THE CREDIBLE DISTRIBUTION

WILLIAM S. JEWELL
U.S.A.

ABSTRACT

Credibility theory is concerned with the problem of forecasting the mean performance (claim frequency, total losses, etc.) of an individual risk, selected from a collective of heterogeneous risks, based upon the statistics of the collective, and upon the individual's experience data. The classic results, derived by American actuaries in the 1920's, were further strengthened by Bailey and Mayerson in 1950 and 1965, who showed that these results were exact Bayesian for certain risk distributions. Bühlmann, in 1967, then showed that the credibility formulae were the best least-squares linearized approximation to the exact Bayesian forecast, for general risk distributions.

This paper extends credibility theory to the problem of forecasting the *distribution* of individual risk, based upon collective statistics and individual experience data. Although the problem is, in principle, solved by finding a Bayesian conditional distribution, this approach requires a detailed knowledge of collective structure. The *credible distribution*, on the other hand, requires fewer prior statistics, and is also a best least-squares linearized approximation to the exact Bayesian distribution.

Following the theoretical development, detailed computational results are given.

1. INTRODUCTION

Credibility theory is the name given to a method of *experience rating* an insurance risk, which was developed by American actuaries in the 1920's. In the classic problem, one begins with a pool, a *collective* of somewhat heterogeneous insurance contracts which are grouped together to "spread the risk"; it is assumed that detailed prior statistics are available from this pool. In particular, the *fair collective premium*, $E\{\xi\}$, is the average value of the risk random variable of interest, such as number of accidents per year, dollar losses per unit exposure, etc.

Now suppose that a new insurance contract of unknown risk characteristics is underwritten, and assigned to this pool. At the beginning, the *fair individual premium* changed would be just the collective premium $E\{\xi\}$; however, as *individual experience data*

$x_1, x_2, \ldots x_n$ is obtained over n years, this data would tend to reflect more nearly the individual risk characteristics.

Using heuristic reasoning on the pooling of data (and considering only the number of claims per year), arguments were advanced in the early literature for a *fair experience premium* for next year's risk, ξ_{n+1}, based on a formula of the form

$$\mathrm{E}\{\xi_{n+1} \mid x_1, x_2, \ldots x_n\} \approx (1 - Z) \cdot \mathrm{E}\{\xi\} + Z \cdot (\sum_{t=1}^{n} x_t/n), \quad (1.1)$$

with

$$Z = \frac{n}{n + N}. \quad (1.2)$$

Z was called the *credibility factor*; it provides for a mixing of the fair collective premium, $\mathrm{E}\{\xi\}$, and the *individual sample mean*, $\Sigma\, x_t/n$, with increasing "credibility" attached to the latter as n increases. The time constant N was essentially determined by trial and error.

This credibility formula was successfully used in American actuarial circles for more than 50 years, with innumerable variation and elaboration. A full survey, with references, may be found in Longley-Cook [11]. However, the modern theory of credibility begins with the resurgence of Bayesian techniques and with the works of Bailey [2] and Mayerson [12], who showed, under certain assumptions regarding the structure of the collective of risks, that (1.1) was an exact formula. Finally, in 1967, Bühlmann [3] showed that the credibility formula was the best least-squares linearized approximation to the exact Bayesian forecast, and gave an explicit formula for N in terms of collective second-order statistics (see formula (4.6) below). A larger survey of this development is in [19].

Since that time, other research has focused on credibility-type forecasts of variance [4], the use of auxiliary data in conditional distributions [5], [6], the "IBNR triangle" of partial data development [17], and multi-dimensional risks [19], [20].

The purpose of this paper is to extend the approach of credibility theory to the problem of estimating the *distribution of individual risk*, based upon collective statistics and individual experience data. Although this problem is, in principle, solved by finding the Bayesian conditional distribution, $Pr\{\xi_{n+1} \leq y \mid x_1, x_2, \ldots x_n\}$,

this approach requires a detailed knowledge of collective structure for every y. We shall see that the credibility approach needs fewer prior statistics for a fixed value of y, and leads to a simplified prediction of credibility type (1.1); furthermore, the *credible distribution* is unbiased, and is a least-squares linearized fit to the exact Bayesian distribution.

First, we consider in more detail the nature of the risk collective, and results of least-squares theory we shall need in the sequel. Following the development of the credible distribution, we consider the problem of forecasting the density, and show how, in the discrete case, more complicated estimates can be made. Various theoretical properties of the credible estimates are presented, followed by computational results for certain well-known distributions. Finally, we briefly consider certain problems in moment estimation.

2. The Risk Collective: Bayesian Results

Consider a *collective of heterogeneous risks*, such as an insurance portfolio, in which each member is characterized by a *risk parameter*, θ. For a given value of θ, the *claims experience* (number of claims or total value of claims) for a certain time period or exposure base t is a random variable, ξ_t, with known distribution

$$P_t(x \mid \theta) = Pr\{\xi_t \leq x \mid \theta\} \qquad (t = 1, 2, \ldots). \tag{2.1}$$

In what follows, it will be assumed that the ξ_t are mutually independent, given θ; the (discrete or continuous) density of (2.1) will be indicated by $p_t(x \mid \theta)$.

If the true value of an individual parameter $\theta = \theta_T$ were known, then the *fair premium* would be:

$$E\{\xi_t \mid \theta_T\} = \int x dP_t(x \mid \theta_T) \tag{2.2}$$

for any time period t.

If θ_T were not known, it would still be possible to infer a fair premium for an individual risk, provided that *a prior distribution*, $U(\cdot)$, on the collective risk parameter was known, and if *experience data* $(\xi_t = x_t \mid t = 1, 2, \ldots, n)$ were available for this individual. By the usual Bayesian argument, the *forecast distribution* of next year's risk would be the conditional distribution:

$$P_{n+1}(y \mid x_1, x_2, \ldots, x_n) = Pr\{\xi_{n+1} \leq y \mid x_1, x_2, \ldots, x_n\}$$

$$= \frac{\int P_{n+1}(y \mid \theta) \prod_{t=1}^{n} p_t(x_t \mid \theta)\, dU(\theta)}{\int \prod_{t=1}^{n} p_t(x_t \mid \theta)\, dU(\theta)}, \qquad (2.3)$$

which is known to be the unbiased, least-squares estimate of $P_{n+1}(y \mid \theta_T)$, given the experience data $\underline{x} = \{x_1, x_2, \ldots x_n\}$.

The *fair experience premium* would then be:

$$\mathrm{E}\{\xi_{n+1} \mid \underline{x}\} = \int y dP_{n+1}(y \mid \underline{x}). \qquad (2.4)$$

The statistical literature emphasizes the behavior of the *density of* θ *posterior to* $\underline{x}$, given by:

$$dU_n(\theta \mid \underline{x}) = \frac{\prod_{t=1}^{n} p_t(x_t \mid \theta)\, dU(\theta)}{\int \prod_{t=1}^{n} p_t(x_t \mid \phi)\, dU(\phi)} \qquad (2.5)$$

It is known (see, for example DeGroot [7]), that for fairly arbitrary priors, $U_n(\theta \mid \underline{x})$ is approximately normal with mean θ_T, variance proportional to n^{-1}, for large enough n, and converges to the degenerate distribution at $\theta = \theta_T$ as $n \to \infty$. Thus the forecast distribution (2.3) also converges to $P(y \mid \theta_T)$ for almost every y.

These Bayesian calculations are laborious, except for certain well-known *conjugate-prior families of priors and likelihoods*, such as Beta-Binomial, Gamma-Poisson, Normal-Normal, etc. (See for example [7]). Furthermore, the problem in insurance and other applications is that, although detailed statistics may be available from the mixed *collective distribution*:

$$P_t(x) = \mathrm{E}_\theta P_t(x \mid \theta) = \int P_t(x \mid \theta)\, dU(\theta), \qquad (2.6)$$

there is very little information available on the internal structure of the collective, between different risks. Thus a full Bayesian forecasting is impossible without additional distributional assumptions.

In the sequel, we shall deal only with time-invariant collectives, for which $P_t(x \mid \theta) = P(x \mid \theta)$ $(\forall\, t)$, and we shall consider the problem of providing a credibility-type approximation to (2.3). The theoretical basis for this approximation is in least-squares theory.

3. Least-Squares Theory

Suppose we have a vector-valued random variable $\underline{\omega}$ from whose observations $\underline{w}$ we are trying to predict another random variable η through a forecast function $f(\underline{w})$. Assuming the joint distribution $P(y, \underline{w}) = Pr\{\eta \leq y; \underline{\omega} \leq \underline{w}\}$ is known, the classical norm to evaluate the forecast is the mean-square error:

$$I = \int (y - f(\underline{w}))^2 \, dP(y, \underline{w}). \tag{3.1}$$

It is known that the integrable function f^0 which minimizes (3.1) at value I^0 is the conditional mean:

$$f^0(\underline{w}) = \mathrm{E}\{\eta \mid \underline{\omega} = \underline{w}\}, \tag{3.2}$$

where E is defined with respect to the measure P. However, in many cases the exact conditional calculation is too difficult and an approximate forecast function f is sought. Since completion of the square shows that

$$\begin{aligned} I &= I^0 + \int (f^0(\underline{w}) - f(\underline{w}))^2 \, dP(\underline{w}) \\ I^0 &= \mathrm{V}\{\eta\} - \mathrm{V}f^0(\underline{\omega}) = \mathrm{E}_{\underline{\omega}}\mathrm{V}\{\eta \mid \underline{\omega}\} \end{aligned} \tag{3.3}$$

for any f, then the approximate forecast is also a least-squares fit to the conditional mean, and one may select arbitrary parameters in the approximation to make the integral in (3.3) as small as possible, or work directly with (3.1).

