by Fisher [31] (and taught in the SOA study material) takes the form of the inverse of the variance and was a first attempt to formalize the notion that data and models in statistics were essentially information transmissions. In the case of the normal distribution, Fisher's variance-covariance based measure turns out to be a reasonable measure of uncertainty, and classical statistics as used in actuarial science research defines (the Fisher) information in terms of such moments.

However, other potentially more appropriate measures of information are available, such as that defined by Shannon and Weaver [72] and developed further by Khinchine [48], [49] and by Kullback and Leibler [51]. As actuarial models become more sophisticated, the restrictions involved in normality assumptions become widely recognized, and the distributions involved in actuarial calculations depart substantially from normality (for example, mortality distributions closely resembling the Gompertz law, sample sizes involved in risk pooling decreasingly subject to central limit arguments due to smaller sample sizes, and so on), we see a move away from strictly normal distribution-based models and, consequently, a move away from variance-based measures of information. In fact, for statistical applications, Wiener [77] remarked quite early (1948) that the Shannon measure of information would eventually replace the Fisher measure of information.

We first present the statistical methodology. In information theoretic notation, the expected amount of information in the observation of a random variable X for distinguishing between two potential probability distributions $\mathbf{p}$ and $\mathbf{q}$ for X is denoted by $I(\mathbf{p}|\mathbf{q})$. Mathematically this expected information is quantified by the expected value of the log-odds ratio (which is a sufficient statistic for this discrimination, compare Cox and Hinkley [27, pp. 20–21]). In shorthand symbolic notation,

$$I(\mathbf{p}|\mathbf{q}) = \int_{-\infty}^{\infty} p(x) \ln \left[\frac{p(x)}{q(x)}\right] \lambda(dx) \tag{2.1}$$

where λ is some dominating measure for both $\mathbf{p}$ and $\mathbf{q}$. Usually $\lambda(dx)$ is Lebesgue measure in the (absolutely) continuous case or counting measure in the discrete case, so that a more familiar concrete representation is:

$$l(\mathbf{p}|\mathbf{q}) = \left\{ \begin{array}{l} \displaystyle\int_{-\infty}^{\infty} p(x) \ln \left[\frac{p(x)}{q(x)}\right] dx \text{ if the variable } X \text{ is continuous} \\ \displaystyle\sum_{i=0}^{\infty} p(x_i) \ln \left[\frac{p(x_i)}{q(x_i)}\right] \text{ if the variable } X \text{ is discrete} \end{array} \right\}.$$

An application of Jensen's inequality applied to the function $h(x)=x\ln x$ suffices to prove that $l(\mathbf{p}|\mathbf{q})\geq 0$ with $l(\mathbf{p}|\mathbf{q})=0$ if and only if $\mathbf{p}=\mathbf{q}$. As a consequence, $l(\mathbf{p}|\mathbf{q})$ can be thought of as the (pseudo-) distance or "closeness" between $\mathbf{p}$ and $\mathbf{q}$ within the space of all probability measures.

Kullback and Leibler [51] have shown that this measure $l(\mathbf{p}|\mathbf{q})$ (called the Kullback-Leibler number) can be used to develop a consistent statistical theory for measuring the expected amount of information given by a set of observations. In addition, the measure $l(\mathbf{p}|\mathbf{q})$ satisfies certain intuitively and theoretically desirable properties of an information measure (Kullback [50]; Haaland, Brockett and Levine [37]; Guiasu [36]; and Sakamoto, Ishiguro and Kitagawa [67]). It has also been used in actuarial science as a measure of equity. In this application it is a special case of the general functions developed by Promislow [66] for this purpose.

If certain characteristics of the distributions are presumed known, such as moments (say, net premiums, life expectancy or expected values of other functions of the variable), percentiles (say, expected mortality rates or survival probabilities), or other characteristics that can be expressed as expected values, then these quantities can also be incorporated as constraints into the analysis. The minimization of $l(\mathbf{p}|\mathbf{q})$ subject to these given constraints results in a convex programming problem (compare Hillier and Leiberman [41, Chapter 14]). The dual of this convex programming problem is actually unconstrained, so that the computation of the minimum $l(\mathbf{p}|\mathbf{q})$ is simply carried out by using elementary numerical techniques (compare Burden and Faires [20]). The constrained minimum $l^*(\mathbf{p}|\mathbf{q})$, called the Minimum Discrimination Information (MDI) statistic, gives the expected amount of information that an observation X yields in favor of the distribution $\mathbf{p}$ as opposed to the distribution $\mathbf{q}$. We elaborate this further as follows.

Minimizing the information $l(\mathbf{p}|\mathbf{q})$ for discrimination between the probability distributions $\mathbf{p}$ and $\mathbf{q}$, subject to any constraints that may apply to the

parameters of **p**, results in an estimate of **p**, say **p***, which is the distribution least distinguishable from **q**, but which satisfies the given constraints (which **q** itself may not do). In many important cases, these information theoretic estimates are also maximum likelihood estimates, and they are in general best asymptotically normal (compare Gokhale and Kullback [32] and Kullback [50]).

The asymptotic (as the sample size increases) distribution theory for $l^*(\mathbf{p}|\mathbf{q})$ (the minimum discrimination information (MDI) value) leads to a chi-squared test of the hypothesis that **p** and **q** are identical, that is, that the observed parameters are consistent with the estimated parameters (compare Golden, Brockett, and Zimmer [34]). From an actuarial perspective, one benefit of the above development is that estimation and hypothesis testing can be achieved simultaneously, consequently making for a more coherent approach to actuarial analysis under uncertainty. In addition, because $l(\mathbf{p}|\mathbf{q})$ is a general measure of the "distance" between **p** and **q**, all estimates and inferred relationships resulting from a constrained MDI problem are valid regardless of whether H_0 is accepted. In fact, if the hypothesis that p and q are identically distributed is not true, then the asymptotic limit of $l^*(\mathbf{p}|\mathbf{q})$ is the "distance" between the convex set of all distributions that *do* satisfy the given constraints and the given hypothesized distribution q. See Sakamoto, Ishiguro and Kitagawa [67] for details and a proof.

Information theory provides a convenient framework for most of the usual problems of statistical inference as taught in the Society of Actuaries syllabus. Akaike [2], for instance, has shown that the principle of maximum likelihood and the Fisher information approach are asymptotically equivalent to the information theoretic method, thus yielding on the one hand, a decision theoretic interpretation of maximum likelihood and, on the other hand, a single rational decision theoretic method of statistical estimation and hypothesis testing. As such, it becomes a unifying principle for the otherwise separate parts of statistics taught and applied in actuarial science.

3. MODEL SELECTION

Using the above developed "distance" interpretation of $l(\mathbf{p}|\mathbf{q})$, Akaike [2] showed how to use information theory to choose rationally among competing stochastic models to obtain a parsimonious parametric representation for a given stochastic phenomenon. Of course selecting an appropriate stochastic model is essential for actuarial analysis to provide useful input for decision-making. The information theoretic approach to model selection pioneered by Akaike proceeds as follows:

Using **q** as a generic potential postulated or stochastic model-based distribution and **p** as the "true" (but unknown) underlying distribution that is to be modeled, Akaike [2] proposes to choose the stochastic model **q** that is as "close" to the true stochastic model as possible, that is, which minimizes $l(\mathbf{p}|\mathbf{q})$. To this end we first observe that

$$l(\mathbf{p}|\mathbf{q}) = \int_{-\infty}^{\infty} p(x) \ln \left[\frac{p(x)}{q(x)}\right] \lambda(dx) \tag{3.1}$$

$$= \int_{-\infty}^{\infty} p(x) \ln [p(x)] \lambda(dx) - \int_{-\infty}^{\infty} p(x) \ln [q(x)] \lambda(dx).$$

The first term in (3.1) depends only upon the distribution of the true (but unknown) stochastic model **p** and hence is common to all the potential postulated densities **q**. To select the best or "closest to correct" postulated stochastic model from among a given class of postulated models, **q** must be chosen to minimize $l(\mathbf{p}|\mathbf{q})$ or, equivalently, to maximize

$$\int_{-\infty}^{\infty} p(x) \ln [q(x)] \lambda(dx),$$

the expected log-likelihood of **q**.

By the law of large numbers, based upon a sample $X_1, X_2, \ldots, X_n$ from the true probability distribution **p**, the expectation

$$\int_{-\infty}^{\infty} p(x) \ln [q(x)] \lambda(dx)$$

can be approximated by $1/n$ times

$$\sum_{i=0}^{n} \ln [q(x_i)] = \ln \left[\prod_{i=0}^{n} q(x_i)\right],$$

that is, by the sample average log-likelihood of the postulated model.

In addition, there are many important situations in the actuarial context in which the postulated stochastic model is a *parametric* model with parameters $\theta_1, \theta_2, \ldots, \theta_m$. For example, multivariate regression, polynomial

regression, and other linear models are often used in insurance ratemaking (compare Kalbfleisch [47], Sampson and Thomas [68]); ARIMA time series models (compare Miller and Wichern [61]) are used in models of interest rate structures (compare, Panjer and Bellhouse [64] and Bellhouse and Panjer [6]); Markov models are used in multivariate increment-decrement marriage or working life tables (compare Schoen and Land [71] and Hoem [42]); and factor analysis models are used in arbitrage price models for financial valuation of equity and bond prices (compare Martin, Cox and MacMinn [57]). In these parametric situations Akaike shows that $1/n$ times the *maximum* log-likelihood

$$\ln\left[\prod_{i=0}^{n} q(x_i:\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)\right]$$

(which is obtained by substituting the maximum likelihood parameter estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$ into the likelihood function) is, in fact, (asymptotically) upwardly biased as an estimator of the *expected* log-likelihood

$$\int_{-\infty}^{\infty} p(x) \ln [q(x)] \lambda(dx),$$

and that the size of this bias is precisely m/n, where m is the number of free parameters in this particular postulated stochastic model $\mathbf{q}$. Incorporating this bias correction into the maximum log-likelihood estimate gives the desired estimate of the expectation

$$\int_{-\infty}^{\infty} p(x) \ln[q(x)] \lambda(dx),$$

(which as we have noted, differs only by a constant common to all postulated models from the information distance $I(\mathbf{p}|\mathbf{q})$ between the true but unknown distribution $\mathbf{p}$ and postulated model $\mathbf{q}$).

Thus, using a sample $X_1, X_2, \ldots, X_n$, selecting the "closest" (minimum information distance) postulated model from among a given class of postulated models leads, upon multiplying the bias-corrected estimate through by the constant n, to the following criteria for stochastic model selection: Choose the stochastic model $\mathbf{q}$ that maximizes the quantity

maximum log-likelihood − number of free parameters.

For consistency with other statistical usage, this quantity is usually multiplied by -2 in order to arrive at the equivalent so-called Akaike Information Criteria (AIC):

minimize {-2(maximum log-likelihood $-$ number of free parameters)}.

Numerous examples that use the AIC criteria for model selection problems of interest to actuarial science researchers are worked out in Sakamoto, Ishiguro and Kitagawa [67]. These include such fundamental problems as determining the number of variables to include in a multiple regression, selecting the order of a polynominal regression, deciding upon the number of factors to include in a factor analysis, selecting the order of an ARIMA model, choosing the order of dependence in a Markov chain model, and other applications. Because most commercially available statistical computer programs output the log-likelihood value, the implementation of this stochastic model selection procedure is straightforward.

Although the above development was based upon the asymptotically unbiased estimation of the expected log-likelihood of the postulated model, a further "small sample" justification of the use of Kullback-Leibler information statistics for model selection is given by Larimore [53]. There it is shown that the considerations of likelihood and sufficiency lead naturally, even in the dependent variable setting, to the use of information theoretic measures as approximations to the actual predictive distributions obtained in repeated sampling, and leads to the same general information theoretic method of parsimonious stochastic model selection as above.

Note that for developing parsimonious parametric models, penalized log-likelihood methods proposed previously have been of the form:

Maximize (log-likelihood $-$ K $\times$ number of free parameters)

for some number K (for example, Schwarz [70] and Smith and Spiegelhalter [73]). However, the information theoretic method removes the ad-hoc nature of the selection of the value of the parameter K (it should be 1) and, moreover, develops this selection rule from a single unified philosophical information theoretic approach to the statistical analysis and consolidates it with other aspects of the statistical modeling and testing framework.

4. INFORMATION THEORY AND BAYES' THEOREM

Bayesian methods have a long history of use in actuarial science, possibly because these methods provide a well-advertised method for incorporating the prior experience and knowledge of the actuary into the model formulation

and estimation (compare Heilmann [40] for an example in a credibility theory context). Johns [44], however, points out in Kahn [46] that " ... insurance ratemaking provides one of the best examples where the pure Bayesian approach based on subjective prior probabilities is not appropriate." As we have seen, constrained information theoretic methods provide an alternative method of obtaining the same objective of incorporating the prior expertise of the actuary into the calculations. In the information theoretic formulation the prior information is incorporated through the use of the "standard" or "goal" distribution **q** and through the use of constraints that express properties that are known to hold. This section discusses further connections between the Bayesian and information theoretic approaches.

According to Kullback [50, p. 4], the first simple relationship between the information theoretic method and Bayes' Theorem (and hence a connection between information theoretic methods and the close cousin of Bayesian methods-decision theory) is obtained by writing

$$p(x) = Pr(x|H_1)$$

$$q(x) = Pr(x|H_2)$$

where $p(x)$ and $q(x)$ are the (either discrete or continuous) statistical distributions associated with two competing hypothesized distributions H_1:**p** and H_2:**q** for the random vector X. By using Bayes' Theorem, the likelihood ratio of the two hypothesized distributions **p** and **q** can be assessed in light of the sample data x and the given prior likelihoods $Pr(H_1)$ and $Pr(H_2)$ as follows:

$$\frac{Pr(H_1|x)}{Pr(H_2|x)} = \frac{Pr(x|H_1)Pr(H_1)}{Pr(x|H_2)Pr(H_2)} = \frac{p(x)Pr(H_1)}{q(x)Pr(H_2)},$$

or

$$\log \frac{p(x)}{q(x)} = \log \frac{Pr(H_1|x)}{Pr(H_2|x)} - \log \frac{Pr(H_1)}{Pr(H_2)}.$$

The expression on the left is the "log-odds" ratio in favor of the distribution **p** against the distribution **q** on the basis of the observation $X=x$ and is a sufficient statistic for this discrimination (compare Cox and Hinkley [27, pp. 20–21]). It is evidently the difference between the posterior and prior distribution log-odds ratios and hence can be interpreted as the information gained *in favor* of **p** by additional knowledge of the observation x.

The statistic

$$l(\mathbf{p}|\mathbf{q}) = \left\{ \begin{array}{l} \displaystyle\int_{-\infty}^{\infty} p(x) \ln \left[\frac{p(x)}{q(x)}\right] dx \text{ if the variable } X \text{ is continuous} \\ \displaystyle\sum_{i=0}^{\infty} p(x_i) \ln \left[\frac{p(x_i)}{q(x_i)}\right] \text{ if the variable } X \text{ is discrete.} \end{array} \right\}$$

thus represents the expected value (using **p**'s distribution) of this gain and hence can be interpreted as the average information gained in favor of **p** per observation of X, assuming that H_1 is the true state of nature.[1] See Golden, Brockett, and Zimmer [34] for more details.[2] Thus, Bayes' Theorem can be used to obtain an interpretation of the information theoretic functional as the expected information gain from repeated sampling. Even more fundamental connections exist between information theoretic methods and Bayesian analysis, however. For example, Zellner [78] has shown that if one *starts* with an information theoretic formulation and seeks an optimal information processing principle that transmits all the "information" in the prior distribution for updating the prior distribution in order to obtain a posterior distribution, then in fact *Bayes' Theorem is a consequence* of the information theoretic method. This is important in actuarial science because of the numerous applications that take Bayesian methods as a starting point for their analysis. Although Bayesian techniques have been criticized in science for their ostensible subjectivity, the "objective" and "frequentist" information theoretic method can be used to provide an alternative approach to virtually all nonsubjective applications of Bayesian techniques in actuarial science.

