Credibility

By Gary Venter

CREDIBILITY

BY GARY VENTER

Section 1 - Introduction and History

Until recent years, classical statistics had focussed on estimating a quantity based only on directly relevant observations; peripherally relevant or seemingly unrelated series which may provide further information had been excluded. Since at least the early 1900's, however, casualty actuarial practice has incorporated related information, sometimes in a fairly ad hoc manner, under the name of "credibility."

Classical statistical procedures estimate a value, such as the average age of a group, by taking a sample from the group and using the mean value of that sample as the estimate. Credibility estimation makes use of the sample value, but may incorporate other information as well, such as the average age of similar groups. In ratemaking, for example, the experience of the latest period might be regarded as a sample from all possible time periods. Rather than using this by itself, even properly adjusted for premium and loss levels, to determine the new rate, other information might be incorporated, such as the old rate, or rates for related exposures.

If the new rate is taken to be a weighted average between the indication from the data and the old rate or some other estimate, the weight applied to the data is called the credibility weight, or sometimes, more loosely, the credibility of the data. The

latter terminology may be misleading, however, in that it seems to imply that the credibility weight is an inherent property of the data. This will turn out to not be the case. In addition to any features of the data itself, the context in which it will be used, including what it is to be weighted against, will explicitly or implicitly influence the credibility to be assigned.

Credibility theory incorporates the entire study of this weighting process, including development of the formulas for assigning the credibility weights, as well as estimation of the parameters or values that appear in these formulas.

Although pragmatically motivated, credibility weighting now has both theoretical and practical justification. Credibility formulas can be derived from statistical assumptions, and they have proven useful in application. This chapter outlines the background and use of credibility theory. Being an overview, the results are in many cases given without proof, or the proofs are just outlined. Underlying assumptions are included, however. As with many disciplines, the real world is often more complex than the initial assumptions, and more intricate models are often needed in order to be truly practical. The more practical models are presented in the later sections, but their exposition will benefit from the simpler paradigms covered first.

To illustrate the type of related information that may be useful, imagine that an estimate is desired for the quantity of ice cream

a particular person will consume next year. The average consumption for that individual for the last few years might be selected as the estimate. However this estimate could probably be improved by giving some degree of weight to the average consumption of the population at large.

Another example, examined in greater detail below, is to estimate a baseball player's season batting average from the early season performance. In this case it has been shown that giving weight to the early season averages of other players can considerably improve the estimate.

More typical insurance examples include estimation of claim frequency, severity, or total loss cost for an insured, a class, or a rating territory. Experience for other insureds, classes, states, insurers, etc. may be the auxiliary data incorporated.

In all of the above cases the auxiliary information comes from a wider, more stable population. There is, however, another type of credibility application; rather than incorporating a wider population, earlier observations of a single series may be used. For something like claim frequency countrywide, for example, the latest observation by itself could be regarded as sufficient; however if this is subject to significant random fluctuations, some weight may be given to prior years, perhaps with the weights decreasing to zero after some time. Credibility formulas used in

this case are somewhat different from those incorporating a wider population.

Limited Fluctuation vs. Greatest Accuracy

By the mid 1920's, two fairly different approaches to credibility had been established. The terminology noted was introduced by Arthur Bailey in his far reaching 1945 paper. Basically speaking, the limited fluctuation approach aims to limit the random component of an estimate; the greatest accuracy approach attempts to make the estimation error as small as possible. The example below shows how each of these might be applied in a single series case.

The series in question, N_i, could be anything of interest, e.g., state loss ratios, countrywide frequency, etc. To have a concrete example, let N_i denote the number of doctor visits made by members of the U.S. Congress in year i. C_i will denote the credibility estimate of N_i made based on the data through year i-1. In a single series situation, both the limited fluctuation and greatest accuracy approaches to credibility make use of a credibility weight Z_i between 0 and 1 so that:

$$C_{i+1} = (1-Z_i)C_i + Z_iN_i \qquad (1.1)$$

This weight, however, has a different purpose and derivation in the two approaches.

The limited fluctuation approach seeks to limit the fluctuations in the series of estimates C_j, at least insofar as those fluctuations are due to the randomness inherent in the series of observations N_j. The greatest accuracy approach, on the other hand, seeks to minimize the estimation errors. To be more precise, this approach specifies then seeks to minimize an error function. Usually the expected squared difference between the estimated and actual value is the function to be minimized. In the current example this would be denoted as $E(C_i-N_i)^2$. With this error function, the greatest accuracy approach is referred to as "least squares" credibility.

To illustrate the formulas for Z_i that arise from these two approaches, a few additional assumptions will be introduced. N_i is hypothesized to be approximately normally distributed with mean M_i and constant variance v. (Constant in that it is the same for each year.) The mean M_i is hypothesized to change each year by the random amount D_i, that is, $M_{i+1}=M_i+D_i$. D_i is a random variable with mean zero and variance d. The D's are assumed to be independent of each other and of M_1. Because of the mean zero, each M_j has the same unconditional expected value, denoted by m, i.e., $E(M_i)=m$. (M_i is treated as a random variable because its value is not known, which in part is due to the random term D_i.) Because the M's change each year by the D's, the variance of M increases each year by d. E.g., $Var(M_{i+1})=Var(M_i)+d$. Thus $Var(M_j)=w+(j-1)d$, with $w=Var(M_1)$.