A typical choice of an approximate forecast is a *linear function*

$$f(\underline{w}) = a_0 + \Sigma\, a_j w_j. \tag{3.4}$$

In this case it is well known that the optimal parameters a_j^* are given by solutions of linear equations of the form:

$$\sum_{j \neq 0} \mathrm{C}\{\omega_i, \omega_j\} \cdot a_j^* = \mathrm{C}\{\eta; \omega_i\} \qquad (\forall\, i \neq 0) \tag{3.5}$$

with a_0^* selected so as to make the average forecast $\mathrm{E}\{f(\underline{\omega})\}$ unbiased, e.g.,

$$\mathrm{E}\{f(\underline{\omega})\} = \mathrm{E}\{\eta\}; \; a_0^* = \mathrm{E}\{\eta\} - \sum_{j \neq 0} a_j^* \mathrm{E}\{\omega_j\}. \tag{3.6}$$

The prior variance of the optimal linear forecast is:

$$\mathrm{V}\{f(\underline{\omega})\} = \sum_{i,j \neq 0} \sum a_i^* a_j^* \mathrm{C}\{\omega_i; \omega_j\} = \sum_{j \neq 0} a_j^* \mathrm{C}\{\eta; \omega_j\} \tag{3.7}$$

and the approximation error turns out to be:

$$I - I^0 = Vf^0(\underline{\omega}) - \sum_{j \neq 0} a_j^* C\{\eta; \omega_j\}. \tag{3.8}$$

As before, all operators are defined on the measure $P = P(y, w)$.

It is sometimes not realized that the above approximation is a linearization only in terms of the parameters a_j, and not necessarily in terms of the observables. For, suppose there is an underlying vector random variable $\underline{\xi}$, with observations $\underline{x}$, and there are known transformations

$$\eta = g_0(\underline{\xi}); \quad \omega_j = g_j(\underline{\xi}) \quad (j = 1, 2, \ldots). \tag{3.9}$$

Then the above theory applies directly to the prediction of $g_0(\underline{\xi})$ by a forecast function

$$f(x) = a_0 + \sum_{j \neq 0} a_j g_j(\underline{x}), \tag{3.10}$$

by making the obvious extension of the operators in (3.5) and (3.6). In many cases a further simplification results if the $g_j(\cdot)$ are functions of only a single component of $\underline{\xi}$.

Another modification of linear least-squares theory occurs when one *constrains* the variables:

$$c_0 a_0 + \sum_{j \neq 0} c_j a_j = \text{Constant}. \tag{3.11}$$

Here one defines a Langrange multiplier μ, adds $\mu \cdot c_0$ to the definition (3.6) of a_0^*, and adds $\mu(c_i - c_0 E\{\omega_i\})$ to the ith equation of (3.5); μ is then adjusted until (3.11) holds.

A special case of the above occurs when a subset of the a_j $(j \neq 0)$ are constrained to be equal to each other. One can show that the columns in the constraint matrix $[C\{\xi_i, \xi_j\}]$ corresponding to the common a_j are first added together to form the coefficients of the common variable a_c; then the coefficients of the rows corresponding to the constrained subset (in both the constraint matrix and the RHS, $[C\{\eta; \xi_i\}]$) of variables are aggregated by addition into a single equation, thus making the system (3.5) again square. If there were m equations, and $1 < r \leq m$ variables are set equal to each other, the resulting system is $(m - r + 1) \times (m - r + 1)$, and the coefficient of a_c in its own row consists of the sum of r^2 old coefficients.

All of these constraints increase the mean-square error.

Finally, we make the observation that *superposition* holds, i.e. suppose we have made a linear forecast of a certain random variable η^1 by finding parameters $\{a_j^1\}$ from (3.5) using a RHS of $C\{\eta^1; \omega_i\}$. We then repeat the process, finding other sets of parameters $\{a_j^k \mid k = 2, 3, \ldots\}$ using a RHS of $C\{\eta^k; \omega_i\}$ $(k = 2, 3, \ldots)$. Not only is only one inversion of the constraint matrix of (3.5) required for all the parameter sets, but any linear combination of predictands, say of:

$$\eta' = \sum_k c_k \eta^k \tag{3.12}$$

will have optimal values

$$a_j' = \sum_k c_k a_j^k \qquad (j = 0, 1, 2, \ldots). \tag{3.13}$$

We now apply these results to the model of the collective.

4. The Credible Mean

In the collective model, there are underlying random variables $\xi_1, \xi_2, \ldots, \xi_n; \xi_{n+1}$ which are mutually independent (and, here, identically distributed), given 0, the risk parameter. To predict the *mean of the next observation*, given the n observed values $\xi_t = x_t$ $(t = 1, \ldots, n)$, we take the simplest case of (3.9):

$$\eta = \xi_{n+1}; \quad \omega_t = \xi_t \ (t = 1, \ldots, n). \tag{4.1}$$

Using the fact that

$$C\{\xi_i; \xi_j \mid \theta\} = \begin{cases} V\{\xi_i \mid \theta\} & (i = j) \\ 0 & (i \neq j), \end{cases} \tag{4.2}$$

we find

$$C\{\omega_i; \omega_j\} = \begin{cases} V\{\xi_i\} = E_0 V\{\xi_i \mid \theta\} + V_0 E\{\xi_i \mid \theta\} & (i = j) \\ C_0\{E\{\xi_i \mid \theta\}; E\{\xi_j \mid \theta\}\}. & (i \neq j) \end{cases} \tag{4.3}$$

The second case also holds for $C\{\eta; \omega_i\}$.

If the ξ_t are identically distributed $(t = 1, 2, \ldots n + 1)$, we find that the optimal a_j $(j \neq 0)$ are identical, with:

$$a_0 = E\{\xi\}[1 - na_1] \tag{4.4}$$

$$a_1 = \frac{1}{n + N} \qquad N = \frac{V\{\xi\}}{V_0 E\{\xi \mid \theta\}} - 1 = \frac{E_0 V\{\xi \mid \theta\}}{V_0 E\{\xi \mid \theta\}}. \tag{4.5}$$

In other words, we get the classical credibility forecast:

$$\mathrm{E}\{\xi_{n+1} \mid \underline{x} = \underline{\xi}\} \approx f(\underline{x}) = (1 - Z)\ \mathrm{E}\{\xi\} + Z\left(\frac{\sum_{t=1}^{n} x_t}{n}\right). \tag{4.6}$$

This result is due to Bühlmann [3].

Where necessary in the sequel, we shall distinguish the above—given N and Z from others as

$$N_{[1]} = \frac{\mathrm{E}_0\mathrm{V}\{\xi \mid \theta\}}{\mathrm{V}_0\mathrm{E}\{\xi \mid \theta\}};$$

$$Z_{[1]} = \frac{n}{n + N_{[1]}}, \tag{4.7}$$

the *mean-credible time constant* and *credibility factor*, respectively. The corresponding forecast function in (4.6) will be referred to as $f_{[1]}(\underline{x})$. The forecast of the mean is *a priori* unbiased,

$$\mathrm{E}f_{[1]}(\underline{\xi}) = \mathrm{E}\{\xi\}. \tag{4.8}$$

It is easy to show that the mean square error is:

$$I = \mathrm{V}\{\xi\} - Z_{[1]}\ \mathrm{V}_0\mathrm{E}\{\xi \mid \theta\}, \tag{4.9}$$

so that error starts at $\mathrm{V}\{\xi\}$ (variance using the collective mean as forecast), and decreases with increasing n to $\mathrm{E}_0\mathrm{V}\{\xi \mid \theta\}$ (irreducible variance in sample mean). This error is, of course, *a priori*; Section 8 examines the forecast behavior when θ is known.

5. The Credible Distribution

We now consider the central problem of this paper, which is to find a credibility approximation to the true distribution of the next observation:

$$P_{n+1}(y \mid \theta_T) = Pr\{\xi_{n+1} \leq y \mid \theta_T\}. \tag{5.1}$$

The analysis is greatly facilitated if we use the generalized least-squares formulae (3.9) (3.10) and set:

$$\eta = g_0(\xi_{n+1}) = I(y - \xi_{n+1}) \tag{5.2}$$

for a fixed value of y, where $I(\cdot)$ is the unit step, unity for non-negative arguments, zero otherwise.