[1]If we are interested in distinguishing *between* hypotheses (as opposed to favoring one particular hypothesis), then we may introduce the symmetric statistic that Kullback [50] refers to as the "divergence measure" $J(p,q) = l(p|q) + l(q|p) =$

$$\int_{-\infty}^{\infty} p(x) \ln \left[\frac{p(x)}{q(x)}\right] \lambda(dx) + \int_{-\infty}^{\infty} q(x) \ln \left[\frac{q(x)}{p(x)}\right] \lambda(dx) = \int_{-\infty}^{\infty} [p(x) - q(x)] \ln \left[\frac{p(x)}{q(x)}\right] \lambda(dx).$$

This represents a generalization of the usual Mahalanobis distance statistic for normal distributions. The terms $l(p|q)$ and $l(q|p)$ represent what might be called "directed divergences," and $J(p,q)$ is a measure of the divergence between H_1 and H_2 on the basis of $X = x$. $J(p,q)$ has all the properties of a distance measure, except that it need not satisfy the triangle inequality. See Appendix A in Charnes and Cooper [22].

[2]Note that in this context $I[p|q]$ is a "frequentist" approach as opposed to a subjective Bayesian approach.

Other authors have also noted some strong similarities between Bayesian methods and the information theoretic approach. For example, Murray [62] showed that when developing a predictive distribution for a multivariate normal distribution with unknown mean and covariance matrix as, for example, in Bayesian graduation techniques (compare London [56]), the MDI estimate over the class of invariant distributions coincides with the Bayesian predictive distribution obtained by using an invariant prior distribution. Ng [63] generalized this approach and showed that a number of common distributions lead to the same predictive distributions from either a Bayesian or a repeated sampling-information theoretic method. Thus, information theoretic methods can unify and extend certain Bayesian as well as frequentist approaches used in actuarial science. Akaike [3], [4] provides further insight into the information theoretic connections with Bayesian procedures. The information theoretic method can also be used to enable the Bayesian to construct a unimodal prior distribution that expresses the assessed information elicited from the decision-maker. Prior distribution determination is a critically important problem that must be addressed in the implementation of any Bayesian technique. See Brockett, Charnes, Golden and Paick [16] for details.

5. THE CANONICAL MDI PROBLEM

In the previous sections we have considered information theoretic approaches involving analysis of the functional $I(\mathbf{p}|\mathbf{q})$ viewed as a pseudo-distance measure over various classes of probability distributions. This gave a unified approach to many problems of interest, such as model selection, and generalized other approaches, such as Bayesian methods. We now introduce the notion of *constrained* information theoretic analysis and show that this further unifies and generalizes many important methods in actuarial science.

We consider first the continuous situation of estimating the density $p(x)$, which is as close as possible to a given density $q(x)$ but which satisfies the additional expectational constraints $E[a_j(X)] = \theta_j$, $j = 1, 2, \ldots, m$; that is,

$$\text{Minimize } I(p|q) = \int_{-\infty}^{\infty} p(x) \ln \left[\frac{p(x)}{q(x)} \right] dx$$

subject to

$$1 = \theta_0 = \int_{-\infty}^{\infty} p(x)\,dx \tag{5.1}$$

$$\theta_j = \int_{-\infty}^{\infty} a_j(x)\,p(x)dx \quad j = 1, 2, \ldots, m.$$

By selecting $a_j(x) = x^j$, we may introduce expectational constraints implying given values of the moments, for example, means (net single premiums) or variances, of the modeled probability distribution. Similarly, by selecting $a_j(x)$ as the indicator function of an interval, we may introduce constraints involving probabilities (survival probabilities or stop loss values, for example).

For ease of presentation, we also specify the discrete analogue of (5.1). Accordingly, we consider a discrete probability distribution $\mathbf{q} = (q_1, q_2, \ldots, q_n)$. The discrete version of (5.1) is:

$$\text{Minimize } I(\mathbf{p}|\mathbf{q}) = \sum_{i=1}^{n} p_i \ln\left[\frac{p_i}{q_i}\right]$$

subject to

$$\sum_{i=1}^{n} p_i = 1$$

$$\sum_{i=1}^{n} a_{ij}\,p_i = \theta_j, j = 1, 2, \ldots, m,$$

$$p_i \geq 0, i = 1, 2, \ldots, n,$$

This can be written succinctly in matrix form as Min $I(\mathbf{p}|\mathbf{q})$ subject to

$$\begin{aligned} \mathbf{Ap} &= \theta \\ \mathbf{p} &\geq \mathbf{0}. \end{aligned} \tag{5.2}$$

The matrix $\mathbf{A}$ is $(m+1)\times n$ with first (zero-th) row set to be all ones, and the first constant $\theta_0 = 1$ appended to ensure the constraint

$$\sum_{i=1}^{n} p_i = 1,$$

that is, to ensure that the optimizing vector $\mathbf{p}^*$ forms a probability distribution. (If one is dealing with numerical values $\{p_i\}$ that do not necessarily sum to one, for example, when dealing with mortality rates as opposed to probabilities, then the zero-th row constraint can be deleted and the MDI problem is still well-defined.)

Problems of the form (5.1) or (5.2) are referred to as Minimum Discrimination Information (MDI) problems. In the above discrete MDI formulations, the probabilities q_i and the constraint values θ_j are constants. The q_i may be hypothesized values and the θ_j sample statistics or vice versa. They may be sample determined or exogenously determined. In any case, the null hypothesis to be examined is H_0:$\mathbf{p}=\mathbf{q}$; that is, the observed and expected probability distributions are not statistically distinguishable.

The information theoretic framework outlined previously allows for statistical testing of the hypothesized relation that the postulated and the observed distributions are statistically compatible (that is, a test of the above H_0). If we denote by $l^*(\mathbf{p}|\mathbf{q})=l(\mathbf{p}^*|\mathbf{q})$ the constrained minimum value of (5.2) and if we suppose that the observed distribution does, in fact, arise from a random sample of size M from the probability distribution $\mathbf{q}$, then since $\mathbf{p}^*\to\mathbf{q}$, the empirically calculated value of $l^*(\mathbf{p}|\mathbf{q})$ converges to zero at a rate of $1/M$). Under these conditions, $2Ml^*(\mathbf{p}|\mathbf{q})$ can be shown to be asymptotically distributed as a chi-square random variable, with degrees of freedom that depend upon n, the number of probabilities in the vector $\mathbf{q}$ and upon the number of linearly independent constraints (see Kullback [50], Gokhale and Kullback [33], and Phillips [65]). Examples of these statistical tests are given in Golden, Brockett, and Zimmer [34] and in Gokhale and Kullback [33], where exact formulas are given for the degrees of freedom involved in certain common tests (such as independence and conditional independence in contingency tables).

6. THE LOGLINEAR (MULTIPLICATIVE) MODEL IS A CONSEQUENCE OF INFORMATION THEORETIC MODELING

The loglinear model has application in several aspects of actuarial science research (for example, the automobile ratemaking models proposed by Chang and Fairley [21], Fairley, Thomberlin and Weisberg [29], Weisberg, Thomberlin, and Chatterjee [76], Coutts [26], or Harrington [38]). The exponentiated analogues—the multiplicative models—are also common in actuarial science. By introducing a Lagrange multiplier for each of the constraints in the MDI formulation, (5.1) or (5.2), it can be shown (compare Brockett, Charnes and Cooper [15] or Gokhale and Kullback [32]) that the optimum

probability distribution obtained by using an information theoretic approach is precisely of loglinear form.

To see that this is true in the continuous case (5.1), we introduce the Lagrange multiplier z_j for each constraint j and then minimize

$$\int_{-\infty}^{\infty} p(x) \ln\left[\frac{p(x)}{q(x)}\right] dx - z_0 \int_{-\infty}^{\infty} p(x)\, dx - z_1 \int_{-\infty}^{\infty} a_1(x)\, p(x)\, dx - \ldots - z_m \int_{-\infty}^{\infty} a_m(x)\, p(x)\, dx.$$

Letting $a_0(x) = 1$ and multiplying by -1 yields the equivalent maximization problem:

$$\text{maximize} \int_{-\infty}^{\infty} p(x) \ln\left[\frac{q(x)}{p(x)}\right] dx + \sum_{j=0}^{m} z_j \int_{-\infty}^{\infty} a_j(x)\, p(x)\, dx$$

$$= \int_{-\infty}^{\infty} p(x) \left\{ \ln\left[\frac{q(x)}{p(x)}\right] + \sum_{j=0}^{m} z_j\, a_j(x) \right\} dx$$

$$= \int_{-\infty}^{\infty} p(x) \left\{ \ln\left[\frac{q(x) \exp\left\{ \sum_{j=0}^{m} z_j\, a_j(x) \right\}}{p(x)} \right] \right\} dx$$

$$\leq \int_{-\infty}^{\infty} p(x) \left\{ \frac{q(x) \exp\left\{ \sum_{j=0}^{m} z_j\, a_j(x) \right\}}{p(x)} - 1 \right\} dx$$

where the inequality follows since $\ln Z \leq Z - 1$ with equality if and only if $Z = 1$. Thus the upper bound is actually obtained with

$$Z = \frac{q(x) \exp\left\{ \sum_{j=0}^{m} z_j\, a_j(x) \right\}}{p(x)} = 1,$$

that is, when

$$p(x) = q(x) \exp\left\{\sum_{j=0}^{m} z_j\, a_j(x)\right\}. \tag{6.1}$$

Thus (6.1) is the desired maximum and is as close as possible (in an information theoretic distance sense) to the distribution $\mathbf{q}$ subject to the constraints. Note that

$$\ln\left[\frac{p(x)}{q(x)}\right] = \sum_{j=0}^{m} \{z_j\, a_j(x)\}, \tag{6.2}$$

so the resulting representation for $p(x)$ is in loglinear form.

For the discrete situation modeled by (5.2), the analogous loglinear representation obtains by virtually the same mathematics. In this case the optimal probability values are of the form

$$p_i = q_i \exp\{{}_i\mathbf{A}\mathbf{z}\} \tag{6.3}$$

where ${}_i\mathbf{A}$ denotes the i-th column of the matrix $\mathbf{A}$ and $\mathbf{z} = \{z_j\}$ is the collection of Lagrange multipliers. Again, we note that this is the loglinear model as described in, for example, Bishop, Fienberg and Holland [8]. Thus loglinear modeling is a *particular case* of the information theoretic approach, but one in which the parameterization is determined from a fundamental approach. Note also that the coefficients of the various parameters in the loglinear representation of the probabilities can be immediately read from the columns of the matrix $\mathbf{A}$ in the information theoretic formulation.

If we write the loglinear model in the form

$$\ln p_i - \ln q_i = \ln\left[\frac{p_i}{q_i}\right] = \sum_{j=1}^{m} z_j\, a_{ij},$$

where the z_j are the loglinear parameters, we observe that the z_j are actually Lagrange multipliers determined so that the resulting probabilities p_i satisfy the given moment constraints. Consequently, the p_i in this loglinear model automatically sum to one, because this is a condition of the MDI problem. Gokhale and Kullback [32] stress that this loglinear model of the p_i is a *consequence* of the MDI formulation and is not derived from seemingly arbitrary assumptions of convenience. Moreover, the MDI approach shows that the term $\ln q_j$ in the log-odds equation is not merely a constant of fit. Its meaning as a prior probability is implied by the derivation of $l(\mathbf{p}|\mathbf{q})$ via Bayes' Therorem in the earlier section.

If we now apply an exponential transformation to both sides of the log-linear equation, we arrive naturally at the multiplicative model that is also used in actuarial science (compare Almer [5], Jung [45], Ajne [1], and Sant [69]).

$$p_i = q_i \exp\left\{\sum_{j=1}^{m} z_i\, a_{ij}\right\} = q_i \prod_{j=1}^{m} e^{z_j\, a_{ij}} = q_i \prod_{j=1}^{m} \{e^{a_{ij}}\}^{z_j}$$

$$= q_i \prod_{j=1}^{m} \{y_{ij}\}^{z_j}$$

where $y_{ij} = e^{a_{ij}}$.

The information theoretic formulation also allows for simplified computation of the loglinear or multiplicative model parameters. This is due to the following duality result from convex programming. Charnes and Cooper [23] and Brockett, Charnes and Cooper [15] prove that the following two problems form a duality pair:

Primal Problem	Dual Problem
$\text{Max } \nu(\mathbf{p}) \equiv -\sum_i p_i \ln (p_i/eq_i)$	$\text{Min } \xi(\beta) \equiv -\sum_i q_i\, e^{i\mathcal{A}\beta - 1} - \theta^T\beta$
Subject to	
$\mathbf{Ap} = \theta$	β unconstrained
$\mathbf{p} \geq \mathbf{0}$	

Here e is the base of the natural logarithm and enters the formulation because of the duality proof. Whenever the constraint set implies that the two distributions have the same total mass (as is the situation when dealing with probability distributions for example), then the number e can be eliminated from the formulation provided one is only concerned with the primal problem and not the duality relationship.

These duality results imply in particular that, at their optimum values, $\xi(\beta^*) = \nu(\mathbf{p}^*)$, and

$$p_i^* = q_i\, e^{i\mathcal{A}\beta - 1}.$$

(See Brockett, Charnes and Cooper [15] for a complete statement.) Two important things to note are: (1) the condition

$$p_i^* = q_i\, e^{i\mathcal{A}\beta - 1}$$

is a reparameterization of the multiplicative (loglinear) model of the estimates p_i^*, and (2) the loglinear parameters $\mathbf{z}$ in the multiplicative model (which are the Lagrange multipliers in the primal formulation of the corresponding MDI problem) can be easily calculated numerically from the optimal solution to the *unconstrained* dual convex programming problem, which involves only linear and exponential terms. Equating the results obtained using Equation (6.3) with the above duality results implies the first parameter representing the constant normalizing constraint is $z_0 = \beta_0 - 1$ in the dual, while the remaining loglinear parameters (or Lagrange multipliers) are $z_i = \beta_i$ in terms of the corresponding dual parameters. Thus, information theoretic formulation at once simply renders the interpretation and computation of the loglinear parameters. Later sections of this paper link the loglinear model to logit models and cite further actuarial science applications. The computational simplifications available due to the duality formulation are also passed along to the computation of the logit models and eliminate the use of ad-hoc or approximate methods.

7. MDI AND THE ENTROPY MODEL

Maximum entropy models have become well-known in physical science research and more recently in actuarial science research (see Martin-Lof [58], [59] for additional references). The discrete entropy model involves maximizing the "entropy" function of a distribution $\mathbf{p}$:

$$\text{Maximize } H(\mathbf{p}) = -\sum_{i=1}^{n} p_i \ln [p_i]$$

subject to linear constraints. It is easily seen that if we let $q_i = 1/n$ for $i = 1, \ldots, n$, then $H(\mathbf{p})$ finds its extremum at the same point as does $l(\mathbf{p}|\mathbf{q})$ because

$$\begin{aligned} l(\mathbf{p}|\mathbf{q}) &= \sum_{i=1}^{n} p_i \ln (p_i n) \\ &= \sum p_i \ln p_i + \ln n \\ &= -H(\mathbf{p}) + (\text{constant}). \end{aligned}$$

$H(\mathbf{p})$ and $l(\mathbf{p}|\mathbf{q})$ are therefore measures of the deviation of $\mathbf{p}$ from a discrete uniform distribution over n points. (The same relationship exists between continuous entropy and the information theoretic distance to the continuous uniform function $g(x) = 1$.) In this regard the entropy function is clearly a special case of the discrimination information statistic. Moreover, the MDI

formulation is more general and offers greater flexibility, because with the MDI formulation the null-hypothesis function **q** can represent any probability function (not just a uniform distribution). In addition, the MDI formulation has a complete and rigorous theoretical foundation in statistics, so that we need not be troubled by nonrigorous analogies from thermodynamics—as is so often the case with "entropic" models (see Phillips [65] and Haynes, Phillips and Mohrfeld [39]).