The estimation process has to start somewhere, and so C_1 is the estimate of N_1 made before any of the observations N_j are available. This estimate could have been based on previous knowledge of similar processes, for example. C_1 can also be considered as an estimate of M_1. The variance w can be interpreted as an expression of the uncertainty about the value of M_1 before N_1 is observed; as such it may influence how willing we will be, when estimating N_2, to give up on C_1 in favor of N_1 once it becomes available.

Given these model assumptions, the calculation of the credibility factor Z_i under the two approaches can be addressed. The limited fluctuation approach calculates Z_i based on the conditional distribution of N_i given M_i. It seeks to limit the impact on the credibility estimator (1.1) of random deviations of the observation N_i from its conditional expected value M_i. In other terms, it seeks to guarantee, at least to an acceptably high probability, that the quantity:

$$Z_i(N_i-M_i) \qquad (1.2)$$

stays within certain bounds.

The criterion for limiting the deviation is established by first specifying a probability level p, e.g., p=.95, and then requiring, with a probability of at least p, that (1.2) be no greater than some prespecified maximum. In this case that maximum will

be taken to be km, where k is a selected small number, e.g., k=.05. Recall that m is the unconditional mean $E(M_i)$. In other words, Z_i is sought so that $Pr(Z_i(N_i-M_i)\leq km)=p$.

To see the impact of this criterion, the credibility estimate (1.1) can be rewritten as $C_{i+1}=(1-Z_i)C_i+Z_iM_i+Z_i(N_i-M_i)$. These three terms can be regarded as representing stability, truth, and random noise. Since truth and noise cannot be observed separately, the same factor Z_i applies to both. The highest possible factor is sought, so that truth will be emphasized, as long as noise can be kept within acceptable bounds. Thus the value of Z_i is sought that will keep $Z_i(N_i-M_i)$ below 100k% of the expected value m with probability p.

Since N_i has a symmetric distribution about its mean M_i, that Z_i will also ensure that the absolute value of the random component is less than km with probability 2p-1. Limited fluctuation credibility as so formulated emphasizes the conditional distribution of N_i given M_i, but the conditioning is not always noted explicitly.

The value of Z_i that meets the criterion is $Z_i= km/y\sqrt{v}$, where y is the 100pth percentile of the standard normal distribution. To show this, by hypothesis, given M_i, $(N_i-M_i)/\sqrt{v}$ has the standard normal distribution, and so from the definition of y, we have

$Pr[(N_i-M_i)/\sqrt{v} < y] = p$. Multiplying both sides of the inequality by $Z_i\sqrt{v}$ then gives $(Pr[Z_i(N_i-M_i)<km]=p$, as desired.

In most applications of this approach, Z_i is regarded as a function of m, and the values of m that lead to different credibility levels are sought. The variance v is often taken to be proportional to m, e.g., v=cm. This yields $Z_i = (k/y)\sqrt{m/c}$.

Z_i is capped at 1, even though the formula value may be higher. For selected p and k, the value of m that yields $Z_i=1$ is referred to as the full credibility value, and is given by $m_F=c(y/k)^2$. Then for $m<m_F$, Z_i can be conveniently computed by the square root rule:

$$Z_i = \sqrt{m/m_F} \qquad (1.3)$$

which can be verified by substituting $c(y/k)^2$ for m_F in (1.3).

The least squares approach for determining Z_i does not start with formula (1.1), but derives it as the result of a more general estimation problem: N_{i+1} is to be estimated as a linear combination of the previous observations $N_1,\ldots N_i$, with the expected squared error to be minimized. That is, coefficients b_j are sought to minimize:

$$E[N_{i+1}-(b_0 + \Sigma_{j=1}^{i} b_jN_j)]^2 \qquad (1.4)$$

It turns out, after much algebra, that the solution to this estimation problem can be expressed in the form (1.1), that is, as a credibility formula.

The Z_i that do this are computed recursively by:

$$Z_1 = 1/[1 + K] \qquad (1.5a)$$

$$Z_{i+1} = 1/[1 + 1/(J+Z_i)] \qquad (1.5b)$$

where K=v/w and J=d/v. The details of the derivation, including more general conditions under which (1.4) leads to (1.1), are found in Gerber and Jones (1974). For the interested reader, a sketch of the proof is below.

The minimization of (1.4) is accomplished by first setting its partial derivatives with respect to the b_j to zero. This produces i+1 equations, one for each b_j. For example the partial of (1.4) with respect to b_1 produces the equation:

$$E[N_1(N_{i+1}-b_0-\Sigma_{j=1}^{i} b_j N_j)]=0.$$

All these equations involve terms like $E(N_j)$ and $E(N_jN_h)$, which are then evaluated in order to solve for the b_j's.