The optimal predictor of η, in terms of $\underline{\xi} = \underline{x}$ (now referring to the first n samples only), is by (3.2):

$$f^0(\underline{x}) = \mathrm{E}\{\eta \mid \underline{\xi} = \underline{x}\} = \mathrm{E}\{I(y - \xi_{n+1} \mid \underline{\xi} = \underline{x}\} = P\{y \mid \xi = x\}, \tag{5.3}$$

the Bayesian conditional distribution! Thus, a credibility forecast of type (4.5) will approximate the Bayes distribution, if suitable transformations (3.9) can be chosen. (5.2) above suggests we also choose

$$\omega_t = I(y - \xi_t) \qquad (t = 1, \ldots, n) \tag{5.4}$$

Using the independence and identical distribution properties of the collective described previously, we find the (prior) moments:

$$\mathrm{E}\{\eta\} = \mathrm{E}\{\omega_t\} = P(y); \tag{5.5}$$

$$\mathrm{Cov}\{\omega_i; \omega_j\} = \begin{cases} P(y)\,(1 - P(y)) & (i = j) \\ \mathrm{V}_0 P(y \mid \theta). & (i \neq j) \end{cases} \tag{5.6}$$

(The last case also covers $\mathrm{Cov}\{\eta; \omega_i\}$). These results should be compared with (4.3). It follows, as for the mean, that the optimal coefficients $a_j (i \neq 0)$ are identical, and we obtain the *credible forecast distribution*:

$$P\{y \mid \underline{\xi} = \underline{x}\} \approx f(\underline{x}) = (1 - Z)\,P(y) + Z \left[\frac{\sum_{t=1}^{n} I(y - x_t)}{n} \right] \tag{5.7}$$

with

$$Z = \frac{n}{n + N}; \quad N = \frac{P(y)\,(1 - P(y))}{\mathrm{V}_0 P(y \mid \theta)} - 1. \tag{5.8}$$

Notice how the classical form remains the same; the forecast is a mixture of the collective estimate of the distribution, $P(y)$, and of the *sample distribution*, $\Sigma\, I(y - x_t)/n$. The credibility factor is an increasing function of n, of classical form, but with a different time-constant, N, which in this case depends upon the chosen value of y. And, $\lim_{n \to \infty} Z = 1$.

To distinguish the above results from other credibility formulae, and to emphasize the role of y, we shall henceforth refer to the

above time-constant as $N_P(y)$, the credibility factor as $Z_P(y)$, and the forecast function as $F(y \mid x)$.

A priori, the mean forecast is unbiased:

$$\mathrm{E}\{F(y \mid \underset{\sim}{\xi})\} = P(y), \tag{5.9}$$

and the mean-square error is

$$I = P(y)\,(1 - P(y)) - \mathrm{V}_0 P(y \mid \theta)\ .\ Z_P(y). \tag{5.10}$$

Incidentally, it is easy to see that the credible estimate of the complementary distribution, $P^c\{y \mid \underset{\sim}{x}\} = Pr\{\xi_{n+1} > y \mid \underset{\sim}{\xi} = \underset{\sim}{x}\}$ is the same as (5.7), with the same credibility factor, but with $P(y)$ replaced by $P^c(y)$, and the complementary sample distribution used.

6. Historical Remarks; an Exact Result

The form of the credible distribution has already been hinted at in other works. Whitney [18] in 1918 begins with a normal distribution of "class hazard" and, using a mixture of arguments reminiscent of later Bayesian and maximum likelihood techniques, finds a credibility form to mix "the indicated (individual) risk hazard" with P, "the indicated class hazard". He obtains Z of form (5.8), with, as one approximation,

$$N = \frac{P(1 - P)}{\varepsilon^2}, \tag{6.1}$$

ε^2 being the (normal) "variation of hazard within the class". "We now come to the most difficult question of all, the determination of ε^2. It is obviously impossible to determine ε^2 statistically in each case. Some general assumptions must be made regarding its form and value". [18] Whitney goes on to argue for ε^2 varying as $P^{5/4}$, while others argued for N a constant. (See the discussion to [12], p. 123-4).

The formula (4.5) for the credible mean of the Beta-Bernoulli family, as derived by Bailey [2] and Mayerson [12], is also suggestive:

$$N = \frac{m(1 - m)}{\sigma^2} - 1 \tag{6.2}$$

where m and σ^2 are the mean and variance of "$P(H)$, the prior probability (one is) willing to assign to H", the hypothesis. [12]

In fact, after reflection we see that estimating probabilities is a form of Bernoulli trial, in which we count each sample as a "success" or "failure", depending on which side of y it falls; the long-run frequency of success (which is the mean success) must be the probability sought. Our contribution is thus to point out that (5.7) and (5.8) are the minimal-variance estimators for an *arbitrary* distribution of "class hazard".

Perhaps it is not surprising that the only distributions of $P(x \mid \theta)$ and $U(\theta)$ which the author has been able to find for which (5.7) is exact Bayesian are the Bernoulli $(x \mid \theta)$ — Beta$(\theta \mid \alpha, \beta)$ families, for which:

$$N_P(y) = \frac{\mathrm{E}\{\theta\}\,[1 - \mathrm{E}\{\theta\}]}{\mathrm{V}\{\theta\}} - 1 = \alpha + \beta. \qquad (y = 0,1) \qquad (6.3)$$

Credibility is already only an approximation for slightly enlarged families, such as Binomial-Beta, or Bernoulli-Arbitrary $U(\cdot)$.

7. Computational Considerations

What has been accomplished with (5.7), as compared to the minimal variance Bayesian prediction (2.3)? In the first place, the exact calculation requires knowledge of the structure of the prior and likelihood for all values of the observables, for all θ. Practically speaking, this restricts the computations to the conjugate-prior families of distributions.

The credible forecast (5.7), on the other hand, is a point estimate of $P(y \mid \underline{x})$, which is practically distribution-free, requiring only estimates of $P(y) = \mathrm{E}_\theta P(y \mid \theta)$ and $\mathrm{V}_\theta P(y \mid \theta)$ from the collective at the desired value(s) of y. Even the experience record-keeping is simplified; the sample distribution $\Sigma\, I(y - x_t)\,/n$ only counts the number of samples $\leq y$, and not their exact values.

On the other hand, the credibility approach is somewhat awkward for estimating probabilities for many different values of y, unless there is a simple model for the variation of $P(y \mid \theta)$ over θ. The mean-square error will be larger than obtainable from Bayesian techniques, although the limited computational results in Section 9 seem to indicate that most of the variance is due to the samples, rather than the approximation.

8. Behavior of the Forecast for Known θ

Additional insight into the nature of credible forecasts can be gotten by examining the behavior of $f(\underline{\omega})$, assuming that the true value of $\theta = \theta_T$ is known. From (3.4), the prior unbiasedness of the forecast gives

$$\mathrm{E}\{f(\underline{\omega}) \mid \theta_T\} = \mathrm{E}\{\eta\} + \sum_{j \neq 0} a_j[\mathrm{E}\{\omega_j \mid \theta_T\} - \mathrm{E}\{\omega_j\}]. \tag{8.1}$$

For the credible mean, the results of Section 4 give:

$$\mathrm{E}\{f_{[1]}(\underline{\xi}) \mid \theta_T\} = (1 - Z_{[1]})\,\mathrm{E}\{\xi\} + Z_{[1]}\mathrm{E}\{\xi \mid \theta_T\} \quad (n = 0, 1, 2, \ldots) \tag{8.2}$$

which is itself a "credibility" curve, moving the average estimate from the collective mean to true mean as $n \to \infty$, with time constant $N_{[1]}$.

A similar result and interpretation applies to the credible distribution:

$$\mathrm{E}\{F(y \mid \underline{\xi}) \mid \theta_T\} = [1 - Z_P(y)]\,P(y) + Z_P(y)\,P(y \mid \theta_T), \quad (n = 0, 1, 2, \ldots) \tag{8.3}$$

and with obvious modification, to the credible discrete density (10.6).

The variance of the linear estimator (3.4), given θ_T, is generally:

$$\mathrm{V}\{f(\omega) \mid \theta_T\} = \sum_{i,\, j \neq 0} \sum a_i a_j \mathrm{C}\{\omega_i; \omega_j \mid \theta_T\}; \tag{8.4}$$

however, in the collective models, the transformed random variables (3.9) are independent, given θ_T. For the credible mean:

$$\mathrm{V}\{f_{[1]}(\underline{\xi}) \mid \theta_T\} = \mathrm{V}\{\xi \mid \theta_T\} \cdot \frac{(Z_{[1]})^2}{n}, \quad (n = 0, 1, 2.) \tag{8.5}$$

and for the credible distribution:

$$\mathrm{V}\{F(y \mid \underline{\xi}) \mid \theta_T\} = P(y \mid \theta_T)\,(1 - P(y \mid \theta_T)) \cdot \frac{(Z_P(y))^2}{n}, \quad (n = 0, 1, 2.) \tag{8.6}$$

with, of course, zero variance for $n = 0$.

The function

$$\frac{Z^2}{n} = \frac{(n + N)^2}{n} = \frac{1}{N} \frac{(n/N)}{(1 + (n/N))^2}$$

increases rapidly from zero to its maximum value $1/4N$ when $n = N$, thereafter decreasing slowly to zero as $1/n$.