The *constrained* maximum entropy density estimation may be construed as a new useful extension of Laplace's famous "principle of insufficient reason," which postulates a uniform distribution when no knowledge is available. Here, when only information of the form (5.1) is available, we select the distribution that is as close to uniform as possible subject only to these given constraints.[3]

This formulation gives for the first time a unified characterization in terms of Laplace's information principle of the constant force and uniform distribution of deaths assumptions so commonly used in mortality table analysis. We can for the first time delineate the precise type of information that is being assumed when using these models.

If all that is known about the distribution of deaths within an age interval $[a,b]$ is the probability of death, then there is only a single piece of information that, in an MDI framework, translates into the single constraint

$$\int_{-\infty}^{\infty} a_0(x)\, p(x)\, dx = \theta_0 \tag{7.1}$$

where

$$a_0(x) = \left\{ \begin{array}{l} 0 \text{ if } x < a \\ 1 \text{ if } a \leq x < b \\ 0 \text{ if } x \geq b \end{array} \right\},$$

and θ_0 is the given probability of death. In this case the MDI density in (6.1) has a piecewise constant form and corresponds to the uniform distribution of deaths assumption. The interpretation immediately follows that the

[3]Of course, densities other than the uniform may be more appropriate for reference or comparison in certain situations (such as using a "standard" mortality table or a Gompertz table as argued for by Tenenbein and Vanderhoof [75]. This would lead to viewing the general MDI problem (5.1) as yet a further generalization of Laplace's famous "principle of insufficient reason."

uniform distribution of deaths assumption provides the "least informative (maximum entropy) distribution" possible with these given probabilities.

In the case in which further information *is* available about the mortality during the interval $[a,b]$, the constant force assumption can be derived. Suppose that the mean life length during the interval is also known. This imposes the additional constraint of the form

$$\int_{-\infty}^{\infty} a_1(x)\, p(x)\, dx = \theta_1 \tag{7.2}$$

where

$$a_1(x) = \begin{Bmatrix} 0 \text{ if } x < a \\ x \text{ if } a \leq x < b \\ 0 \text{ if } x \geq b \end{Bmatrix},$$

with θ_1 being the given average age of death during the interval. Now the MDI density in (6.1) has a form that is piecewise equal to $p(x) = \exp[z_0 + z_1 x]$. The constants z_0 and z_1 are selected to ensure that the two constraints (7.1) and (7.2) hold. This piecewise exponential density $\exp[z_0 + z_1 x]$ is precisely the constant force model with exponential survivorship[4] within each interval, and thus can be interpreted as the least informative mortality distribution possible subject only to the knowledge of individual interval death probabilities and average death ages. Any other mortality assumption must involve additional information, which, if we truly had, could also be incorporated into the constraint set (5.1).

8. MDI AND THE LOGIT MODEL

The Logit model is a special case of the loglinear model (see Green, Carmone, and Wachspress [35]) and hence can also be derived as a particular case of the information theoretic technique. In the logistic model the log-odds ratio in favor of occurrence of a particular event, E, is related to several independent variables $x_1, x_2, x_3, \ldots$ via the relationship

$$\ln\left[\frac{P(E)}{1 - P(E)}\right] = \sum_i \beta_i x_i,$$

where the β_i's are parameters to be fit to the observed data.

[4]Note that z_1 is negative and z_0 is merely a normalizing constant that gives the requisite probability of survivorship for the interval.

There have been numerous uses of logit models in the biostatistical and actuarial literature (compare Steenackers and Goovaerts [74] and Elandt-Johnson and Johnson [28]). We present the following univariate logit example to illustrate that the constrained information theoretic generalization of the entropic analysis presented here also contains the logit model for contingency table analysis. Gokhale and Kullback [32, p. 273] provide an MDI derivation of a multivariate logit model.

For our illustration we utilize an example from Berkson [7] and Gokhale and Kullback [32] in which the following four samples are analyzed under different values of a single covariate x:

Sample Number	Value of x	Sample Size	Number of Successes
1	0	10	1
2	1	10	6
3	2	10	3
4	3	10	8

These are transformed into a contingency table format:

Value of x	i	Success ($j=1$)	Failure ($j=2$)	Total
0	1	1	9	10
1	2	6	4	10
2	3	3	7	10
3	4	8	2	10
Total		18	22	40

We solve the maximum entropy problem below,

$$\text{Max } H(\mathbf{p}) = -\sum_{i=1}^{4}\sum_{j=1}^{2} p_{ij} \ln [p_{ij}]$$

subject to

$$\sum_{j=1}^{2} p_{ij} = \frac{10}{40}, \; i = 1, 2, 3, 4$$

$$\sum_{i=1}^{4} p_{i1} = \frac{18}{40}$$

$$\sum_{i=1}^{4} x_i p_{i1} = \frac{36}{40}$$

$$p_{ij} \geq 0 \qquad i = 1, 2, 3, 4; j = 1, 2.$$

If we translate the series of summations into a single matrix equation, it is of the form

$$\begin{vmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ x_1 & 0 & x_2 & 0 & x_3 & 0 & x_4 & 0 \end{vmatrix} \begin{vmatrix} p_{11} \\ p_{12} \\ p_{21} \\ p_{22} \\ p_{31} \\ p_{32} \\ p_{41} \\ p_{42} \end{vmatrix} = \begin{vmatrix} 1 \\ \frac{10}{40} \\ \frac{10}{40} \\ \frac{10}{40} \\ \frac{10}{40} \\ \frac{18}{40} \\ \frac{36}{40} \end{vmatrix}$$

The problem is thus recognized as being of the general MDI form

$$\text{Max } H(\mathbf{p})$$

subject to

$$\mathbf{AP} = \theta$$
$$\mathbf{p} \geq \mathbf{0}.$$

From the loglinear representation of the optimal solution in terms of the columns of the matrix **A** and the Lagrange multipliers, $z_0, z_1, \ldots, z_6$ for each of the constraints, we find:

$$\ln [p_{11}] = x_1 z_6 + z_5 + z_1 + z_0$$

and

$$\ln [1 - p_{11}] = \ln [p_{12}] = z_1 + z_0,$$

so that

$$\ln\left[\frac{p_{11}^*}{1-p_{11}^*}\right] = x_1 z_6 + z_5.$$

A similar calculation of the loglinear representation for p_{i1} and $1-p_{i1}=p_{i2}$ in the cases $i=2, 3, 4$ yields the same coefficients as above, so that in general

$$\ln\left[\frac{p_{i1}}{1-p_{i1}}\right] = x_i z_6 + z_5.$$

This is precisely the logistic model

$$\ln\left[\frac{P(E)}{1-P(E)}\right] = \sum_i \beta_i x_i$$

with the Lagrange multipliers serving the role of the logistic parameters. The desired probability $P(E|x)$ is given by the logistic cumulant function

$$P(E|x) = \frac{1}{1+e^{-(xz_6+z_5)}},$$

and thus logistic modeling can also be developed from the MDI perspective discussed here. The logistic parameters may be easily calculated from the dual convex programming formulation delineated in a previous section. This dual convex programming problem is unconstrained, making it easily amenable to the numerical analysis solution methods (for example, Newton-Raphson, successive bisection, and so on) as taught in SOA courses.

We now consider the MDI extension of the above max-entropy-to-logit sequence. If above we replace Max $H(p_{ij})$ by Min $I(p_{ij}|\pi_{ij})$, the resulting log-odds are

$$\ln\frac{p_{ij}}{1-p_{ij}} = \ln\frac{\pi_{ij}}{1-\pi_{ij}} + xz_7 + z_6.$$

Evidently if the prior log-odds in $\ln \pi_{ij}/1-\pi_{ij}$ is a linear function of x, then $\ln p_{ij}/1-p_{ij}$ will also be a linear function of x. Otherwise, the resulting representation will constitute a nonlinear generalization of the logit model. This more comprehensive framework for the logit model may resolve some of the problems that arise in its application. In particular, this extends and

unifies many of the automobile ratemaking models that involve multiplicative relatives. The extension to include other covariates and prior distributions is also apparent from this new information theoretic formulation.

9. DETERMINING A CLIENT'S LOSS DISTRIBUTION

In this section we explicitly show how to use the previously discussed information theoretic methods for obtaining a loss distribution that is as close as possible to some selected standard or reference distribution and that reflects the individual characteristics of a client's history (which may be inconsistent with or not reflected by the standard distribution). A standard or reference distribution for losses for a particular insurance line may, for example, be adopted by a small insurance company using Insurance Service Office data bases and might not be routinely or immediately applicable to its particular situation without some adjustment to reflect the known characteristics of the clients of the particular small insurance company.

The problem is how to compare and adjust the standard or reference distribution without relying upon ad-hoc ratio or graphical or ratio methods, which are commonly used, compare London [56], but which possess no theoretical or statistical basis or justification for their usage. This application presents the actuary with a new method for adjusting and comparing distributions that has a firm statistical foundation (the information theoretic technology previously discussed) and that is capable of even further statistical analysis and extension. This is not possible with the ad-hoc ratio and graphical methods now taught to actuaries (for example, London [56]). As an illustration, we consider the problem of obtaining a duration table necessary for pricing a client's disability insurance; this example was first considered by Brockett [14].

For simplicity we assume that there is a single furnished constraint: the client's expected duration is $\mu = 21$ days, rather than the standard table mean duration of 31.35 days. The problem is to determine a duration table for the client that is as indistinguishable as possible from the standard table (obtained from internal company data or industry-wide data), but that is subject to the constraint $\mu = 21$ days.

Mathematically, we let q_i denote the probability of a disability lasting a duration of i days according to the standard or reference table, and p_i denote the unknown corresponding probability of a duration of $x_i = i$ days to be developed for the client. We find the values of $\{p_i\}$ by solving

$$\min l(\mathbf{p}|\mathbf{q}) = \sum_{i=1}^{\omega} p_i \ln \left[\frac{p_i}{q_i}\right]$$

subject to

$$1 = \sum_{i=1}^{\omega} p_i$$

$$21 = \sum_{i=1}^{\omega} x_i p_i = \sum_{i=1}^{\omega} i\, p_i.$$

According to Equation (6.3), the solution to this problem is of the form

$$p_i = q_i \exp[z_0 + z_1 x_i] = q_i \exp[z_0 + z_1 i] = \{q_i \exp[z_0]\}\{\exp[z_1]\}^i.$$

The numerical determination of the parameters z_0 and z_1 is easily obtained by using the *unconstrained* dual convex programming problem outlined in Section 6, that is,

$$\text{Min } \xi(\beta_0,\beta_1) \equiv -\sum_i q_i\, e^{A\beta-1} - \theta'\beta$$

where $\beta=(\beta_0,\beta_1)=(z_0+1, z_1)$, $\theta=(1,21)'$, and the matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \begin{vmatrix} 1 & 1 & 1 & \dots & 1 & 1 & 1 \\ 1 & 2 & 3 & \dots & \omega-2 & \omega-1 & \omega \end{vmatrix},$$

that is,

$$\text{Min } \xi(\beta_0,\beta_1) \equiv -\exp[\beta_0 - 1]\sum q_i\{\exp[\beta_1]\}^i - \beta_0 - 21\beta_1.$$

Because the above problem is unconstrained, any of a number of simple numerical algorithms can be used to solve for z_0 and z_1, including taking the derivative and setting it equal to zero, using Newton-Raphson methods, successive bisection, and so on.

In our case with $\mu=21$ days, we obtain $\beta_0=z_0+1=1.387888$ and $\beta_1=z_1=0.0150898$, so that the adjusted duration table probabilities satisfy

$$p_i = q_i \exp[\beta_0 - 1 + i\beta_1] = q_i \exp[z_0 + iz_1]$$
$$= q_i\,(1.473864876)(0.9850235)^i.$$

The resulting client duration table is given in Table 1.

TABLE 1*

Length Y	Standard Table $P[Y=y]$	Adjusted $P[Y=y]$ for $\mu=21$
1	0.03500	0.05081
2	0.03474	0.04968
3	0.03349	0.04717
4	0.03318	0.04604
5	0.03195	0.04367
6	0.03160	0.04254
7	0.03040	0.04031
8	0.03002	0.03921
9	0.02885	0.03712
10	0.02701	0.03423
11	0.02530	0.03159
12	0.02370	0.02915
13	0.02222	0.02692
14	0.02083	0.02485
15	0.01953	0.02295
16	0.01831	0.02120
17	0.01772	0.02021
18	0.01662	0.01867
19	0.01611	0.01783
20	0.01510	0.01646
21	0.01465	0.01573
22	0.01374	0.01453
23	0.01334	0.01390
24	0.01295	0.01329
25	0.01214	0.01227
26	0.01180	0.01175
27	0.01106	0.01085
28	0.01076	0.01039
31	0.06361	0.05873
38	0.04832	0.04014
45	0.03753	0.02805
52	0.02980	0.02004
59	0.02399	0.01452
66	0.01939	0.01056
73	0.01586	0.00777
80	0.01300	0.00573
87	0.01077	0.00427
91	0.12561	0.04690

*Client's derived distribution for the length of claim under group weekly Disability Income Insurance. Column 1 gives length of claim; column 2 represents the standard probabilities taken from Bowers et al. [12, Table 13.2]; and column 3 shows the result of adjusting the standard to obtain mean duration $\mu=21$ days.

10. ADJUSTING MORTALITY TABLES

In many situations an actuary is asked to adjust a standard mortality table to obtain a table appropriate for a particular individual or group of individuals. As an example, we consider an actuary asked to value the lost earnings of an individual in a wrongful death suit. The actuary must select a life table to use; however, there may be additional information such as a physician's estimate of life expectancy together with an estimate of the confidence given in this life expectancy estimate. The actuary must construct a mortality table that has $e_x = \mu$, where x is the age of the decedent and μ is the expectation of life estimated by the physician. If the standard table satisfies this condition, then there is no problem. However, because this is usually not the case, we suppose that for the standard table, $e_x \neq \mu$. Mathematically the problem of adjusting the standard table to reflect the known information is very similar to the problem of adjusting loss distributions discussed in the previous section.

We can use the information theoretic method previously demonstrated to obtain an adjusted table that is as indistinguishable as possible from the standard table and that satisfies the known constraint $e_x = \mu$. This same process is described in more detail in Brockett and Cox [17].

We provide two numerical examples that illustrate different levels of knowledge about the potential survivalship of the person in question. Table 2 presents the results. First, we assume that a physician who examined the individual in question is prepared to testify that the life lost was a male, aged 50, having a curtate future life expectancy of $\mu = 9$ years.