Sometimes direct use of the sample mean as a forecast is advocated,

$$f_{SM}(\xi) = \sum_{t=1}^{n} \xi_t / n \tag{8.8}$$

since it is "fully credible" for all n, that is:

$$\mathrm{E}\{f_{SM}(\xi) \mid \theta_T\} = \mathrm{E}\{\xi \mid \theta_T\} \cdot (\forall\, n) \tag{8.9}$$

However, comparison of the efficiency of (8.8) with (8.5) shows that:

$$\frac{\mathrm{V}\{f_{[1]}(\xi) \mid \theta_T\}}{\mathrm{V}\{f_{SM}(\xi) \mid \theta_T\}} = (Z_{[1]})^2 < 1. \tag{8.10}$$

In other words, the same credibility form which limits the rate of change of the estimator also shows its variance-reduction properties. (8.10) also holds for the credible distribution estimate vis-à-vis the sample distribution.

If the same random variables are used to forecast the distribution at more than one value of y, there is, of course, covariance between the two estimators. Thus.

$$\mathrm{C}\{F(y_1 \mid \xi); F(y_2 \mid \xi) \mid \theta_T\} =$$

$$[P(\min(y_1, y_2) \mid \theta_T) - P(y_1 \mid \theta_T) \cdot P(y_2 \mid \theta_T)] \frac{n}{(n+N_P(y_1))\,(n+N_P(y_2))}. \tag{8.11}$$

Examples of this interrelationship will be seen in the next section.

Finally, it is obvious there is strong dependence between the forecasts made in successive years, since:

$$F(y \mid x_1, x_2, \ldots, x_t; \xi_{t+1}) = \frac{N_P(y) + t}{N_P(y) + t + 1} F(y \mid x_1, x_2, \ldots x_t) + \frac{\xi_{t+1}}{N_P(y) + t + 1} \cdot (t = 0, 1, 2, \ldots) \tag{8.12}$$

It follows that

$$\mathrm{E}\{F(y \mid x_1 \ldots x_t, \xi_{t+1}) \mid F(y \mid x_1, x_2, \ldots, x_t); \theta_T\} = \frac{(N_P(y) + t)\, F(y \mid x_1, x_2, \ldots, x_t) + P(y \mid \theta_T)}{N_P(y) + t + 1} \tag{8.13}$$

and

$$V\{F(y \mid x_1 \ldots x_t;\ \xi_{t+1}) \mid F(y \mid x_1, x_2, \ldots, x_t);\ \theta_T\} = \frac{P(y \mid \theta_T)\,[1 - P(y \mid \theta_T)]}{(N_P(y) + t + 1)^2}. \tag{8.14}$$

A priori, the covariance between successive forecasts slowly diminishes in a manner similar to (8.6)

$$C\{F(y \mid \xi_1, \ldots, \xi_T);\ F(y \mid \xi_1, \ldots, \xi_t, \xi_{t+1}) \mid \theta_T\} = P(y \mid \theta_T)\,[1 - P(y \mid \theta_T)] \cdot \frac{t}{[t + N_P(y)]\,[t + N_P(y) + 1]}. \tag{8.15}$$

9. Computational Results-Credible Distribution

Detailed computations were carried out for three conjugate prior families of distributions for which explicit results are available:

I. Poisson + Gamma = Negative Binomial

$$p(x \mid \theta) = \frac{\theta^x e^{-\theta}}{x!} \quad (x = 0, 1, 2, \ldots)$$

$$E\{\xi \mid \theta\} = \theta$$

$$V\{\xi \mid \theta\} = \theta$$

$$u(\theta) = \frac{b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)} \quad (\theta \geq 0)$$

$$E\{\theta\} = a/b$$

$$V\{\theta\} = a/b^2 \tag{9.1}$$

$$p(x) = \frac{\Gamma(a + x)}{\Gamma(a)\,x!} \left(\frac{b}{b + 1}\right)^a \left(\frac{1}{b + 1}\right)^x \quad (x = 0, 1, 2, \ldots)$$

$$E\{\xi\} = a/b$$

$$V\{\xi\} = \frac{a}{b}\left(1 + \frac{1}{b}\right)$$

Bayesian Conditional Distributions:

$$a \leftarrow a + \sum_{i=1}^{n} x_i; \quad b \leftarrow b + n$$

II. Exponential + Gamma = Shifted Pareto

$$p(x \mid \theta) = \theta e^{-\theta x} \qquad (x \geq 0)$$

$$\mathrm{E}\{\xi \mid \theta\} = \theta^{-1}$$

$$\mathrm{V}\{\xi \mid \theta\} = \theta^{-2}$$

$$u(\theta) = \frac{b^a \theta^{a-1} e^{-b\theta}}{\Gamma(a)} \qquad (\theta \geq 0)$$

$$\mathrm{E}\{\theta^{-1}\} = \frac{b}{a - 1}$$

$$\mathrm{V}\{\theta^{-1}\} = \frac{b^2}{(a-1)^2\,(a-2)}$$

$$p(x) = \frac{ab^a}{(b+x)^{a+1}} \qquad (x \geq 0)$$

$$\mathrm{E}\{\xi\} = \frac{b}{a - 1}$$

$$\mathrm{V}\{\xi\} = \frac{ab^2}{(a-1)^2\,(a-2)} \qquad (9.2)$$

Bayesian Conditional Distributions:

$$a \leftarrow a + n; \quad b \leftarrow b + \sum_{i=1}^{n} x_i$$

III. Uniform + Pareto = (Constant-Pareto)

$$p(x \mid \theta) = \frac{1}{\theta} \qquad (0 \leq x \leq \theta)$$

$$\mathrm{E}\{\xi \mid \theta\} = \frac{\theta}{2}$$

$$\mathrm{V}\{\xi \mid \theta\} = \frac{\theta^2}{12}$$

$$u(\theta) = \begin{cases} 0 & (0 \leq \theta < b) \\ \dfrac{ab^a}{\theta^{a+1}} & (\theta \geq b) \end{cases}$$

$$\mathrm{E}\{\theta\} = \frac{ab}{a - 1}$$

$$V\{\theta\} = \frac{ab^2}{(a-1)^2(a-2)}$$

$$p(x) = \left(\frac{a}{a+1}\, b^a\right) \cdot \begin{cases} b^{-(a+1)} & (0 \le x \le b) \\ x^{-(a+1)} & (x > b) \end{cases}$$

$$E\{\xi\} = \tfrac{1}{2}\,\frac{ab}{a-1}$$

$$V\{\xi\} = \frac{ab^2}{(a-2)}\left[\frac{1}{12} + \frac{1}{4(a-1)^2}\right] \tag{9.3}$$

Bayesian Conditional Distributions

$$a \leftarrow a + n; \quad b \leftarrow \max\ (b; x_1, x_2, \ldots, x_n).$$

The time constants for the credible means are:

$$N_{[1]} = \begin{cases} b & \text{(Poisson-Gamma)} \\ a-1 & \text{(Exponential-Gamma)} \\ \tfrac{1}{3}(a-1)^2 & \text{(Uniform-Pareto).} \end{cases} \tag{9.4}$$

The credible mean is exact Bayesian for the first two families: the correct Bayesian conditional mean for the Uniform-Pareto is:

$$E\{\xi_{n+1} \mid \underline{x}\} = \tfrac{1}{2}\,\frac{a+n}{a+n-1}\ \max\ (b; x_1, x_2, \ldots, x_n). \tag{9.5}$$

Figures 1, 2, and 3 show the time constant $N_P(y)$ for the above three cases, with the hyperparameters (a, b) adjusted so that $E\{\xi\} = 1$ always, and $V\{\xi\} = 2, 4, 8$. This would result in mean time constants, for example, of:

$$N_{[1]} = \left\{ \begin{array}{llll} 1, & 1/3, & 1/7; & \text{(Poisson-Gamma)} \\ 3, & 5/3, & 9/7; & \text{(Exponential-Gamma)} \\ 0.600, & 0.455, & 0.391. & \text{(Uniform-Pareto)} \end{array} \right\}$$

$$(V(\xi) = 2, 4, 8).$$

Thus, in all these cases, $N_{[1]} < N_P(y)$ for all y.

In Figure 1, we see that the $N_P(y)$ for the Poisson-Gamma have their largest values and most marked variation over v for small collective variance. This is consistent with the idea that when the inter-risk variance is small ($N_P(y)$ large), the occurrence of a large

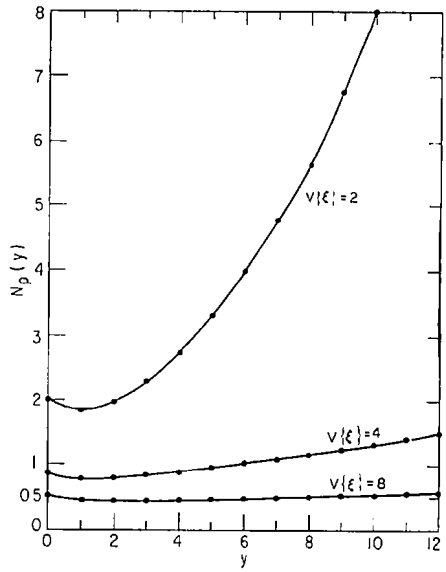

Fig. 1. Credible distribution time constant, $N_P(y)$, for different collective variances. Poisson-Gamma distributions. $E\{\xi\} = 1$ (Straight lines for clarity only).