In this example we solve the mathematical programming problem

$$\text{Min } I(\mathbf{p}|\mathbf{q}) = \sum_{i=0}^{\omega} p_i \ln \left[\frac{p_i}{q_i}\right]$$

subject to

$$1 = \sum_{i=0}^{\omega} p_i$$

$$9 = \sum_{i=0}^{\omega} x_i p_i = \sum_{i=0}^{\omega} i\, p_i,$$

where q_i is the probability of death i years from now (during age $50+i$), and p_i is the desired probability of death in year i, which has been adjusted to reflect the known information. The standard table to be used for illustrative purposes in this example is the PBGC Table V.

TABLE 2

Age	Standard Probability q_i from PBGC	Probability p_i Adjusted for $\mu = 9$	Probability p_i Adjusted for $\mu = 9$ and Even Odds of Dying between 55th and 64th Birthdays
50	0.054711	0.086346	0.0734290
51	0.052871	0.080354	0.0683716
52	0.051146	0.074855	0.0637287
53	0.048879	0.068890	0.0586828
54	0.047452	0.064404	0.0548919
55	0.045562	0.059550	0.0715825
56	0.043648	0.054937	0.0660745
57	0.041633	0.050461	0.0607257
58	0.039687	0.046322	0.0557761
59	0.037742	0.042422	0.0511081
60	0.035806	0.038756	0.0467182
61	0.033890	0.035324	0.0426057
62	0.031694	0.031813	0.0383918
63	0.030149	0.029142	0.0351884
64	0.028303	0.026345	0.0318291
65	0.026290	0.023566	0.0202098
66	0.024334	0.021005	0.0180240
67	0.022640	0.018819	0.0161577
68	0.019377	0.015511	0.0133246
69	0.019377	0.014937	0.0128386
70	0.018002	0.013363	0.0114926
71	0.016731	0.011960	0.0102916
72	0.015552	0.010706	0.0092175
73	0.014360	0.009519	0.0082006
74	0.013207	0.008431	0.0072671
75	0.012155	0.007472	0.0064443
76	0.011236	0.006652	0.0057399
77	0.010307	0.005876	0.0050733
78	0.009380	0.005150	0.0044486
79	0.008894	0.004702	0.0040643

Solving the dual problem (or simply using Newton-Raphson or successive bisection techniques) yields the two parameters necessary for adjustment.[5] From (6.3) we have

$$p_i = q_i \exp(\beta_0 - 1) \exp(i\beta_1) = q_i(1.57823)(0.9629889)^i$$

The standard table values and the values adjusted to reflect the knowledge of life expectancy are listed in columns 2 and 3 of Table 2.

[5]This problem was actually solved quite easily by using a spreadsheet program (Excel). Using the equivalent formulation $p_i = aq_ib^i$ from (6.1), the constant term "a" is found from the first constraint to be $a = 1/\Sigma_i q_k b^k$, so that the second constraint translates into one equation in one unknown b: that is, $\mu = \Sigma_i\, i[q_ib^i/(\Sigma_k q_k b^k)]$, or equivalently $\Sigma_i iq_ib^i - \mu\Sigma_i\, q_ib^i = 0$. Successive bisection on a spreadsheet easily determines the parameter b, and hence a is also determined. Again by using the spreadsheet, the adjusted probability distribution $p_i = aq_ib^i$ is calculated.

TABLE 2—*Continued*

Age	Standard Probability q_i from PBGC	Probability p_i Adjusted for $\mu = 9$	Probability p_i Adjusted for $\mu = 9$ and Even Odds of Dying between 55th and 64th Birthdays
80	0.008616	0.004382	0.0037936
81	0.008274	0.004056	0.0035102
82	0.007865	0.003713	0.0032150
83	0.007393	0.003361	0.0029118
84	0.006860	0.003003	0.0026034
85	0.006286	0.002650	0.0022985
86	0.005673	0.002303	0.0019987
87	0.005032	0.001967	0.0017082
88	0.004382	0.001650	0.0014333
89	0.003728	0.001352	0.0011749
90	0.003098	0.001082	0.0009408
91	0.002512	0.000845	0.0007350
92	0.001979	0.000641	0.0005579
93	0.001510	0.000471	0.0004102
94	0.001108	0.000333	0.0002900
95	0.000782	0.000226	0.0001972
96	0.000526	0.000146	0.0001278
97	0.000336	0.000090	0.0000787
98	0.000202	0.000052	0.0000456
99	0.000113	0.000028	0.0000246
100	0.000058	0.000014	0.0000122
101	0.000027	0.000063	0.0000055
102	0.000139	0.000031	0.0000271
103	0.000004	$8.75E-07$	0.0000008
104	$1.23E-06$	$2.53E-07$	0.0000002

Assume now that the physician, when pressed under examination, also gives a confidence measure for the curtate life expectancy measure. Accordingly, he states that in addition to a curtate life expectancy of 9 years, he would have given even odds that the person in question would have died somewhere between his 55th and 65th birthdays. Taking this set of information into account in the adjustment process yields the second example. The physician's last revelation of information implies the constraint

$$0.5 = \sum_{i=5}^{14} p_i = \sum_{i=0}^{\omega} a(i)\, p_i$$

where

$$a(x) = \begin{Bmatrix} 0 \text{ if } x < 5 \\ 1 \text{ if } 5 \leq x \leq 14 \\ 0 \text{ if } x > 14 \end{Bmatrix},$$

so that there are now three parameters needed to achieve the necessary adjustment of the mortality table. From (6.3) the adjusted mortality table is of the form

$$p_i = q_i \exp(\beta_0 - 1) \exp(i\beta_1) \exp[\beta_2 a(i)]$$

$$= \begin{Bmatrix} q_i\,(1.342125)(0.96353)^i & \text{if } i < 5 \\ q_i\,(1.891812)(0.96353)^i & \text{if } 5 \le i \le 14 \\ q_i\,(1.342125)(0.96353)^i & \text{if } i > 14 \end{Bmatrix}.$$

The parameters can be easily determined by using the unconstrained dual mathematical programming problem.[6] The mortality table adjusted to reflect the physician's expert testimony is now reproduced in the final column of Table 2. The resulting table exactly satisfies the constraints that there be a mean of 9 years and that there be equal odds of the person in question dying between ages 55 and 65.

Some situations might require an adjusted life table or loss distribution in which the constraint set is not of the linear equality form given in (5.1) and (5.2). For example, uncertainty about the constraints may take the form of linear *inequality* constraints. The duality theory and corresponding computational advantages of the dual program in the linear inequality constrained case are exposed in Charnes et al. [25]. The computation of the optimum adjusted table or distribution is still readily carried out by using existing nonlinear programming codes (for example, the generalized reduced gradient method described in Lasdon et al. [52], and Liebman et al. [54], which is also available for the personal computer). The example in Section 11 includes linear equality, inequality, and nonlinear constraints and is still readily computed.

11. INFORMATION THEORETIC GRADUATION

Because there is no known universally applicable "law" of mortality, observable data must be used to bring out an underlying pattern in the data sufficient for probabilistically predicting future outcomes. Graduation is the process used by actuaries to develop a unified series of observations from the observed data.

In the case of mortality tables, the observed data often are of the form of the number of deaths at specific age intervals among a group of individuals

[6]In our calculation a spreadsheet program was used to determine the unknown parameter values, as described in footnote 5. The calculation was again simply performed by using successive bisection.

recorded over time. If graphed, the set of data would have a jagged pattern; these irregularities are due to the limited amount of data and the statistical sampling variations. Via the process of graduation, these irregularities are smoothed into a curve that "fits" the data while reflecting a desire for "smoothed" transitions from age to age, and any other characteristics that are assumed known about the mortality rate sequencing.

The most common methods of graduation are graphic, interpolation, adjusted average, difference equations, and graduation by mathematical formula (compare London [55]). This section introduces a new method (also discussed in Brockett and Zhang [18] and Zhang and Brockett [79]) based upon an information theoretic approach, which allows the inclusion of constraints such as isotonicity, convexity, or any other of a variety of desired attributes of the graduated series. This technique is easier to interpret and computationally simpler and faster than the Bayesian isotonic graduation method presented by Broffitt [19]. In addition, it can handle convexity constraints, intervals of isotonicity and can even be extended to multidimensional graduation (for example, select and ultimate mortality tables) quite simply.

Our technique is quite general and applicable to loss distributions, to term structures of interest rates, and to many other relevant issues for the actuary. For concreteness, however, we phrase our discussion in the sequel in terms of mortality tables (death rates). Thus we can also illustrate a situation in which the information theoretic technology is applied to a sequence of numbers (the mortality rates) that are not constrained to sum to one, as was the case with the previous examples.

In graduating mortality data, the entries are often the mortality rates rather than some other quantity (such as the probabilities discussed in previous sections), because it is usually the rates about which we have some prior information (based upon biological or other considerations). The mortality rate is defined to be the number of deaths divided by some measure of the number of lives exposed to death during the year. The most simple such mortality rate (called the crude mortality rate by biostatisticians and actuaries) is just the observed number of deaths during the year (age) x divided by the number of people alive at the beginning of the year (age interval) x. We denote this crude rate for age x by u_x. Thus, the graduation process can succinctly be phrased as follows: Given the observed series $\{u_x\}$, construct a "smooth" series $\{\delta_x\}$, which "closely approximates" the observed series $\{u_x\}$ and which satisfies certain other (biological or actuarial) constraints known to exist.

The goals of "fit" to the observed series, "smoothness" of the graduated series, and the desire to make the graduated series reflect the known prior information (for example, mortality rates increasing with age) can lead to conflicting goals. Clearly we could choose a very smooth series that does not adequately represent the data, or a series that perfectly represents the data but that is as irregular as the data themselves, or that does not exhibit the prior information known to hold. The problem is to choose a parsimonious tradeoff between these conflicting goals. We achieve this end through the use of information theoretic techniques.

In addition to the obvious nonnegativity constraint for the mortality rates, $\delta_x \geq 0$, we also have prior knowledge that the true underlying pattern of mortality rates is (a) smooth, (b) increasing with age (that is, $\Delta\delta_x \geq 0$ where $\Delta\delta_x = \delta_{x+1} - \delta_x$ is the usual forward difference operator), and (c) more steeply increasing at the higher ends of the age range (that is, $\Delta^2\delta_x \geq 0$). We would like the graduated table to reflect this prior information and also (as is standard in actuarial work) to satisfy the additional constraints that (d) the graduated number of deaths equals the observed number of deaths and (e) the total of the graduated ages at death equals the observed total ages at death. Constraints (d) and (e) together imply that the average age at death is required to be the same for the graduated and empirically derived tables.

A measure of smoothness often used by actuaries (compare London [55]) is the size of the sum of squares of the third differences of the derived series, namely, $\Sigma(\Delta^3\delta_x)^2$. The smaller this sum of squares, the smoother the graduation is judged to be. The measure of "fit to observed data" that we use is the informational distance $I(\delta|\mathbf{u}) = \Sigma\delta_i \ln[\delta_i/u_i]$ between the graduated series $\{\delta_i\}$ and observed crude mortality rate series $\{u_i\}$. The fact that $I(\delta|\mathbf{u})$ is still a measure of fit even in the nonprobability situation holds because the mortality rates are non-negative and because of the assumed constraints. For concreteness we reproduce the example given in Brockett and Zhang [18] involving the graduation of data from Miller [60] as considered in London [55]; Table 3 gives the raw data for graduation.

The graduation problem outlined above can be formulated as a constrained information theoretic problem:

$$\min_{\delta} I(\delta|\mathbf{u}) = \sum \delta_i \ln\left[\frac{\delta_i}{u_i}\right]$$

subject to

(a') Smoothness: $(\mathbf{A}\delta)'(\mathbf{A}\delta) = \delta'\mathbf{A}'\mathbf{A}\delta \leq M$

TABLE 3

Age	Exposed to Risk	Actual Deaths	Ungraduated Mortality Rate
70....	135	6	0.044
71....	143	12	0.084
72....	140	10	0.071
73....	144	11	0.076
74....	149	6	0.040
75....	154	16	0.104
76....	150	24	0.160
77....	139	8	0.058
78....	145	16	0.110
79....	140	13	0.093
80....	137	19	0.139
81....	136	21	0.154
82....	126	23	0.183
83....	126	26	0.206
84....	109	26	0.239
	2,073	237	

where

$$\mathbf{A} = \begin{vmatrix} -1 & 3 & -3 & 1 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & -1 & 3 & -3 & 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & -1 & 3 & -3 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & & & & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 3 & -3 & 1 \end{vmatrix}$$

and M is a smoothness constant that can be adjusted to trade "fit" for "smoothness." The constraints (b) through (e) can be written as:

(b′) Increasing with age: $\mathbf{B}\delta \geq 0$, where

$$\mathbf{B} = \begin{vmatrix} -1 & 1 & 0 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & -1 & 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & -1 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\ 0 & 0 & 0 & 0 & 0 & \cdot & \cdot & -1 & 1 \end{vmatrix}$$

(c″) More steeply increasing at the higher ends of the age range (convexity): $\mathbf{C}\delta > \mathbf{0}$, where

$$\mathbf{C} = \begin{vmatrix} 1 & -2 & 1 & 0 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 1 & 2 & 1 & 0 & \cdot & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 2 & 1 & \cdot & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\ \cdot & \cdot & \cdot & \cdot & \cdot & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -2 & 1 \end{vmatrix}$$

(d′) The graduated number of deaths equals the observed number of deaths: $\mathbf{D}'\delta = \mathbf{D}'\mathbf{u}$, where $\mathbf{D} = (l_{70}, l_{71}, \ldots, l_{84})'$, and l_x is the number exposed to risk at age x.

(e′) The total graduated ages at death equals the observed total ages at death: $\mathbf{E}'\delta = \mathbf{E}'\mathbf{u}$, where $\mathbf{E} = (70 \cdot l_{70}, 71 \cdot l_{71}, \ldots, 84 \cdot l_{84})'$.

The computation of the MDI estimate δ (the graduated mortality rates) in the mathematical programming problem developed above can be numerically evaluated by using any of a number of nonlinear programming computer codes.[7] Table 4 presents the results obtained using the information theoretic mortality table graduation technique on the crude mortality rates from Table 3 and the smoothness constant $M = 2 \times 10^{-4}$. The ad-hoc graphical graduation of London [55] is included for comparison.

Figure 1 presents a graphical representation of the graduated mortality rates from Table 4. As is readily apparent, the information theoretic method is easily computerized. This method yields graduated series as output that are remarkably similar to those which would be obtained by a very experienced graduation actuary. Moreover, unlike many other graduation techniques, this information theoretic technique can be extended to the multivariate setting without any theoretical difficulty. This extension is outlined in the next section.

12. MULTIVARIATE GRADUATION

The multivariate graduation problem poses many difficulties for most traditional multivariate techniques such as Whittaker-Henderson graduation.

[7] In our analysis, we used GRGII computer code due to L. Lasdon. This technique (generalized reduced gradient) has been discussed in a number of articles, such as Lasdon et al. [52] and Liebman et al. [54] and is available on the PC as well as the mainframe computer. Using a CDC 6600 computer, the computation time for our graduation in the above problem was less than 10 seconds CPU time. Other nonlinear programming methods such as the Sequential Unconstrained Minimization Techniques (SUMPT) by Fiacco and McCormick [30] could also be used. Note that Zhang and Brockett [79] present the duality states for this general inequality quadratic constraint case and give the computational form for the dual parameters.