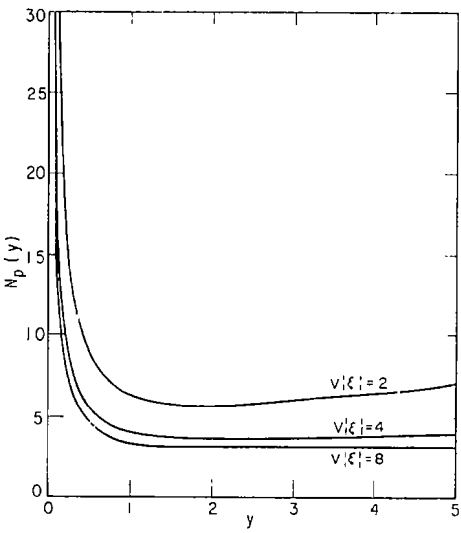

Fig. 2. Credible distribution time constant, $N_P(y)$, for different collective variances. Exponential-Gamma distributions. $E\{\xi\} = 1$.

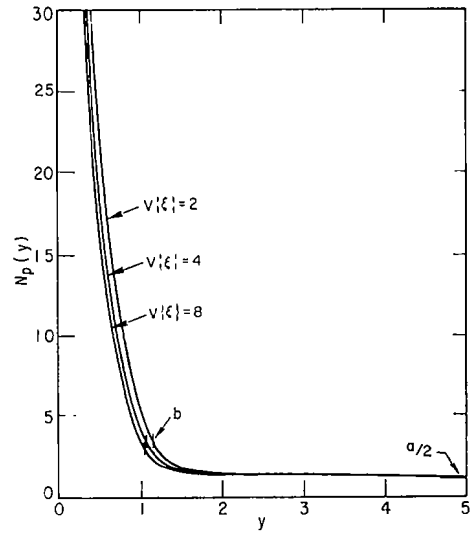

Fig. 3. Credible distribution time constant, $N_P(y)$, for different collective variances. Uniform-Pareto distributions. $E\{\xi\} = 1$.

sample is not weighted heavily; it is likely due to chance. And since $P(y \mid \theta_T)$ is likely "close to" $P(y)$, it takes many more samples to accredit the sample distributions. Conversely, for a very heterogeneous collective, samples for any value of y are treated more evenly and with more credibility.

Figure 2, for the Exponential-Gamma shows much the same behavior as the previous case, except the time constants are larger, in general. Information close to the origin is practically disregarded, as all risks have a preponderance of samples there, due to the exponential form. Tail values are only slightly deemphasized, relative to middle values. The Uniform-Pareto curves, Figure 3, are practically indistinguishable from one another, and are constantly decreasing towards the asymptote $N_P(y) = a/2 = (1.171, 1.084, 1.042)$. Since all risks in the collective differ only by their range $[0, \theta]$, it follows that information below the minimum range $\theta = b$ $(= 1.146, 1.077, 1.04)$ is pretty much disregarded. Thus a good approximation to the Uniform-Pareto is $N_P(y) = a/2$ $(y > b)$, ∞ otherwise.

A variety of simulations were run using these distributions, and

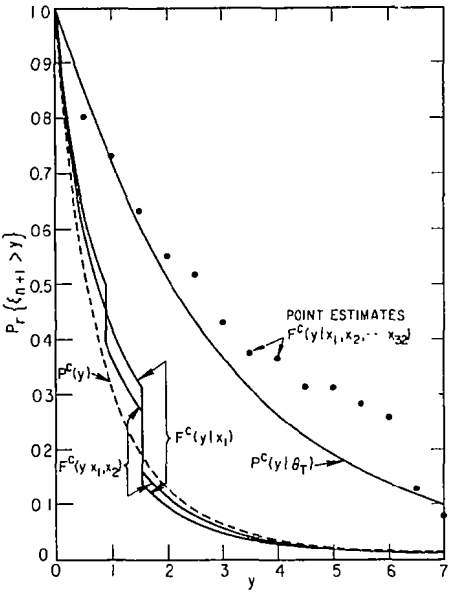

Fig. 4. Credible distribution forecasts Exponential-Gamma distributions $E\{\xi\} = 1$. $V\{\xi\} = 2$. $\theta_T^{-1} = 3$. $n = 1, 2, 32$.

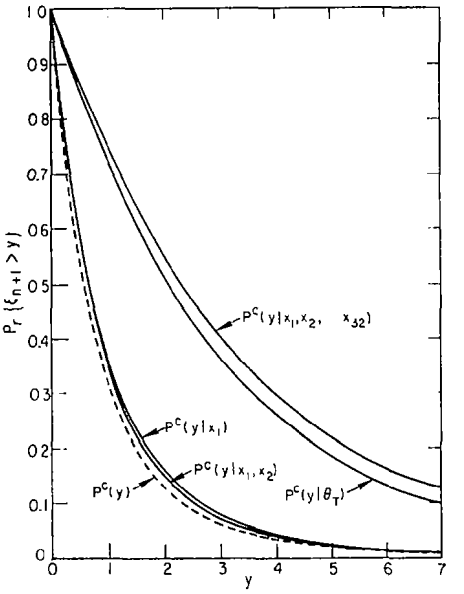

Fig. 5. Bayesian distribution forecasts Exponential-Gamma distributions $E\{\xi\} = 1$. $V\{\xi\} = 2$. $\theta_T^{-1} = 3$. $n = 1, 2, 32$.

comparisons made with the known Bayesian results. Figures 4 and 5 illustrate typical results, using the Exponential-Gamma distributions, with a collective mean of 1, collective variance of 2, and samples from an exponential with mean of 3. Complementary distributions, $P^c(\cdot) = 1 - P(\cdot)$, are used throughout. In Fig. 4, the dotted line represents the prior collective distribution. The first sample drawn was 1.549, and the credible estimate results in a mixed distribution with a discontinuity at that point. The next sample was 0.891, giving the two jumps shown in $F^c(y \mid x_1, x_2)$. Thus far, there has not been much prediction, because the random samples were all low. However, after 32 draws, the sample mean is 3.56, giving the point estimates shown: the actual curve is not drawn because it has 32 jumps, of magnitude $\frac{1}{32 + N_P(y)}$, at the values of the random variates. The dramatic drop between the estimates for $y = 6.0$ and 6.5 is because 5 of the 32 first samples fell here.

Fig. 5 contrasts the result when true Bayesian forecasting is used with the same samples. From (9.2) we see that the conditional distribution is a Shifted Pareto distribution, with updated parameters $a + n$, $b + \Sigma x_t$. This always gives a smooth curve for $P^c(y \mid \nu)$, as shown in Fig. 5. Note that the curves move with the sample mean -too low at first, overestimating the true curve with 32 samples.

It is perhaps unfair to compare the curve of $F^c(y \mid \underline{x})$ with that of $P^c(y \mid \underline{x})$, since the credible distribution only minimizes variance for a fixed value of y. The next example is chosen from some Poisson-Gamma simulations, in which $E\{\xi\} = 1$, $V\{\xi\} = 2$ as before, but where $\theta_T = 2$. In Figs. 6 and 7, we have plotted five simulation runs for $n = 1$ to 16, estimating $P^c(0 \mid \underline{x})$ and $P^c(2 \mid \underline{x})$. There are, in fact five sample paths connected by straight line segments, but they overlap whenever the number of counts $> \nu$ catches up with the number in another sequence of draws, which happens often for an integer-valued random variable. In general, the credible forecast for a given value of y starts at $P^c(y)$ and "relaxes" towards zero whenever no counts $> y$ occur, getting boosted up again whenever a count occurs; this phenomenon is considered further in Section 13.

The Bayesian forecast in Fig. 7, however, will only have the same

values on different simulation runs when the sample means equal each other—only for small n, if at all. For the same random draws the sample path is smoother, however it tends to wander more up and down, instead of following a relaxation curve. Also, there is obviously more correlation between $P^c(y \mid \underline{x})$ for two values of y, since the sample mean is used as a parameter. One can easily trace out corresponding sample paths for $y = 0$ and 2.

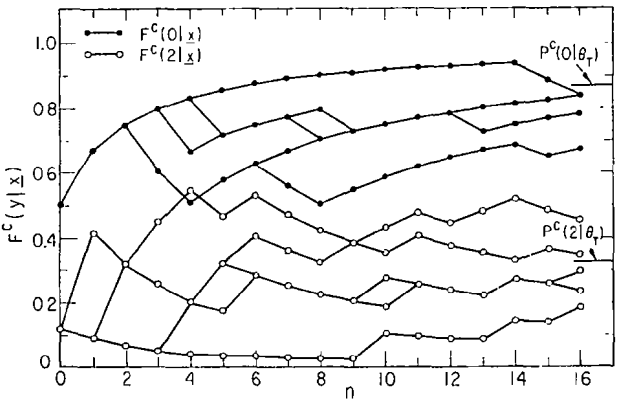

Fig. 6. Credible distributions $F^c(0 \mid x_1, x_2, \ldots, x_n)$ and $F^c(2 \mid x_1, x_2, \ldots, x_n)$ versus n. Five simulations of Poisson-Gamma distributions with $E\{\xi\} = 1$, $V\{\xi\} = 2$, $\theta_T = 2$. (Straight lines for clarity only).

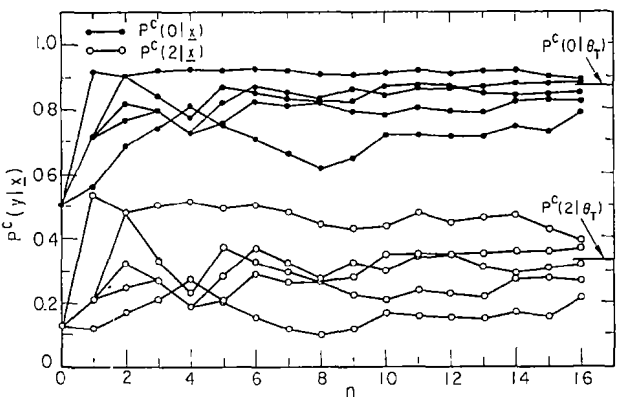

Fig. 7. Bayesian conditional distributions (negative Binomial) $P^c(0 \mid x_1, x_2, \ldots, x_n)$ and $P^c(2 \mid x_1, x_2, \ldots, x_n)$ versus n. Five simulations of Poisson-Gamma distributions with $E\{\xi\} = 1$, $V\{\xi\} = 2$, $\theta_T = 2$. (Straight lines for clarity only).

The variance using the credible forecast is not too large compared with exact Bayesian; only about 70% more for the example shown. By far the biggest contribution to variance is seen to be the random deviates themselves. Even in the Bayesian case, there is still a lot of variance at $n = 16$.

10. Credible Densities

It is difficult to see how to get a credible estimate of the density of a continuous distribution, because of the lack of a natural sample density to replace $\Sigma\, I(y - x_t)/n$. Differentiation leads to unit impulses $\delta(y - x_t)$, and a forecast which is a mixed density at observation points!

However, one can formally use only $\eta = \delta(y - \xi_{n+1})$, and look for a forecast still in terms of the number of counts $\leq y$. The RHS of (3.5) now becomes

$$C\{\eta;\, \omega_t\} = C_0\{p(y \mid \theta);\ P(y \mid \theta)\} \tag{10.1}$$

and we have the formal result

$$p(y \mid \underline{x}) \approx p(x) + \frac{C_0\{p(y \mid \theta);\, P(y \mid \theta)\}}{V_0 P(y \mid \theta)} \cdot Z_P(y) \cdot [\Sigma\, I(x_t - y)/n - P(y)]. \tag{10.2}$$

With a density of a discrete distribution, on the other hand, we are on much safer ground. One can forecast $p(y \mid \underline{x}) = P(y \mid \underline{x}) - P(y - 1 \mid \underline{x})$ by:

(1) A direct credibility approach, counting the number of samples *equal to* y;

(2) An approach similar to (10.2), using the number of samples $\leq y$; or

(3) By differencing the credible distribution (5.7).

We consider the three cases in turn.

For the direct approach, we use:

$$\eta = \delta^y_{\xi_{n+1}};\quad \omega_t = \delta^y_{\xi_t}\quad (t = 1, \ldots, n) \tag{10.3}$$

where δ^j_i is the indicator function, equal to unity if $i = j$, zero otherwise. The analogues of (5.5) and (5.6) carry through in terms of discrete densities:

$$E\{\eta\} = E\{\omega_t\} = p(y); \tag{10.4}$$

$$\operatorname{Cov}\{\omega_i; \omega_j\} = \begin{cases} p(y)\,(1 - p(y)) & (i = j) \\ V_0 p(y \mid 0) & (i \neq j) \end{cases} \tag{10.5}$$

The credible discrete density is then of form:

$$p(y \mid \underline{x}) \approx f(y \mid \underline{x}) = (1 - Z_p(y))\, p(y) + Z_p(y)\, [\sum_{i=1}^{n} \delta^y_{\xi_i}/n], \tag{10.6}$$

with new credibility factors:

$$Z_p(y) = \frac{n}{n + N_p(y)}; \quad N_p(y) = \frac{p(y)\,(1 - p(y))}{V_0 p(y \mid 0)}. \tag{10.7}$$

It is easy to see how this might be estimated in collectives with discrete data, such as automobile claim frequency; only counts of claims for the desired value of y are used in setting up the predictor.

If we adopt the second approach, we keep $\eta = \delta^y_{\xi_{n+1}}$, but use:

$$\omega_i = I(y - \xi_i) = \sum_{j \leq y} \delta^j_{\xi_i}, \tag{10.8}$$

and get the same formal result as (10.2) above. In discrete density notation, this is rewritten.

$$p(y \mid \underline{x}) \approx p(y) + \frac{\sum_{j \leq y} C_0\{p(j \mid 0); p(y \mid 0)\}}{\sum_{i \leq y} \sum_{j \leq y} C_0\{p(i \mid 0); p(j \mid 0)\}} \cdot Z_p^{(y)} \left[\frac{1}{n} \sum_{i=1}^{n} \sum_{j \leq y} \delta^j_{x_i} - \sum_{j \leq y} p(j)\right]. \tag{10.9}$$

Clearly this method uses the internal covariances of all discrete probabilities $\leq y$ and the counts of all observations $\leq y$.

In the third approach, we simply difference (5.7), and get

$$\begin{aligned} p(y \mid \underline{x}) &\approx F(y \mid \underline{x}) - F(y - 1 \mid \underline{x}) \\ &= (1 - Z_P(y))\, p(y) + Z_P(y) \left[\frac{1}{n} \sum_{i=1}^{n} \delta^y_{\xi_i}\right] \\ &+ [Z_P(y) - Z_P(y-1)] \cdot \left[\frac{1}{n} \sum_{i=1}^{n} \sum_{j \leq y-1} \delta^j_{x_i} - \sum_{j \leq y-1} p(j)\right]. \end{aligned} \tag{10.10}$$

This is certainly simpler than (10.9), even though all counts $\leq y$ are used. However, we have not been able to prove that $F(y \mid \underline{\xi})$ is monotone in y for all $\underline{\xi}$, and thus this method might give a negative

forecast. There is also obvious covariance between $F(y \mid \xi)$ and $F(y - 1 \mid \xi)$, and we expect this forecast to have greater variance than (10.9).

11. Use of all Sample Values for Discrete Distributions

An interesting overview of credibility models for discrete distributions can be obtained if we expand our forecast functions to include *all* the discrete values of observations.

Suppose ξ_t attains only discrete values, say $\xi_t \in R$. Define:

$$\omega_i = \sum_{t=1}^{n} \delta_{\xi_t}^{i} \quad (i \in R); \qquad \sum_{i \in R} \omega_i = n. \tag{11.1}$$

This is just the number of samples which attain value i in n trials.

From (3.5), any least-squares prediction problem using the $|R|$ observed values $w_i = \sum_{t=1}^{n} \delta_{x_t}^{i}$ $(i \in R)$, requires the inversion of an $|R| \times |R|$ covariance matrix, whose elements can be shown to be:

$$C\{\omega_i; \omega_j\} = \begin{cases} p(i)\,[1 - p(i)] + (n-1)\,V_\theta p(i \mid \theta) & (i = j) \\ -p(i)\,p(j) + (n-1)\,C_\theta\{p(i \mid \theta); p(j \mid \theta)\} & (i \neq j) \end{cases} \tag{11.2}$$

for the homogeneous collective.

If $|R|$ is finite, the inversion of (11.2) may be carried out by digital computer; for semi-infinite ranges, such as the Poisson, one may deliberately truncate the distribution, or hope for analytic simplifications. (Some special multi-dimensional credibility models are discussed in [20]). In any case it should be noted that because of the constraint (11.1), the matrix is not of full rank; $\sum_{j \in R} C\{\omega_i; \omega_j\} = 0$ $(\forall\, i)$. Thus the rank will be $\leq |R| - 1$. In addition, certain points of mass may not have any across-the-collective variances, and these values of y cannot be predicted beyond $p(y)$; counts at these values may still be useful, however. In the sequel, we shall assume that the range R has been diminished appropriately to the collective structure and predictand and will continue to use notation R.