FIGURE 1

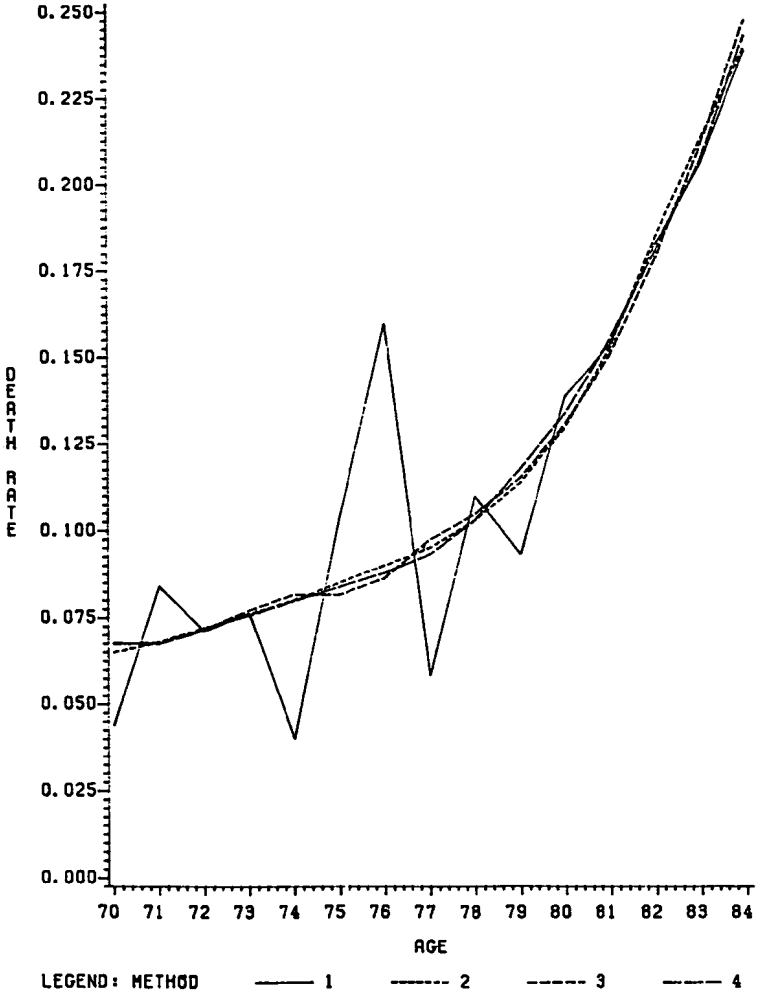

0.250
0.225
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
DEATH RATE
70 71 72 73 74 75 76 77 78 79 80 81 82 83 84
AGE
LEGEND: METHOD 1 2 3 4
1: ORIGINAL DATA
2: LONDON S GRADUATION
3: MDI WITH 3RD ORDER DIFFERENCE CONSTRAINT ONLY
4: MDI WITH 3RD ORDER DIFFERENCE AND CONVEX CONSTRAINT

TABLE 4

GRADUATED RATES

Age	Undergraduated Rate	London's Graduation	Information Theoretic Graduation with 3rd Difference Constraint	Information Theoretic Graduation with Convexity and 3rd Difference Constraints
70	0.044	0.065	0.06766	0.06752
71	0.084	0.068	0.06756	0.06756
72	0.071	0.072	0.07143	0.07161
73	0.076	0.076	0.07696	0.07565
74	0.040	0.080	0.08154	0.07970
75	0.104	0.085	0.08144	0.08375
76	0.160	0.090	0.08627	0.08779
77	0.058	0.095	0.09734	0.09316
78	0.110	0.103	0.10479	0.10297
79	0.093	0.114	0.11539	0.11795
80	0.139	0.130	0.13102	0.13405
81	0.154	0.153	0.15094	0.15549
82	0.183	0.185	0.17897	0.18121
83	0.206	0.213	0.21094	0.20692
84	0.239	0.240	0.24774	0.24358

TABLE 5

Selection Age	Years since Selection				
	0	1	2	3	Ultimate
x	$\delta_{[x]}$	$\delta_{[x]+1}$	$\delta_{[x]+2}$	$\delta_{[x]+3}$	δ_{x+4}
$x+1$	$\delta_{[x+1]}$	$\delta_{[x+1]+1}$	$\delta_{[x+1]+2}$	$\delta_{[x+1]+3}$	δ_{x+5}
$x+2$	$\delta_{[x+2]}$	$\delta_{[x+2]+1}$	$\delta_{[x+2]+2}$	$\delta_{[x+2]+3}$	δ_{x+6}
$x+3$	$\delta_{[x+3]}$	$\delta_{[x+3]+1}$	$\delta_{[x+3]+2}$	$\delta_{[x+3]+3}$	δ_{x+7}
$x+4$	$\delta_{[x+4]}$	$\delta_{[x+4]+1}$	$\delta_{[x+4]+2}$	$\delta_{[x+4]+3}$	δ_{x+8}
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

For example, consider graduating a bivariate select and ultimate mortality table of mortality rates $\{\delta\}$ such as those illustrated in Table 5.

It is desirable to obtain a graduated series δ_x in which several constraints (such as monotonicity in rates down the columns, monotonicity of rates along the rows, and also monotonicity along the upward diagonals) hold simultaneously. It is usually difficult to develop a single method that achieves these monotonicity constraints simultaneously, and often an ad-hoc adjustment or a sequence of ad-hoc adjustments are made iteratively to the initial graduation in order to develop the desired graduation. Usually it is considered too difficult to also insist upon convexity along rows and columns, smoothness, and average-age-at-death type of constraints.

The information theoretic graduation technique can accomplish the desired goals. For convenience of notation, let $\delta_{[x+j]+k}$ denote the graduated mortality rate for an individual selected at age $x+j$ who has attained age $x+j+k$. The row, column, and diagonal forward difference operators are defined by

$$\Delta_j(\delta_{[x+j]+k}) = \delta_{[x+j+1]+k} - \delta_{[x+j]+k},$$

$$\Delta_k(\delta_{[x+j]+k}) = \delta_{[x+j]+k+1} - \delta_{[x+j]+k},$$

and

$$\Delta_{jk}(\delta_{[x+j]+k}) = \delta_{[x+j-1]+k+1} - \delta_{[x+j]+k}.$$

If $u_{[x+j]+k}$ represents the corresponding crude (observed) mortality rate, then the desired graduated series can be obtained by solving the convex programming problem:

$$\min_{\delta} l(\delta|\mathbf{u}) = \sum_{j\&k} \delta_{[x+j]+k} \ln \left[\frac{\delta_{[x+j]+k}}{u_{[x+j]+k}} \right]$$

subject to

$$\Delta_j(\delta_{[x+j]+k}) > 0 \qquad \text{for all } k,$$

$$\Delta_k(\delta_{[x+j]+k}) > 0 \qquad \text{for all } j,$$

and

$$\Delta_{jk}(\delta_{[x+j]+k}) > 0 \qquad \text{for all } j \text{ and } k.$$

The convexity and smoothness constraints can also be inserted in a manner exactly analogous to that done in Section 11.

13. CONCLUSIONS

This paper presents a single easily understood philosophical approach to modeling and analyzing data that, for the first time, unifies several different fundamental areas of actuarial science. This technique, constrained information theoretic analysis, provides a non-Bayesian statistical method for approaching numerous problems of interest to the actuarial community. Several examples are explicitly worked out that extend or improve upon existing actuarial methods. At the same time, these methods provide firm statistical foundations and clear signposts for subsequent actuarial applications.

14. ACKNOWLEDGMENT

This research was supported by the Actuarial Education and Research Fund, whose assistance is gratefully acknowledged.

REFERENCES

1. AJNE, B. "A Note on the Multiplicative Ratemaking Model," *ASTIN Bulletin* 8 (1974): 144–53.
2. AKAIKE, H. "Information Theory and an Extension of the Maximum Likelihood Principle." In *2nd International Symposium on Information Theory.* Edited by B.N. Petrov and F. Csaki. Budapest: Akademiai Kiaido, 1973, 267–81.
3. AKAIKE, H. "An Extension of the Method of Maximum Likelihood and the Stein's Problem," *Annals of the Institute of Statistical Mathematics* 29, Part A (1977): 153–64.
4. AKAIKE, H. "A New Look at the Bayes Procedure," *Biometrika* 65, no. 1 (1978): 53–59.
5. ALMER, B. "Risk Analysis in Theory and Practical Statistics," *Transactions of the 15th International Congress of Actuaries* 2 (1957): 314–53.
6. BELLHOUSE, D.R., AND PANJER, H.H. "Stochastic Modelling of Interest Rates with Applications to Life Contingencies," *Journal of Risk and Insurance* 67 (1980): 91–610.
7. BERKSON, J. "Minimum Discrimination Information, the 'No-Interaction' Problem, and the Logistic Function," *Biometrics* 28, no. 2 (1972): 443–68.
8. BISHOP, Y.M.M., FIENBERG S., AND HOLLAND, P.W. *Discrete Multivariate Analysis.* Cambridge, Mass.: MIT Press, 1975.
9. BORCH, K. "The Utility Concept Applied to the Theory of Insurance," *ASTIN Bulletin* 1 (1961): 245–55.
10. BORCH, K. *The Economics of Uncertainty.* Princeton, N.J.: Princeton University Press, 1968.
11. BORCH, K. "Risk Theory and Serendipity," *Insurance: Mathematics and Economics* 5, no. 1 (1986): 103–12.
12. BOWERS, N.L., JR., ET AL. *Actuarial Mathematics.* Itasca, Ill.: Society of Actuaries, 1986.
13. BRIYS, ERIC P. "Investment Portfolio Behavior of Non-Life Insurers: A Utility Analysis," *Insurance: Mathematics and Economics* 4, no. 2 (1985): 93–98.
14. BROCKETT, P.L. "Using a Standard Distribution and Client Data to Obtain a Client Loss Distribution," *Conference of Actuaries in Public Practice* 34 (1985): 503–11.
15. BROCKETT, P.L., CHARNES, A., AND COOPER, W.W. "MDI Estimation via Unconstrained Convex Programming," *Communications in Statistics* B9, no. 3 (1980): 223–34.
16. BROCKETT, P.L., CHARNES, A., GOLDEN, L., AND PAICK, K. "A Method for Constructing a Unimodal Prior Distribution," *Center for Cybernetic Studies Research Report,* in preparation.

17. BROCKETT, P.L., AND COX, S. "Statistical Adjustment of Mortality Tables to Reflect Known Information" *TSA* 36 (1984): 63–71; Discussion, 73–75.
18. BROCKETT, P.L., AND ZHANG, J. "Information Theoretic Mortality Table Graduation," *Scanadinavian Actuarial Journal* (1986): 131–40.
19. BROFFITT, J.D. "A Bayes Estimator or Ordered Parameters and Isotonic Bayesian Graduation," *Scandinavian Actuarial Journal* (1984): 231–47.
20. BURDEN, R.L., AND FAIRES, J.D. *Numerical Analysis*. 4th ed. Boston, Mass.: PWS Publishers, 1989.
21. CHANG, L., AND FAIRLEY, W. "Pricing Automobile Insurance under Multivariate Classification of Risks: Additive Versus Multiplicative," *Journal of Risk and Insurance* 46 (1979): 73–96.
22. CHARNES, A., AND COOPER, W.W. *Management Models and Industrial Applications of Linear Programming*. New York: John Wiley & Sons, Inc., 1961.
23. CHARNES, A., AND COOPER, W.W. "An Extremal Principle for Accounting Balance of a Resource-Value Transfer Economy; Existence, Uniqueness and Computations," *Rendiconti di Accademia Nazionale dei Lincei* (1974): 556–78.
24. CHARNES, A., COOPER, W.W., AND SEIFORD, L. "Extremal Principles and Optimization Dualities for Khinchine-Kullback-Leibler Estimation," *Math. Operationsforsch. Statist., Ser. Optimization* 9, no. 1 (1978): 21–29.
25. CHARNES, A., COOOPER, W.W., AND TYSSEDAL, J. "Khinchine-Kullback-Leibler Estimation with Inequality Constraints," *Math. Operationsforsch. Statist., Ser. Optimization* 14 (1983): 1–4.
26. COUTTS, S. "Motor Premium Rating," *Insurance: Mathematics and Economics* 3 (1984): 73–96.
27. COX, D.R., AND HINKLEY, D.V. *Theoretical Statistics*. London: Chapman and Hall, 1974.
28. ELANDT-JOHNSON, R.C., AND JOHNSON, N.L. *Survival Models and Data Analysis*. New York: John Wiley and Sons, 1980.
29. FAIRLEY, W., THOMBERLIN T., AND WEISBERG, H. "Pricing Automobile Insurance under a Cross-Classification of Risks: Evidence from New Jersey," *Journal of Risk and Insurance* 48 (1981): 505–41.
30. FIACCO, A., AND MCCORMICK, G.P. *Nonlinear Programming: Sequential Unconstrained Minimization Techniques*. New York: John Wiley & Sons Inc., 1968.
31. FISHER, R.A. "The Logic of Inductive Inference," *Journal of the Royal Statistical Society Vol. 98. (1935): Contributions to Mathematical Statistics*. New York: John Wiley & Sons, 1950, paper 26.
32. GOKHALE, D.V., AND KULLBACK, S. *The Information in Contingency Tables*. New York: Marcel Dekker, 1978.
33. GOKHALE, D.V., AND KULLBACK, S. "The Minimum Discrimination on Information Approach in Analyzing Categorical Data," *Communications in Statistics—Theory & Methods* A7, no. 10 (1978): 987–1005.

34. GOLDEN, L.L., BROCKETT, P.L., AND ZIMMER, M. "An Information Theoretic Approach for Identifying Shared Information and Asymmetric Relationships among Variables," *Multivariate Behavioral Research* 25 (1990): 479–502.
35. GREEN, P.E., CARMONE F.J., AND WACHSPRESS, D.P. "On the Analysis of Qualitative Data in Marketing Research," *Journal of Marketing Research* XIV (Feb. 1977): 52–59.
36. GUIASU, S. *Information Theory with Application*. Great Britain: McGraw-Hill Book Co., Inc., 1977.
37. HAALAND, P.D., BROCKETT, P.L., AND LEVINE, A. "A Characterization of Divergence with Applications to Questionnaire Information," *Information and Control* 41, (1979): 1–8.
38. HARRINGTON, S. "Estimation and Testing for Functional Form in Pure Premium Regression Models," *ASTIN Bulletin* 16 (1986): S31–43.
39. HAYNES, K.E., PHILLIPS, F.Y., AND MOHRFELD, J.W. "The Entropies: Some Roots of Ambiguity," *Socio-Economic Planning Sciences* 14, no. 3 (May 1980): 137–45.
40. HEILMANN, W.-R. "Decision Theoretic Foundations of Credibility Theory," *Insurance: Mathematics and Economics* 8, 1 (1989): 77–95.
41. HILLIER, F., AND LEIBERMAN, G. *Introduction to Operations Research*. 4th ed. New York: McGraw-Hill Book Co., and McGraw-Hill Ryerson, Ltd., 1986.
42. HOEM, JAN M. "A Markov Chain Model of Working Life Tables," *Scandinavian Actuarial Journal* (1977): 1–20.
43. HURLIMANN, W. "A Numerical Approach to Utility Functions in Risk Theory," *Insurance: Mathematics and Economics* 6, no. 1 (1987): 19–31.
44. JOHNS, M.V. "Discussion of Buhlmann's Paper 'Minimax Credibility'." In *Credibility: Theory and Applications*. Edited by P.M. Kahn. New York: Academic Press, 1975.
45. JUNG, J. "On Automobile Insurance Ratemaking," *ASTIN Bulletin* 5 (1968): 41–48.
46. KAHN, P.M. "Credibility: Theory and Applications." In *Proceedings of Actuarial Research Conference on Credibility*. Berkeley, California, New York: Academic Press, Sept. 1974.
47. KALBFLEISCH, J.G. "Comments on Trending Methods in Automobile Insurance Ratemaking," *Actuarial Research Clearing House* 1983.2: 253–77.
48. KHINCHINE, A.I. *Mathematical Foundations of Statistical Mechanics*. New York: Dover Publishers, 1948.
49. KHINCHINE, A.I. *Mathematical Foundations of Information Theory*. New York: Dover Publ. Co., 1957. (New translation of Khinchine's papers "The Entropy Concept in Probability Theory" and "On the Fundamental Theorems of Information Theory," originally published in Russian in *Uspekni Matematicheskikh* 7, no. 3 (1953) and 11, no. 1 (1956), respectively.)
50. KULLBACK, S. *Information Theory and Statistics*. New York: John Wiley, 1959.

51. KULLBACK, S., AND LEIBLER, R.A., "On Information and Sufficiency," *Annals of Mathematical Statistics* 22 (1951): 79–86.
52. LASDON, L.S., WAREN, A.D., JAIN, A., AND RATNER, M. "Design and Testing of a Generalized Reduced Gradient Code for Non-Linear Programming," *ACM Transactions on Mathematical Software* 4, no. 1 (1978): 34–50.
53. LARIMORE, W.E. "Predictive Inference, Sufficiency, Entropy and an Asymptotic Likelihood Principle," *Biometrika* 70 (1983): 175–81.
54. LIEBMAN, J., LASDON, L.S., SCHRANGE, L., AND WAREN, A. *Introduction to Modeling and Optimization in GINO*. Redwood, Calif.: Scientific Press, 1986.
55. LONDON, D. *Graduation: The Revision of Estimates*. Winsted, Conn.: Actex Publications, 1985.
56. LONDON, D. *Survival Models and Their Estimation*. 2nd ed. Winsted, Conn.: Actex Publications, 1988.
57. MARTIN, J.D., COX, S.H., AND MACMINN, R.D. *The Theory of Finance—Evidence and Applications*, New York: The Dryden Press, 1988.
58. MARTIN-LOF, A. "Entropy Estimates for Ruin Probabilities." In *Probability and Mathematical Statistics*. Edited by A. Gut and L. Holst. Uppsala, Sweden: Dept. Mathematics, Uppsala University, 1983.
59. MARTIN-LOF, A. "Entropy Estimates for the First Passage Time of a Random Walk to a Time Dependent Barrier," *Scandinavian Journal of Statistics* 13 (1986): 221–29.
60. MILLER, M.D. *Elements of Graduation*. New York: Actuarial Society of America, 1949.
61. MILLER, R., AND WICHERN, D. *Intermediate Business Statistics*. New York: Holt Reinhart & Winston Publ. Co., 1977.
62. MURRAY, G.D. "A Note on the Estimation of Probability Density Functions," *Biometrika* 64 (1977): 150–52.
63. NG, V.M. "On the Estimation of Parametric Density Functions," *Biometrika* 67 (1980): 505–506.
64. PANJER, H.H., AND BELLHOUSE, D.R. "Stochastic Modelling of Interest Rates with Applications to Life Contingencies," *Journal of Risk and Insurance* 47 (1980): 91–110.
65. PHILLIPS, F.Y. "A Guide to MDI Statistics for Planning and Management Model Building," *Institute for Constructive Capitalism Technical Series #2*, University of Texas at Austin, 1980.
66. PROMISLOW, S.D. "Measurement of Equity," *TSA* XXXIX (1987): 215–37.
67. SAKAMOTO, Y., ISHIGURO, M., AND KITAGAWA, G. "Akaike Information Criterion Statistics." In *Mathematics and Its Applications*. Tokyo, Japan: Institute of Statistical Mathematics, 1986.
68. SAMPSON, D., AND THOMAS, H. "Claim Modeling in Auto Insurance." Urbana, Ill: Department of Business Administration, University of Illinois, 1984.

69. SANT, D. "Estimating Expected Losses in Auto Insurance," *Journal of Risk and Insurance* 47 (1980): 133–51.
70. SCHOEN, R., AND LAND, K.C. "A General Algorithm for Estimating a Markov-Generated Increment-Decrement Life Table with Applications to Marital-Status Patterns," *Journal of American Statistical Association* 74, no. 368 (1979): 761–76.
71. SCHWARZ, G. "Estimating the Dimension of a Model," *Annals of Statistics* 6 (1978): 461–64.
72. SHANNON, C., AND WEAVER, W. *The Mathematical Theory of Communication*. Urbana, Ill.: University of Illinois Press, 1949.
73. SMITH, A.F.M., AND SPIEGELHALTER, D.J. "Bayes Factors and Choice Criteria for Linear Models," *Journal of the Royal Statistical Society* B 42 (1980): 213–20.
74. STEENACKERS, A., AND GOOVAERTS, M.J. "A Credit Scoring Model for Personal Loans," *Insurance: Mathematics and Economics* 8, no. 1 (1989): 31–34.
75. TENENBEIN, A., AND VANDERHOOF, I.T. "New Mathematical Laws of Select and Ultimate Mortality," *TSA* XXXII (1980): 119–59.
76. WEISBERG, H., THOMBERLIN, T., AND CHATTERJEE, S. "Predicting Insurance Losses under Cross-Classification: A Comparison of Alternative Approaches," *Journal of Business and Economic Statistics* 2 (1984): 170–78.
77. WIENER, N. *Cybernetics*. New York: John Wiley and Sons Inc., 1948.
78. ZELLNER, A. "Optimal Information Processing and Bayes' Theorem," *The American Statistician*, 42, no. 4 (Nov. 1988): 278–80.
79. ZHANG, J. AND BROCKETT, P.L. "Quadratically Constrained Information Theoretic Analysis," *SIAM Journal of Applied Mathematics* 47, no. 4 (August 1986): 871–85.

DISCUSSION OF PRECEDING PAPER

BRADLEY P. CARLIN:

First, congratulations to Dr. Brockett on a fine paper, which pulls together a great many results from the information theory literature and shows how they may be fruitfully applied to problems in actuarial science. My remarks on the paper's approach to model selection are from a more fully Bayesian point of view.

In Section 3, the discussion dwells on the Akaike Information Criterion (AIC), while only briefly mentioning the Schwarz criterion, or Bayesian Information Criterion (BIC), as given by Schwarz [3]. The author discards BIC as a model selection criterion because of its "ad-hoc nature" and because it is not equal to the Akaike criterion. This dismissal seems premature, especially in view of the attempt made in Section 4 to link the author's approach with the traditional Bayesian methodology. In the notation of this section, suppose we have parametric models H_1 and H_2 for data x, and the two models have respective parameter vectors λ_1 and λ_2. Under prior densities $\pi_k(\lambda_k)$, $k=1,2$, the marginal distributions of x are found by integrating out the parameters,

$$p(x|H_k) = \int p(x|\lambda_k,H_k)\pi_k(\lambda_k)d\lambda_k, \qquad k = 1,2.$$

Bayes' Theorem may then be applied to obtain posterior probabilities $P(H_1|x)$ and $P(H_2|x)=1-P(H_1|x)$ for the two models. The quantity commonly used to summarize these results is the Bayes factor, B, which is the ratio of posterior to prior odds in favor of H_1 relative to H_2 indicated by the data, and is given by

$$B = \frac{p(H_1|x)/p(H_1)}{p(H_2|x)/p(H_2)} = \frac{p(x|H_1)}{p(x|H_2)},$$

the ratio of the observed marginal densities for the two models. Assuming the two models are *a priori* equally probable,

$$B = P(H_1|x)/[1 - P(H_1|x)],$$

the posterior odds in favor of H_1. Now, Schwarz [3] showed that for large sample sizes n, an approximation to $-2 \log B$ is given by

$$\Delta BIC = W - (p_2 - p_1)\log n, \tag{1}$$

where p_1 and p_2 are the numbers of parameters in models 1 and 2, respectively, and W is the usual likelihood ratio test statistic,

$$W = -2 \log[sup_{\lambda 1} L(\lambda_1;x)/sup_{\lambda 2} L(\lambda_2;x)].$$

Although the approximation in (1) is rather crude, Schwarz [3] showed that ΔBIC does tend to $-2 \log B$ as n tends to infinity. Thus the definition of BIC is not ad-hoc, but rather is done deliberately to satisfy this asymptotic property. The criterion also enjoys the property of being independent of the chosen prior distributions $\pi_k(\lambda_k)$, $k = 1,2$. As such, Bayesians have long used ΔBIC as a quick way to discover the relative weight of evidence between two models (although the recent increasing availability of computing power has lessened its importance somewhat).

Now, analogous to Equation (1), the change in AIC can be written in our notation as

$$\Delta AIC = W - (p_2 - p_1). \tag{2}$$

Clearly as n increases, ΔAIC and ΔBIC will obtain different results. Raftery [1] remarks that using AIC is asymptotically equivalent to choosing the model with the highest posterior probability *only* when the information in the prior increases at the same rate as the information in the likelihood, a rather unusual assumption. Shibata [4], working in the area of time series, pointed out a bias in AIC toward overparametrization, and indeed applied researchers in many fields have noticed that the Akaike criterion does tend to "keep too many terms in the model." To get a rough idea in our setting of why this might be the case, imagine the usual nested model setting where H_2 is the "full" model while H_1 is the "reduced" model. Then the term subtracted from W in the expressions for ΔAIC and ΔBIC is a penalty term that corrects for the advantage the full model naturally enjoys over the smaller reduced model. The aforementioned work of Schwarz [3] shows that the penalty in Equation (2) is too small when n is large.

Of course, various arguments on behalf of AIC may be made as well, but this inconsistency of AIC for $-2 \log B$ leaves many Bayesians unconvinced, preferring the Bayes factor itself or at the least the crude approximation offered by the Schwarz criterion. Linking the frequentist and Bayesian approaches to statistical inference in the presence of a set of competing models is a difficult topic that continues to generate research interest; examples include the recent papers by Rissanen [2] and Woodroofe [5].

REFERENCES

1. RAFTERY, A.E. "Approximate Bayes Factors for Generalized Linear Models," *Technical Report 121*, University of Washington, Department of Statistics, 1988.
2. RISSANEN, J. "Stochastic Complexity" (with discussion), *Journal of the Royal Statistical Society, Series B,* 49 (1987): 223–39 and 252–65.
3. SCHWARZ, G. "Estimating the Dimension of a Model," *Annals of Statistics* 6 (1978): 461–64.
4. SHIBATA, R. "Selection of the Order of an Autoregressive Model by Akaike's Information Criterion," *Biometrika* 63 (1976): 117–26.
5. WOODROOFE, M."On Model Selection and the Arc Sine Laws," *Annals of Statistics* 10 (1982): 1182–94.

E.S. ROSENBLOOM* AND ELIAS S.W. SHIU:

Dr. Brockett has given a fine exposition of information theory and its actuarial applications.

We have a comment on Sections 11 and 12 of the paper. In the earlier sections, the arguments $\mathbf{p}$ and $\mathbf{q}$ of the function

$$l(\mathbf{p}|\mathbf{q}) = \sum_i p_i \ln(p_i/q_i)$$

have an identical sum, that is,

$$\mathbf{1}'\mathbf{p} = \mathbf{1}'\mathbf{q}.$$

However, it does not seem reasonable to require that

$$\mathbf{1}'\boldsymbol{\delta} = \mathbf{1}'\mathbf{u}.$$

In view of constraint (d'), we suggest that the objective function be formulated as

$$\text{Min } l(\mathbf{D}'\boldsymbol{\delta}|\mathbf{D}'\mathbf{u}) = \sum_i l_i \, \delta_i \ln(\delta_i/u_i).$$

We would also like to know how the value $M = 2 \times 10^{-4}$ for constraint (a') is determined.

We wish to suggest that a potential application of the methodology presented in the paper is the problem of cash-flow matching. Various methods for cash-flow matching and dedication can be found in the *TSA* papers [4] and [5] and their references and in the papers [1], [2] and [3]. None of these uses the information theoretic approach.

*Dr. Rosenbloom, not a member of the Society, is Associate Professor, Department of Actuarial and Management Sciences, University of Manitoba.

REFERENCES

1. FABOZZI, T.D., TONG, T., AND ZHU, Y. "Extensions of Dedicated Bond Portfolio Techniques." In *The Handbook of Fixed Income Securities,* 3rd ed., edited by F.J. Fabozzi, T.D. Fabozzi and I.M. Pollack. Homewood, Illinois: Dow Jones–Irwin, 1991, 959–71.
2. FABOZZI, T.D., TONG, T., AND ZHU, Y. "Symmetric Cash Matching," *Financial Analysts Journal* 46, no. 5 (September-October 1990): 46–52.
3. HILLER, R.S., AND SCHAACK, C. "A Classification of Structured Bond Portfolio Modeling Techniques," *Journal of Portfolio Management* 17, no. 1 (Fall 1990): 37–48.
4. KOCHERLAKOTA, R., ROSENBLOOM, E.S., AND SHIU, E.S.W. "Algorithms for Cash-Flow Matching," *TSA* XL, Part 1 (1988): 477–84.
5. KOCHERLAKOTA, R., ROSENBLOOM, E.S., AND SHIU, E.S.W. "Cash-Flow Matching and Linear Programming Duality," *TSA* XLII (1990): 281–93.

THOMAS N. HERZOG:

I thank Dr. Brockett for writing a stimulating and informative paper. My only point is that all the results in his paper are based on the assumption that the minimum discrimination information statistic

$$p(x) \ln [p(x)/q(x)]$$

is the only loss function that needs to be considered. For certain applications, the actuarial analyst may want to examine the sensitivity of his/her results to alternate loss functions. Sometimes the conclusions are highly sensitive to the choice of loss function selected. One such example of this is described in Herzog [1].

REFERENCE

1. HERZOG, T.N. Discussion of Harry H. Panjer, "AIDS: Exponential vs. Polynomial Growth Models," *Insurance: Mathematics and Economics* 9 (1990): 291–93.

WOJCIECH SZATZSCHNEIDER*:

The concept of entropy is an important and useful tool in many branches of theoretical and applied mathematics, for example, the Boltzmann H-Theorem and its impact on modern probability theory or the Kolmogorov-Ornstein theorem for dynamic flows. Incidentally, the Boltzmann H-Theorem is getting new life in view of recent investigations of molecular chaos. In risk

*Dr. Szatzschneider, not a member of the Society, is Professor of Probability and Risk Theory, School of Actuarial Sciences, Universidad Anahuac, Mexico City.

theory, entropy is proving to be useful for estimating the probability of ruin; see, for example, Martin-Löf [4]. Now is just the right time to apply entropy or, more generally, the information theoretic approach to actuarial practice. The paper by Dr. Brockett aims precisely in this direction, and it fulfills this job admirably.

The main advantage of the information theoretic approach is that it is a simple method of adjusting the probability distribution due to new information; it is like Bayesian statistics, but it is classical. Instead of the pass (transformation)

$$\Pi(\omega) \rightarrow \Pi(\omega|\text{information})$$

from a prior distribution to the posterior one, we have

$$P \rightarrow P(^*|\text{information})$$

from one probability distribution to another.

This approach will be welcomed by orthodox frequentist statisticians and, possibly, will not be rejected by, at least, the less orthodox Bayesians. There are important links between strict Bayesian statistics and the theory of information. The power and efficiency of the information theoretic approach result, not from the conflict between frequentist and Bayesian statisticians, but mainly from its simple presentation.