The RHS to be used depends on the objective. Suppose we are trying to forecast $p(y \mid \theta_T)$ for fixed y. Then:

$$\eta = \delta^y_{\xi_{n+1}}; \quad \mathrm{Cov}\{\eta; \omega_i\} = n\mathrm{C}_0\{p(i \mid 0); p(y \mid \theta)\}. \tag{11.3}$$

The coefficients $\{a^y_0; a^y_j (j \in \mathrm{R})\}$ are found from:

$$a^y_0 = p(y) - \sum_{j \in \mathrm{R}} a^y_j p(j) \tag{11.4}$$

$$\sum_{i \in \mathrm{R}} \mathrm{C}\{\omega_i; \omega_j\} \cdot a^y_j = n\mathrm{C}_0\{p(i \mid 0); p(y \mid \theta)\}. \quad (i \in \mathrm{R}) \tag{11.5}$$

and used in a forecast form:

$$p(y \mid \theta_T) \approx f(x) = a^y_0 + \sum_{j \in \mathrm{R}} a^y_j \left(\sum_{i=1}^{n} \delta^j_{x_i}\right). \tag{11.6}$$

We shall refer to this set of coefficients as the *full multi-dimensional solution to the discrete density*, for fixed y. In a certain sense, it is the best possible solution to the prediction problem, using only a linear function of the individual counts at each value in R.

Now, suppose that the above analyses have been repeated many different times, finding the sets of coefficients $\{a^y_0; a^y_j (j \in \mathrm{R})\}$ *for every value of* $y \in \mathrm{R}$. This requires changing only the RHS of (11.5), and no further inversions of $\mathrm{C}\{\omega_i; \omega_j\}$. Assuming all these sets of coefficients have been found, we can now show the interrelationship between many of the previous models.

First, because of superposition, it is clear that the *full multi-dimensional solution to the discrete cumulative distribution* has the form:

$$P(y \mid \theta_T) \approx A^y_0 + \sum_{j \in \mathrm{R}} A^y_j \left(\sum_{i=1}^{n} \delta^j_{x_i}\right) \tag{11.7}$$

with coefficients:

$$A^y_j = \sum_{\substack{k \leq y \\ k \in \mathrm{R}}} a^k_j \qquad \begin{matrix} (j = 0) \\ \text{or } (j \in \mathrm{R}). \end{matrix} \tag{11.8}$$

In other words, we just cumulate the coefficients from (11.4) and (11.5). Alternately, we can solve the system (11.5) with an RHS of $n\mathrm{C}_0\{p(i \mid 0); P(y \mid 0)\}$.

To obtain the simpler forms given earlier, we merely *constrain* or *eliminate* certain of the coefficients $\{a^y_j\}$ using the remarks in Section 3. The price of these simplifications is, of course, an increase in forecast variance.

For example, when predicting discrete densities, if we set

$$a_j^y = 0 \qquad (j \in R - \{y\}). \tag{11.9}$$

then from (11.4) and (11.5), we get

$$a_y^y = \frac{n V_\theta\{p(y \mid \theta)\}}{p(y)[1 - p(y)] + (n-1) V_\theta\{p(y \mid \theta)\}}; \quad a_0^y = p(y)[1 - a_y^y]. \tag{11.10}$$

This is exactly the credible discrete density (10.6) and (10.7), which uses only counts of observations equal to y.

If, on the other hand, we set

$$a_j^y = 0 \qquad (j > y;\ j \in R), \tag{11.11}$$

and further constrain the nonzero coefficients to be identical:

$$a_j^y = a^y \qquad (j \leq y,\ j \in R), \tag{11.12}$$

a simple calculation will give the second formula for the density (10.9).

Similar remarks apply to estimates of the complete distribution, via the formulas (11.7) and (11.8). For instance, if we set:

$$A_j^y = 0 \quad (j > y; j \in R); A_j^y = A^y \quad (j \leq y;\ j \in R) \tag{11.13}$$

then we get our basic credible distribution formulas (5.7) and (5.8).

To summarize briefly, we see that in the discrete case, the most general way to predict the density, cumulative distribution, or other function of ξ_{n+1} is to solve a multi-dimensional credibility problem, using the counts of observations at all values of y. However, this leads to a requirement for estimating many covariances from the collective and a matrix inversion problem of high order. Simplified formulae and data requirements are obtained by further constraining these least-squares solutions, at the price of increased variance, the results coinciding with those obtained by direct argument.

12. Computational Results-Credible Discrete Density

Computations were carried out for the Poisson-Gamma distributions of (9.1). The density was computed using (10.2), the differences of the credible distribution, and the exact Bayesian forecast.

Typical results are shown in Fig. 8. Generally, the results using (10.2) are farther away from the exact Bayesian forecasts, versus differencing the credible distribution, except for estimates of $p(0 \mid \underline{x}) = 1 - P(0 \mid \underline{x})$, when the results are identical. This performance is due to the limited information used from the samples (counts equal to y), and to the larger time constants, shown below in Table 1.

TABLE 1

Credibility time constants for distribution and discrete-density forecasts. Poisson-Gamma-distribution with $E\{\xi\} = 1$, $V\{\xi\} = 2$.

y	$N_P(y)$	$N_P(y)$
0	2.000	2.000
1	1.793	15.200
2	1.969	11.064
3	2.300	10.185
4	2.748	10.735
5	3.307	12.052
6	3.979	13.949
7	4.773	16.377
8	5 698	19.338

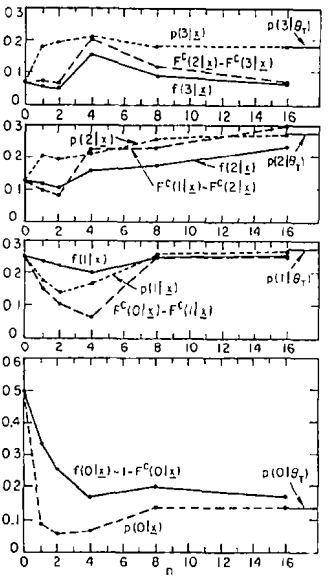

Fig. 8. Discrete probability forecasts for selected values of n. Single simulation of Poisson-Gamma distributions with $E\{\xi\} = 1$, $V\{\xi\} = 2$, $\theta_T = 2$. (Straight lines for clarity only).

13. Rare Counts

To illustrate a limitation of the credible distribution, consider estimating $P^c(y \mid \theta_T)$ for a large value of y, writing the formula:

$$1 - F(y \mid \xi) = \frac{N_P(y)\, P^c(y) + (\#\text{ of } \xi > y)}{n + N_P(y)} . \qquad (13.1)$$

Possible sample paths are illustrated on Fig. 9, assuming n is continuous

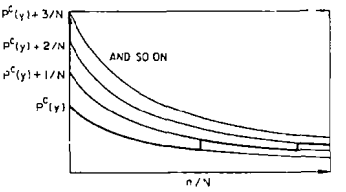

Fig. 9. Sample paths for credibility estimate of $1 - P(y \mid \theta_T)$. $N = N_P(y)$.

We see the familiar "relaxation" of the forecast from the initial estimate of $P^c(y)$ following the curve $P^c(y)/(1 + (n/N))$, until the first count $> y$ causes the estimate to jump up to a curve of similar form which starts at $P^c(y) + (1/N)$. The curve then relaxes again towards zero until the next count occurs. In other words, a given sample path never really converges, but must continually jump up to the neighboring path to stay in the neighborhood of $P^c(y \mid \theta_T)$.

If $P^c(y \mid \theta_T)$ is sufficiently small, then for fixed n not too large, the first jump may not occur with high probability. To a good approximation, then, the credible forecast is a *Bernoulli distribution*, i.e.:

$$1 - F(y \mid \xi) = \begin{cases} \dfrac{P^c(y)}{1 + (n/N)} & \text{with probability } 1 - nP^c(y \mid \theta_T); \\ \dfrac{P^c(y) + 1/N}{1 + (n/N)} & \text{with probability } nP^c(y \mid \theta_T). \end{cases} \qquad (13.2)$$

The mean of this distribution, given θ_T, is just the complement of (8.3), but the variance slightly underestimates the true result (8.6), having instead a leading coefficient $P^c(y \mid \theta_T)\,[1 - nP^c(y \mid \theta_T)]$.

This type of behavior may not be sufficiently accurate for extremely rare events, and suggests estimating more covariances from the collective, and using more complex formulae to obtain more continuously correcting estimates.

The ultimate would be a complete Bayesian analysis which uses the value of every sample at every step to adjust the forecast. However, this requires drastic assumptions about $P(y \mid 0)$ for all values of y.

14. Credible Moments

We conclude with some remarks concerning the problem of estimating various moments of ξ_{n+1}.

First, for the forecast of the mean value, there is the classic formula (4.6), which is known to be exact for most of the well-known conjugate prior distributions such as Beta-Binomial, Gamma-Poisson, Normal-Normal, etc. [8] and [12]; a more general result is shown in [21]. It is easily shown to be incorrect for the Uniform-Pareto and for other families for which the sample mean is not a sufficient statistic [15].

One could also estimate the mean by using the credible distribution or density formulae, (5.7) and (10.6), etc.; numerical integration is necessary in the continuous case because of the awkward dependence of Z upon y.

As an example, Fig. 10, shows the mean for the Gamma-Poisson ($\mathrm{E}\{\xi\} = 1$; $\mathrm{V}\{\xi\} = 2$; $\mathrm{E}\{\xi \mid 0_{77}\} = 2$) calculated four ways:

(1) mean-credible forecast (4.6) (Exact Bayesian);
(2) credible distribution (5.7);
(3) credible density (10.6);
(4) sample mean.