I am more a probabilist than a statistician, but even so, I would like to write a few comments about this conflict in an actuarial context. The failure in applying Bayesian statistics to the solution of practical problems frequently results from the lack of good training; see Efron [1] or Lindley [3], both with discussions. In the discussion of Lindley's paper, a good part of the criticism was derived from the proposed exchangeability. In actuarial problems, this exchangeability is usually assumed (Goovaerts et al. [2]), especially if we apply credibility theory. Even a pure Bayesian approach, based on the prior subjective probabilities, should not find many opponents, if we accept that generally no statistics procedure is free from defects.

In the information theoretic approach, the situation is quite different. One also needs good training, but it is an easy task, and the method should have a real and immediate impact in applications. There are many antecedents but most were probably reduced to investigations and not to actuarial practice. For example, no one in Mexico is using this approach.

My only disagreement is with the statement that under constrained maximum entropy, "We select the distribution that is as close to uniform as possible subject only to these given constraints." We often *must* use

constraints in order to be able to solve a given problem, for example, considering distributions over an infinite interval. It is not the case of an improper uniform distribution used by Bayesians. Even for probabilities concentrated over a finite interval, constraints form an intrinsic part of a problem and should not be considered of secondary importance.

To clarify my point, I propose a different proof of the theorem on page 15. This proof does not need Lagrange multipliers (to sell an idea, one should present it in the simplest way). We use the so-called maximization entropy principle, instead of an explicit use of Jensen's inequality. Start with the elementary inequality

$$x \ln x - x \ln y - x + y \geq 0, x, y > 0.$$

Therefore, if $f(x)$ and $g(x)$ are probability densities, then

$$\int f(x) \ln f(x)\, dx \geq \int f(x) \ln g(x)\, dx.$$

Now we say that $f \in H$ if

$$\int a(x) f(x) dx = \theta,$$

where $a(x)$ and θ might be vectors, like they usually are in the author's presentation. The result is that the density $f_M(x)$, which minimizes $\int f(x) \ln f(x)\, dx$ or maximizes the entropy $-\int f(x) \ln f(x)\, dx$ for densities $\in H$, is given by

$$f_M(x) = \exp \Sigma a_i(x) Z_i.$$

Note that $\int g(x) \ln f_M(x)\, dx$ does not depend on $g \in H$, and so

$$\int f_M(x) \ln f_M(x)\, dx = \int g(x) \ln f_M(x)\, dx \leq \int g(x) \ln g(x) dx.$$

For example, in the case of a distribution over $(0, \infty)$ with a fixed mean, we find that maximal entropy

$$-\int f_M(x) \ln f_M(x)\, dx$$

is obtained for exponential distribution. For other constraints, other probability distributions result. Almost the same proof can be given in the author's case (p. 87).

It is enough to write

$$\int_{-\infty}^{+\infty} p(x) \ln \frac{p(x)}{q(x)}\, dx = \int_{-\infty}^{+\infty} \frac{p(x)}{q(x)} \ln \frac{p(x)}{q(x)}\, q(x)\, dx = \int_{-\infty}^{+\infty} \frac{p(x)}{q(x)} \ln \frac{p(x)}{q(x)}\, \lambda(dx)$$

with $\lambda(dx) = q(x)dx$. This expression is minimized just by $p(x)/q(x)$ given by the author. In this proof, which is not original, the role of constraints is more explicit. Finally, I hope that the information theoretic approach will spread quickly, and its applications will not be postponed for another decade.

REFERENCES

1. EFRON, B. "Why Isn't Everyone a Bayesian?" *The American Statistician* 40 (1986): 1–11.
2. GOOVAERTS, M.J., KAAS, R., VAN HEERWAARDEN, A.E., AND BAUWELINCKX, T. *Effective Actuarial Methods*. New York: North-Holland Division of Elsevier Science Publishing Co., 1990.
3. LINDLEY, D.V., AND SMITH, A.F.M. "Bayes Estimates for the Linear Model," *Journal of the Royal Statistical Society, Series B,* 34 (1972): 1–41.
4. MARTIN-LÖF, A. "Entropy, a Useful Concept in Risk Theory," *Scandinavian Actuarial Journal* 69 (1986): 223–35.

WILLIAM L. ROACH:

Section 10 of the paper characterizes a mortality adjustment as the solution to an optimization problem

$$\text{Min } l(\mathbf{p}|\mathbf{q}) = \sum_0^{\omega} p_i \ln \frac{p_i}{q_i}$$

subject to

$$1 = \sum_0^{\omega} p_i$$

$$9 = \sum_0^{\omega} i p_i.$$

The q_i's represent the probability of death i years from now, during age $50+i$. The solution to the optimization problem is the best way of adjusting the q_i to reflect the known constraints. The form of the solution is given as

$$p_i = q_i \exp(\beta_0 - 1) \exp(i\beta_1)$$

$$p_i = q_i (1.57823) (0.9629889)^i.$$

The form of the problem is artificial and conceals the form of the natural solution. An actuary or a physician considering an individual whom he or she expects to experience nonstandard mortality is likely to express that

intuition as a mortality ratio. One assumes nonsmokers experience a mortality that is 50 percent of the standard mortality. An impaired life might experience a mortality of 150 percent of standard mortality. Determining a mortality ratio is the natural solution to the optimization problem given above.

The form of the adjustment looks different from a simple mortality ratio, because the problem is defined in terms of q_i, the probability of death i years from now, during age $50+i$, rather than the probability of death at the attained age.

The assumed form of the solution

$$p_i = q_i\,(1.57823)\,(0.9629889)^i$$

has a great deal to do with the answer. In fact, *there is no optimization going on here*. The formula for the adjustment has two parameters with two constraints. The probabilities must sum to 1, and the expected value must be 9. Two nonredundant equations in two unknowns determine a unique solution. Figure 1 shows lines corresponding to the β_0 and β_1 pairs, which give probabilities summing to 1, and the β_0 and β_1 pairs, which give expected values of 9. The intersection of those lines at (1.57823, 0.9629889) is the solution given in the paper. There are no degrees of freedom left to optimize over.

The same analysis applies to the example presented in Section 9 of the paper.

I tried to solve the optimization problem as a series of linear programming problems. The coefficients of the objective function were $\ln(p_i/q_i)$. Initially, I calculated

$$p_i = \frac{\mu'_i}{\mu_i} * q_i$$

The resulting p_i's typically do not sum to 1. The solution to the first iteration will assure p_i's that sum to 1 and have the appropriate expected value.

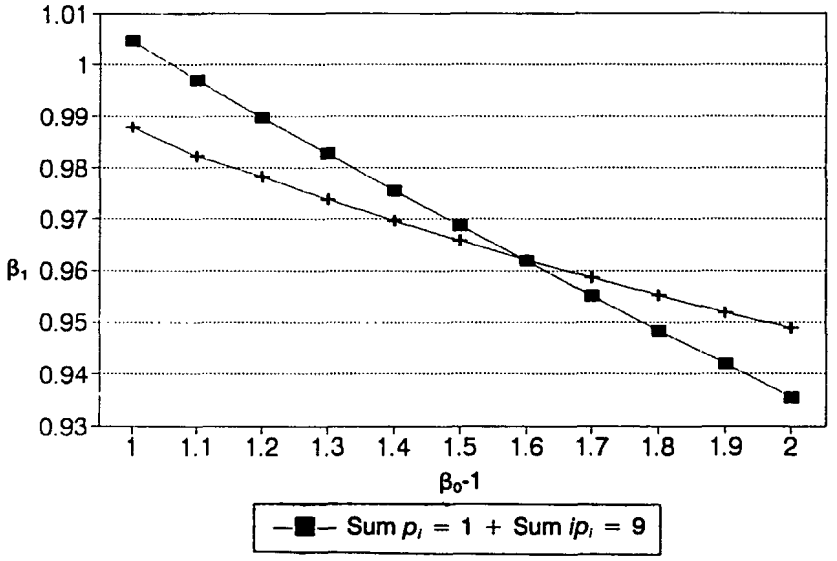
As formulated, the constraints of the problem are linear.

$$1 = \sum_0^\omega p_i$$

$$9 = \sum_0^\omega i p_i$$

FIGURE 1

SOLVING FOR ADJUSTMENT
NO OPTIMIZATION



With no additional constraints, the solutions were very disappointing. Only two p_i need to be positive to satisfy the constraints and minimize the initial objective function. The simple iterative linear programming approach did not converge to an appropriate solution.

I introduced additional constraints into the linear programming formulation

$$p_i - p_{i+1} \geq 0.$$

Frequently, this assumption does not hold true at the end of the mortality table.

Again the results of this iterative linear programming approach were disappointing. For the mortality example in Section 10, this analysis was beyond the numerical capability of my linear programming software (*Quattro Pro* Version 3.0). For the disability example in Section 9, the coefficients of successive objective functions proved to be unstable.

JOSHUA BABIER* AND BEDA CHAN:

We appreciate Dr. Brockett's contribution of systematically treating a broad collection of actuarial problems and of writing an overview on information theory. Although we have written about an information theoretical approach [1], in this discussion we pretend that we do not know of or believe in information theory and redo the problem in Section 9, Determining a Client's Loss Distribution. By providing an alternative solution to the problem in Section 9, we hope to supplement Section 9 to become like Section 12, in which graduation by information theoretic and alternative methods is presented and compared.

If we were to construct an adjusted table with $\mu = 21$ with reference to the Standard Table where $\mu = 31.35$ but not to use an information theoretic approach, what would we do? For simplicity, we first describe a continuous version of our solution. We are given the distribution function

$$F_Y(y),\ 0 \le y \le 91$$

where

$$\int_0^{91} y\, dF_Y(y) + 91\,[1 - F_Y(91)] = 31.35.$$

Denote the adjusted length of claim by Z. We are to find

$$F_Z(z),\ 0 \le z \le 91$$

where

$$\int_0^{91} z\, dF_Z(z) + 91\,[1 - F_Z(91)] = 21.$$

We first concoct an extension of the function $F_Y(y)$ for $y > 91$. We graph $F_Y(y)$ to 91 and then extend our graph by hand to larger values of y. Then, we shrink F_Y: Define

$$F_Z(z) = F_Y(\alpha z)$$

*Mr. Babier, not a member of the Society, is a specialist student in actuarial science at the University of Toronto.

where α is determined by

$$\int_0^{91} z\, dF_Y(\alpha z) + 91\,[1 - F_Y(91\alpha)] = 21.$$

Thus, we need to extend F_Y from [0,91] to [91,91α]. Although our method of extending the graph by hand is rather arbitrary, we only need to use the part of extended F_Y near the known given region of F_Y for the determination of F_Z on [0,91].

We approximate the given Standard Table by a continuous distribution function, extend it, shrink it by finding α numerically, and approximate it back to a discrete distribution. We repeat and fine-tune the process until μ is exactly 21. The repetition and fine-tuning are needed because the exact value of μ is disturbed by the translations between discrete and continuous. For the rest of the discussion we use Z to denote the end product discrete distribution. The Adjusted Table Z we obtained is:

ADJUSTED TABLE

Length z	Prob{$Z = z$}	Length z	Prob{$Z = z$}
1	0.06883	20	0.01437
2	0.06310	21	0.01333
3	0.05787	22	0.01238
4	0.05309	23	0.01150
5	0.04872	24	0.01070
6	0.04473	25	0.01000
7	0.04108	26	0.00929
8	0.03775	27	0.00868
9	0.03471	28	0.00811
10	0.03193	31	0.04437
11	0.02939	38	0.02979
12	0.02706	45	0.02144
13	0.02494	52	0.01644
14	0.02300	59	0.01328
15	0.02122	66	0.01113
16	0.01960	73	0.00957
17	0.01811	80	0.00836
18	0.01675	87	0.00738
19	0.01551	91	0.06249

We compare B, Brockett's adjusted table, and Z by computing the ruin probabilities where B and Z are alternative models for the claim amount. Because disability income insurance benefit amounts are proportional to the

length of claims, our interpretation of claim length as claim amount is a reasonable one. The method and the APL program we used in calculating the values below are from Seah [2]. We used u for initial surplus and θ for loading.

RUIN PROBABILITY

	$\theta = 0.1$	$\theta = 0.3$	$\theta = 0.5$	$\theta = 0.7$	$\theta = 0.9$
$\psi_B(u)$ USING BROCKETT'S ADJUSTED TABLE					
$u = 10$	0.8706	0.6894	0.5691	0.4838	0.4204
$u = 20$	0.8363	0.6228	0.4918	0.4042	0.3420
$u = 30$	0.8045	0.5649	0.4275	0.3404	0.2809
$u = 40$	0.7746	0.5139	0.3735	0.2886	0.2327
$u = 50$	0.7459	0.4675	0.3263	0.2446	0.1927
$u = 60$	0.7178	0.4246	0.2842	0.2064	0.1586
$u = 70$	0.6903	0.3844	0.2461	0.1727	0.1291
$u = 80$	0.6628	0.3462	0.2110	0.1422	0.1028
$\psi_Z(u)$ USING ADJUSTED TABLE Z					
$u = 10$	0.8729	0.6939	0.5746	0.4897	0.4263
$u = 20$	0.8425	0.6349	0.5058	0.4187	0.3563
$u = 30$	0.8148	0.5844	0.4494	0.3623	0.3021
$u = 40$	0.7891	0.5393	0.4009	0.3152	0.2578
$u = 50$	0.7637	0.4969	0.3567	0.2732	0.2190
$u = 60$	0.7383	0.4562	0.3155	0.2348	0.1840
$u = 70$	0.7127	0.4169	0.2766	0.1993	0.1521
$u = 80$	0.6868	0.3785	0.2360	0.1660	0.1226

Our method is conceptually simple. To obtain a Z that is like Y but with a smaller mean, we shrink Y. But since F_Y is given only to 91, we need to extend it over 91 so that we have something to shrink. Our choices in extrapolating F_Y are arbitrary; the fine-tuning that is needed to obtain a discrete table at the end is numerically tedious, especially in comparison with the computational ease of the information theoretic approach as given in Brockett, Charnes and Cooper [15 of the paper].

If information theory is a black box, it is an easy one to use. To come up with a reasonable answer without using information theory requires a great deal more work, and the results are not that different. We note that ψ_Z is always greater than ψ_B in the above tables, but the range of difference is only from 0.26 percent to 19.26 percent.

REFERENCES

1. CHAN, B. "Stability of Premium Principles under Maximum Entropy Perturbations." In *Premium Calculations in Insurance*, edited by F. de Vylder et al. Dordrecht: Reidel, 1984, 381–86.
2. SEAH, E.S. "Computing the Probability of Eventual Ruin," *TSA* XLII (1990): 421–46.

(AUTHOR'S REVIEW OF DISCUSSION)

PATRICK L. BROCKETT:

First, I thank all the discussants for their comments. Second, I address each one's comments in turn.

I appreciate Dr. Carlin's stimulating comments about my paper and his bringing up the Bayesian Information Criterion (BIC). This allows me an additional opportunity to discuss the information theoretic approach versus Bayesian methods and BIC versus AIC.

I did not wish to imply that Bayesian methods are "ad-hoc"; indeed, Bayes' Theorem in its most elementary form is merely a restatement of the fact that joint probability distributions satisfy the logical relationship

$$P(A \text{ and } B) = P(B|A)P(A) = P(A|B)P(B).$$

Dividing by the appropriate marginal distribution yields Bayes' Theorem, and any probability law not satisfying this relationship necessarily fails to satisfy a basic axiom of probability theory.

Although fundamental in this sense, the Bayesian formulation is also special in that the probability distributions are assumed to be known, so that Baye's Theorem merely reflects how one should update the so-called "prior probability" $P(A)$ in the light of new information contained in the event B. Bayes' Theorem does not give any advice about how one should develop the prior probabilities $P(A)$ and $P(B|A)$, especially in the situation frequently encountered in actuarial science in which the "event" B contains vague, incomplete, imprecise, subjective, or inconsistent knowledge about the phenomena under investigation, or where the model itself or the probabilities involved are unknown.