The initial samples in this simulation were quite large, so there are some over-corrections at first; however the response in general is much smoother than that of the distribution forecasts. The Bayesian estimate is, generally the most sensitive to respond to the samples, followed by estimates from the distribution and density. This is obvious from consideration of the magnitudes of the various N.

In his book [4], Bühlmann develops credibility formulae for the conditional variance, $\mathrm{V}\{\xi_{n+1} \mid \underline{x}\}$, based upon separation into a

"variance" part, $E_{\theta|\underline{x}}V\{\xi_{n+1} \mid 0\}$, and a "fluctuation" part, $V_{\theta|\underline{x}}$ $E\{\xi_{n+1} \mid 0\}$. The first part is estimated using the sample variance; the second uses the sample mean. On the other hand, Salmond [15] examines the exact form of the variance for several tractable families, and finds the variance either as a linear or quadratic function of the sample mean only, when the sample mean is a sufficient statistic. Thus the sufficient statistic appears to play the major role in exact results for the variance, but the functional dependence is more complicated.

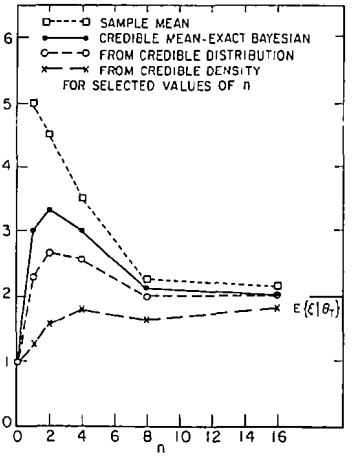

Fig. 10. Forecasts of $E\{\xi_{n+1} \mid x\}$ using four different methods. Gamma-Poisson families. $E\{\xi\} = 1$; $V\{\xi\} = 2$; $\theta_T = 2$.

One can also estimate the variance by estimating $E\{(\xi_{n+1})^2 \mid \underline{x}\}$, and subtracting the square of any estimate of the mean.

If we are trying to estimate the kth moment ($k \geq 2$) then the direct approach via (3.5) and (3.9) is clear. First we set $\eta = (\xi_{n+1})^k$, and select an appropriate predicting function $\omega_1 = g_1(\underline{\xi})$ of the observables.

If the sample mean is known to be a sufficient statistic, one is tempted to set

$$\omega_1 = \frac{1}{n}\sum_{i=1}^{n} \xi_i, \tag{14.1}$$

obtaining the result

$$\mathrm{E}\{(\xi_{n+1})^k \mid \underline{x}\} \approx \mathrm{E}\{\xi^k\} + \frac{\mathrm{Cov}_\theta\{\mathrm{E}\{\xi \mid \theta\}; \mathrm{E}\{\xi^k \mid \theta\}\}}{\mathrm{V}_\theta\{\mathrm{E}\{\xi \mid \theta\}\}} \cdot Z_{[1]} \cdot \left[\frac{1}{n}\sum_{t=1}^{n} x_t - \mathrm{E}\{\xi\}\right]. \qquad (14.2)$$

Thus the fluctuations of the sample mean about the collective mean cause the estimate of the kth moment to change.

Without this foreknowledge, the most natural choice is to take the sample kth moment:

$$\omega_1 = \frac{1}{n}\sum_{t=1}^{n} (\xi_t)^k, \qquad (14.3)$$

obtaining an ordinary credibility formula:

$$\mathrm{E}\{(\xi_{n+1})^k \mid \underline{x}\} \approx (1 - Z_{[k]}) \cdot \mathrm{E}\{\xi^k\} + Z_{[k]} \cdot \frac{1}{n}\sum_{t=1}^{n} (x_t)^k, \qquad (14.4)$$

but with a new time constant:

$$N_{[k]} = \frac{\mathrm{V}\{\xi^k\}}{\mathrm{V}_\theta\mathrm{E}\{\xi^k \mid \theta\}} - 1 = \frac{\mathrm{E}_\theta\mathrm{V}\{\xi^k \mid \theta\}}{\mathrm{V}_\theta\mathrm{E}\{\xi^k \mid \theta\}}. \qquad (14.5)$$

Of course the variances of ξ^k are in fact moments of order $2k$, for which estimates must be found from the collective.

Furthermore, there is no good prior reason why the predictors could not only include *both* (14.1) and (14.3), but *all* sample lth moments, $l = 1, 2, \ldots, k$. Following this approach necessitates estimating all the means of the different moments, as well as covariances of the form:

$$\mathrm{C}\{\xi^i; \xi^j \mid \theta\} \qquad (i, j = 1, 2, \ldots, k). \qquad (14.6)$$

The problem then becomes a multi-dimensional one.

Finally, one can imagine forming the kth moment numerically from the credible distribution.

Regretfully, we must conclude with the observation that there are still many unanswered questions on the efficiency of different approaches. Credibility theory frees us from the distributional assumptions of Bayesian solutions; however, we must now consider

in more detail the form of the approximation, and the availability of statistics from the collective. We must also keep in mind that these estimates are usually made for some decision model in a larger insurance context, and it may be more efficient to examine first the approximations needed at that level.

ACKNOWLEDGEMENT

The credible distribution formula was first presented to a meeting of the Swiss Actuarial Association in Zurich in September, 1972. The various suggestions received since that time are very much appreciated. Special recognition should go to Robert Levin, who efficiently and diligently took responsibility for the computation of the numerical examples.

BIBLIOGRAPHY

[1] BAILEY, A. L. (1945) "A Generalized Theory of Credibility", *PCAS* [1]), Vol. 32, pp. 13-20.

[2] BAILEY, A. L. (1950) "Credibility Procedures, LaPlace's Generalization of Bayes' Rule, and the Combination of Collateral Knowledge with Observed Data". *PCAS*, Vol 37, pp 7-23, (1950); Discussion, *PCAS*, Vol 37, pp. 94-115.

[3] BÜHLMANN, H. (July 1967) "Experience Rating and Credibility", *ASTIN-Bulletin*, Vol 4 Part 3, pp. 199-207.

[4] BÜHLMANN, H. (1970) *Mathematical Methods in Risk Theory*, Springer-Verlag, New York.

[5] BUHLMANN, H. (1971) "Credibility Procedures", *Sixth Berkeley Symposium*, pp. 515-525.

[6] BUHLMANN, H. and STRAUB, E. (1970) "Glaubwürdigkeit fur Schadensätze", *Mitteilungen der Vereinigung Schweizer Versicherungsmathematiker*, Vol. 70, pp. 111-133 Translation by C. E. Brooks, "Credibility for Loss Ratios", (unpublished).

[7] DEGROOT, M. H. (1970) *Optimal Statistical Decisions*, McGraw Hill, New York.

[8] ERICSON, W. A. (1969) "A Note on the Posterior Mean of a Population Mean", *Journal of the Royal Statistical Society*, (B), Vol 31, No 2, pp. 332-334.

[9] FERGUSON, T. S. (1967) *Mathematical Statistics: A Decision Theoretic Approach*, Academic, New York.

[10] HEWITT, C. C., Jr. (1971) "Credibility for Severity", *PCAS*, Vol. 57, pp. 148-171, (1970); Discussion, *PCAS*, Vol. 58, pp. 25-31.

[11] LONGLEY-COOK, L. H. (1962) "An Introduction to Credibility Theory", *PCAS*, Vol. 49, pp. 194-221; Available as a separate report from Casualty Actuarial Society, 200 East 42nd St., New York, N.Y 10017.

1) Proceedings of the Casualty Actuarial Society (and predecessors).

[12] MAYERSON, A. L. (1965) "A Bayesian View of Credibility", *PCAS*, Vol. 51, pp. 85-104, (1964); Discussion, *PCAS*, Vol. 52, pp. 121-127.
[13] MAYERSON, A. L., JONES, D. A. and BOWERS, N. L., Jr. (1969) "On the Credibility of the Pure Premium", *PCAS*, Vol. 55, pp 175-185, (1968); Discussion, *PCAS*, Vol. 56, pp. 63-82.
[14] PERRYMAN, F. A (1932) "Some Notes on Credibility Theory", *PCAS*, Vol. 19, pp. 65-84
[15] SALMOND, D. (February, 1973) "Premium Calculation in Casualty Insurance", ORC Report 73-2, Operations Research Center, University of California, Berkeley.
[16] SEAL, H. L. (1969) *Stochastic Theory of Risk Business*, Wiley, New York.
[17] STRAUB, E. (1972) "On the Calculation of IBNR-Reserves", *IBNR-The Prize-Winning Papers in the Boleslaw Monic Fund Competition Held in 1971*, Nederlandse Reassurantie Groep, N.V., Amsterdam, pp. 132.
[18] WHITNEY, A. W. (1018) "The Theory of Experience Rating", *PCAS*, Vol. 4, pp. 274-292
[19] JEWELL, W. S. (April, 1973) "Multi-Dimensional Credibility", ORC Report 73-7, Operations Research Center, University of California, Berkeley.
[20] JEWELL, W. S., "Models of Multi-Dimensional Credibility", (in preparation).
[21] JEWELL, W. S., "Credible Mean is Exact Bayesian for Simple Exponential Families", (in preparation)