The information theoretic method presented in this paper, however, shows how one *can assign* entire probability *distributions* for events subject to the constraints imposed by prior knowledge. Conversely, the development of *isolated* individual probabilities cannot be addressed within this framework (as it can, for instance, within the Bayesian framework). In addition, the information theoretic method does not deal with how to update the probability structure in

light of new information. For this, either an entirely new probability distribution can be calculated by reapplying the information theoretic variational principle using the new data to augment the constraint set, or Bayes' Theorem might be applied.

Because both the Bayesian approach and, in certain circumstances, the information theoretic approach can be used to update prior probability distributions to obtain posterior distributions and because both techniques are logically consistent, they tend to agree in many specific applications (as noted in my paper). However, the two techniques are designed to address different problems with different goals of analysis and different "given" structures and degrees of subjectivity in the input. My paper was intended to show the wide range of applicability of information theoretic methods as a nonsubjective and unifying principle for actuarial analysis. Information theory has interpretations in terms of information (reduction in uncertainty), estimation facility, stochastic testing facility, parametric model building facility, and unifying ability for many commonly used models such as log-linear, logit, and so on, and hence, it seems to me, to provide a very good nonsubjective philosophical starting point for most aspects of actuarial analysis. This was the purpose of my paper and is why the range of topics covered was so broad.

Having said all that, let me turn to the specifics of the BIC versus the AIC methods comparisons discussed by Carlin. The determination of a criterion for choosing a model selection procedure (such as the asymptotic consistency result cited by Carlin) is not at all clear-cut (and it is for this reason that exogenous philosophical approaches to general problem-solving are needed for guidance). Linhart and Zucchini [3] wrote an entire book on the subject.

With respect to the BIC/AIC controversy, Shibata [4] has shown (as noted by Carlin) that model selection procedures with fixed K (like the AIC) have an asymptotic tendency to overestimate the "true model." However, he has also shown that only the AIC and its variants are asymptotically efficient if the true number of parameters in the model is very large. In fact, Akaike [1] has criticized the relevance of Schwarz's theorem for asymptotically justifying BIC, while Stone [7] has questioned the use of asymptotics as a legitimate selection criterion. It is, in fact, quite possible to criticize the realism of an asymptotic theory which keeps the size of the model fixed as the sample size increases without bound because, as noted by Stone [7], the complexity of the models available in practice would increase as the available data for model refinement increase. Stone [6] has also shown that procedures

like both the BIC and the AIC are asymptotically locally admissible so that no "superior" procedure exists among these methods if the number of variables in the model is held fixed while the sample size increases. Thus, the selection of a criterion depends upon whether you want asymptotic efficiency (possessed by AIC and not BIC) or asymptotic consistency (possessed by BIC and not AIC), and of course all these theorems are based upon asymptotics that, for small samples, may not work so well anyway.

In an empirical investigation Shibata [4] reports that BIC was likely to underestimate the order of the regression model studied. I prefer the AIC because it ties nicely together with all the other topics covered in my paper and is a part of a single philosophical approach to stochastic modeling that goes all the way from selecting models (AIC), to estimating parameters (MDI), to testing hypotheses and setting confidence intervals (loglinear modeling), to incorporating prior knowledge into analysis (MDI), to refining crude observed distributions in light of new or constrained information (graduation), to incorporating and unifying many currently used statistical methods under a single nonsubjective paradigm.

Drs. Rosenbloom and Shiu are correct in their discussion that the condition $l'\mathbf{p} = l'\mathbf{q}$ used in my numerical example is very strong and that the objective function with graduation might be replaced by

$$\sum_{i=-\infty}^{\infty} \ell_i \delta_i \ln(\delta_i/u_i),$$

where ℓ_i could represent, for example, the number of lives exposed to risk at age i. Using a development similar to that in the paper (or that given in Dr. Szatzschneider's discussion), one can derive the explicit parametric form of the solution when this objective function is used subject to the constraint $\ell'\delta = \ell'\mathbf{u}$ (that is, the observed number of deaths equals the expected number of deaths). This was not done in the paper in order to conserve space; however, this analysis is presented below in response to Drs. Rosenbloom's and Shiu's cogent comment.

Because $I(\mathbf{p}|\mathbf{q}) \geq 0$ for all probability measures $\mathbf{p}$ and $\mathbf{q}$, it follows that for arbitrary positive (not necessarily probability) measures, we have

$$\sum_{i=-\infty}^{\infty} p(x_i) \ln[p(x_i)] \geq \sum_{i=-\infty}^{\infty} p(x_i) \ln[q(x_i)] + M_1 \ln[M_1/M_2]$$

where

$$M_1 = \sum_{k=-\infty}^{\infty} p(x_k)$$

denotes the total mass of the first weighted measure and

$$M_2 = \sum_{k=-\infty}^{\infty} q(x_k)$$

denotes the total mass of the second weighted measure. This applies, for example, to the graduation situation suggested by Drs. Rosenbloom and Shiu, in which one uses the weighted mortality rates $\ell_i\delta_i$ and $\ell_i u_i$. When $\mathbf{p}$ and $\mathbf{q}$ are constrained to have the same total mass (for example, if the observed number of deaths is constrained to equal the graduated number of deaths in the weighted mortality table graduation situation $p(x_i) = \ell_i\delta_i$ and $q(x_i) = \ell_i u_i$ suggested by Drs. Rosenbloom and Shiu, or when both $\mathbf{p}$ and $\mathbf{q}$ are probability measures, as in Section 10), the quantity $M_1 \ln[M_1/M_2] = 0$ and the pseudo-distance interpretation of $I(\mathbf{p}|\mathbf{q})$ is retained. It follows that virtually the entire analysis presented in the paper continues after appropriate renormalization. This means, for example, that the extremal solution for the weighted mortality rate graduation problem

$$\text{Minimize} \sum_{i=1}^{n} \ell_i\delta_i \ln\left[\frac{\delta_i}{u_i}\right]$$

subject to

$$\sum_{i=1}^{n} \ell_i\delta_i = \sum_{i=1}^{n} \ell_i u_i = \theta_0$$

$$\sum_{i=1}^{n} a_{ij}\delta_i = \theta_j, \qquad j = 1, 2, \ldots, m,$$

$$\delta_i \geq 0, \qquad i = 1, 2, \ldots, n,$$

is again given by Equation (6.3) with a_{ij} replaced by $a_{ij}M_1/\ell_i$. Note that in this case the resulting representation for δ_i is again in loglinear form, allowing access to the classical statistical analysis of exponential families of distributions. For simplicity of presentation in this paper, we used the same exposure at each age, that is, ℓ_i constant. The above analysis shows this was not really necessary but was done for expositional purposes.

With respect to the question concerning the value of M used in the graduation, first, in general, all graduation techniques require the actuary to make an explicit trade-off between the conflicting goals of fit and smoothness. This is done by selection of parameters in the Whittaker-Henderson methods, and in the information theoretic method the smoothness of the graduation can be varied by varying the value of M. For the analysis presented in the paper, because it was desired to show how an automated statistically based graduation method would compare with ad-hoc graphical methods, we first calculated the value $\Sigma_i (\Delta^3 \delta_i)^2$ implied by the ad-hoc graphical graduation of London. This number then served as a guide as to the magnitude of M necessary for comparing the resultant information theoretic graduation with that of the ad-hoc graphical techniques. In general, a quick ad-hoc graphical comparison such as that done in the paper can be used to determine M, or several different graduations can be done with selected different choices of M and the best graduation selected. We have found that sometimes when the convexity constraints are inserted in addition to monotonicity constraints, the smoothness parameter M is not all that important in the information theoretic graduation.

I thank Dr. Herzog for his comment, with which I concur. The results of *any* analysis are dependent upon the underlying assumptions involved, and the actuary should attempt to be as "unprejudiced" as possible in determining his/her modeling assumptions. In fact, this is actually a further strength of the information theoretic technique.

In addition to the "loss function" or pseudo-distance interpretation of $I(\mathbf{p}|\mathbf{q})$ exploited in the first portion of the paper, the information measure has a "quantification of uncertainty" interpretation as well. It is this interpretation that shows that the minimum discrimination information approach is a generalization of Laplace's famous "principle of insufficient information." The information theoretic method can be viewed as choosing the model that is as "close as possible" to the postulated distribution q subject to the constraints. Essentially the method chooses a model to satisfy the constraints that are known to hold while *maximizing the uncertainty (entropy) of that which we don't assume we know*. Some authors use this argument to justify calling the maximum entropy distribution the most unprejudiced distribution possible subject to constraints. Least squares and other loss functions do not have this interpretation of choosing maximally unprejudiced distributions (again, in the sense of entropy).

While conclusions may change according to the loss function selected, the information theoretic loss function has quite defensible characteristics

supporting its selection, in addition to its being a part of a unified philosophical approach to modeling, estimating, and testing stochastic quantities. Of course I agree with Dr. Herzog that one might try other loss measures, and indeed one could do standard actuarial analysis (for example, Whittaker-Henderson graduation), making sure that this analysis is done subject to the known constraints. Computationally this can be done by using existing nonlinear programming software if desired.

I appreciate the very favorable comments of Dr. Szatzschneider. I agree emphatically with his comment that the constraints of the problem are an integral part of the problem formulation and, in fact, have included the constraints in all my analyses in the paper. It is for this reason that the existence of readily available mathematical programming codes should be of such interest to actuaries. There is no longer any excuse for ignoring important knowledge or constraints solely in order to obtain mathematically elegant closed form solutions (which, if the solution violates the ignored pertinent constraints, results in the analysis becoming irrelevant for practical use anyway). We all use the computer anyway, so why not include the constraints and let the computer find the "best" solution subject to the constraints? This was one of the thrusts of my paper.

I thank Dr. Roach for his comments. The alternative derivation of the extremal solution of the MDI problem presented by Dr. Szatzschneider can be used to comment on the remark by Dr. Roach that there was no optimization involved in the MDI problem. This is wrong, as the derivation in the paper and that given by Dr. Szatzschneider show. Dr. Roach himself seems to have recognized this when he stated that the form of the solution (to the optimization problem) was as given in the paper. The MDI problem is a problem in the calculus of variations, and solving the extremization yields a function as opposed to a number. Of course, once the precise parametric functional form of the solution is known (which requires extremization), the exact parameter values are indeed uniquely specified by the constraints (as noted by Dr. Roach and as noted in my paper). Without extremization, however, we would not know that the parametric form we are looking for is that given by (6.1) or perhaps some other form, and we could proceed no further. There are an infinite number of two-parameter families of distributions whose parameters could be uniquely determined by the two constraints; however, it is the MDI extremization that specifies which is the pertinent family. This becomes even more important as the actuary obtains more information (and hence more constraints). The equation $p_i = q_i(1.57823)(0.9629889)^i$ is not "the assumed form" but rather "the

consequent form" of the adjustment. We are *not* doing parameter optimization here (as Dr. Roach seems to imply in his assertion) but rather *functional form* optimization with the parameters subsequently intrinsically uniquely determined).

I also take issue with the statement that "the form of the problem is artificial and conceals the form of the natural solution." There is nothing "natural" about using mortality ratios to adjust mortality tables in the manner prescribed. Common? Yes. Natural? No. If one examines the information content in a statement like "nonsmokers experience a mortality that is 50 percent of standard mortality," we are left in a quandary because the age interval considered is left out. If the age interval is ages 0 to 200 years, then the statement is ambiguous (because all smokers and nonsmokers are dead by age 200) because there is no "excess mortality" over this interval. If the age interval is not explicitly given, then simply multiplying the standard probability of mortality by the given constant is ad-hoc, does not result in a probability distribution for the adjusted table, and leads to the absurdity that the adjusted probability of death at older ages is greater than 1. The usual technique of truncating such adjusted mortality tables when the probability of death becomes greater than 1 is an ad-hoc adaptation of the method necessary to plug the holes intrinsic in the method. I know of no theoretical justification for simply multiplying the conditional probabilities by a common constant and truncating the table when the probabilities of death exceed one.

An alternative approach is the information theoretic method presented in this paper. If the age interval used in the study that resulted in the empirical statement about the mortality ratios was a to b, then this may be translated into a constraint that the desired adjusted table must satisfy, and the analyses can be performed by methods similar to those of Sections 9, 10, or 11. The adjustment of the mortality tables by using mortality rates rather than time-to-death probabilities can be performed as well. Again the mortality ratio would be inserted as a constraint, and information theoretic selection of the "closest possible" distribution subject to the constraints would be performed.

With respect to the linear programming iteration, I believe the original objective function formulation is sufficiently simple to handle the computations exactly, and I see no particular point in trying to approximate the problem simply so one can use linear programming codes. This is especially true because Dr. Roach finds the approximating solutions either unstable, beyond the numerical capability of his software, or very disappointing. Convex (or nonlinear) programming codes are readily available for the PC, and

when the problem involved equality constraints (like the problems in Section 10), the dual problem is unconstrained so ordinary calculus and successive bisection methods can be used. I did the calculations on a spreadsheet program very quickly by hand. (For some inequality constrained problems, however, I have found the computation facilitated by the reparameterization $p_i = \exp(-X_i)$ with X_i unconstrained.)

The discussion of Babier and Chan presents another way to graduate subject to a mean constraint. If one focuses on the continuous case (as they do at the start of their discussion), then their method reduces to finding an adjusted variable Z which satisfies $F_Z(z) = F_Y(\alpha z)$ for a parameter α selected so that the given mean is achieved. One readily observes that their specified relationship between the two distribution functions is equivalent to the relationship $Z = Y/\alpha$ for the random variables. This would imply that $E[Z] = E[Y]/\alpha$ or that the required value of α is $\alpha = E[Y]/E[Z]$. While the process becomes more complicated when one restricts Z and Y to be integer valued (as done by Babier and Chan), the problem I perceive with their method is the lack of a theoretic rationale for assuming that the observed and graduated variables must be scale multiples of each other. I cannot readily think of any criteria under which this gives the "best" or "natural" estimate of F_Y subject to the mean constraint. Moreover, without such a general rationale for their method, the extension to situations in which more than one constraint is known (as occurs often in actuarial science) becomes problematic. I do find their use of ruin probability estimates as a comparison criteria very interesting and worthy of theoretical investigation. The article by Brockett, Goovaerts and Taylor [2] discusses properties of random variables such that their ruin probabilities are ordered.

REFERENCES

1. AKAIKE, H. "A Bayesian Analysis of the Minimum AIC Procedure," *Annals of the Institute of Statistical Mathematics* 30, Part A (1978): 9–14.
2. BROCKETT, P.L., GOOVAERTS, M., AND TAYLOR, G. "The Schmitter Problem," *ASTIN Bulletin*, 21, no. 1 (1991): 129–32.
3. LINHART, H., AND ZUCCHINI, W. *Model Selection*. New York: John Wiley and Sons 1986.
4. SHIBATA, R. "An Optimal Selection of Regression Variables," *Biometrics* 68 (1981): 45–54.
5. SHIBATA, R. "Regression Variables, Selection of." In *Encyclopedia of Statistical Sciences, Volume 7*, edited by S. Kotz and N.L. Johnson. New York: John Wiley and Sons, 1986, 709–14.

6. STONE, C. "Admissible Selection of an Accurate and Parsimonious Normal Linear Regression Model," *Annals of Statistics* 9 (1981): 475–85.
7. STONE, M. "Comments on Model Selection Criteria of Akaike and Schwarz," *Journal of the Royal Statistical Society, Series B,* 41 (1979): 276–78.