To illustrate this procedure, evaluating the $E(N_jN_h)$ type term is outlined. Note that, given M_j, N_j and N_{j+h} are independent, and the conditional expected value of each is M_j, i.e., $E(N_j|M_j)=M_j$ and $E(N_{j+h}|M_j)=M_j$. This is because $M_{j+h}=M_j+D_j+\ldots+D_{j+h-1}$, and the D_i's have mean zero. Then it follows that $E(N_jN_{j+h}|M_j)= M_j^2$, and eventually that $E(N_jN_{j+h})= Var(M_j)+m^2= w+(j-1)d+m^2$. After evaluating all such terms and combining them algebraically, (1.5) is produced. (End of sketch of proof).

A heuristic interpretation of (1.5) can be made. Note that Z_1 is an increasing function of w and a decreasing function of v. The uncertainty about M_1 is measured by w; thus the greater this uncertainty, the greater is the weight given to the observation N_1. But the uncertainty about M_i is not the only thing considered, the stability of N_1 is greater with lower v and this also leads to greater weight on N_1.

Z_{i+1} is an increasing function of Z_i and J. Greater stability (low v) continues to give greater credibility through a higher J; a higher d also increases the credibility which makes sense as follows: a high d indicates that the M's are greatly subject to change, so the older estimates should be given less weight, with more to the current observation, i.e., higher Z_j's.

It might also be noted that if d happens to equal $w^2/(v+w)$, all the Z's are the same. This can be verified by finding Z_2 from (1.5); it seems an unlikely coincidence, however.

The formulas above are fairly representative of what is produced by the least squares and limited fluctuation approaches to credibility. The limited fluctuation approach will always involve a full credibility value, representing the degree of random fluctuation deemed acceptable. The square root rule for partial credibility is also fairly typical of this approach. The only variance explicitly treated is v, which represents the

random fluctuation of a single observation around its own generally unknown mean.

Formula (1.5a) is fairly typical of least squares credibility; often it is somewhat generalized to Z=P/[P+K], where P is a measure of the volume of data observed. Besides recognizing the random fluctuation measured by v, this formula also incorporates the relevancy of the previous estimate, which w quantifies. Also quantifying changes in the process over time, which d achieves, is a further step not always incorporated into least squares analysis. Thus (1.5b) is a less typical but more general example of a least squares credibility formula.

It should be noted that while the limited fluctuation approach does not explicitly recognize the relevance of the previous estimate or the degree of likely process change over time, judgments about these issues may be incorporated into the selection of the degree of random fluctuation deemed acceptable, as specified by the fluctuation k allowable with probability p.

Historical Perspective

Credibility as known today is generally traced to Mowbray (1914), writing in volume I of the Proceedings of the Casualty Actuarial Society. Decidedly in the limited fluctuation camp, Mowbray's article approximates an assumed binomial claim count process by the normal distribution to derive the full credibility standard

relative to p and k. The goal of the limited fluctuation approach as practiced today is suggested by the title of Mowbray's article: "How extensive a payroll exposure is necessary to give a dependable pure premium?"

The greatest accuracy approach was introduced by Whitney (1918), writing in volume IV of the CAS Proceedings. Whitney assumed that the number of claims for an employer with P employees is binomially distributed with parameters (P,M), and that M itself is normally distributed. The resulting credibility for that employer's experience can be expressed as Z=P/[P+K], with K a function of the binomial and normal variances. The complement 1-Z is applied to the experience of the entire class, as indicated by the class rate, rather than to the previous experience of the employer.

Both Mowbray and Whitney were addressing Workers Compensation experience rating. An application of the limited fluctuation paradigm to automobile classification ratemaking can be found in Stellwegen (1925). Group life insurance experience rating using greatest accuracy credibility was explored by Keffer (1929), who assumed a Poisson claim count distribution with a gamma distribution on the Poisson parameter. Perryman (1932) addressed a number of then current issues, including an interpretation of the limited fluctuation square root rule similar to that discussed

above, i.e., a way to give the credibility estimate no larger a random component than a risk with full credibility would have.

The least squares approach to greatest accuracy credibility was established in Bailey (1945), although the notation was cumbersome. Buhlmann and Straub (1970) formalized the derivation of Z=P/[P+K], K=v/w, from a least squares error criterion, showed that this was valid for all finite variance distributions, and discussed a method of estimating the variances v and w.

Least squares credibility was recognized by Bailey (1950) to replicate the Bayesian posterior mean for the normal-normal and beta-binomial models. Keffer's result essentially shows this for the gamma-Poisson case, and it is also known for the gamma-gamma pair. Essentially, the posterior mean is the best least squares estimator; credibility provides the best linear least squares estimator. Thus when the posterior mean is a linear function of the observations, the two estimators are the same. Ericson (1970) characterized a family of distributions for which this is the case.

More recent research has emphasized generalizations and applications of the original models. Topics include improved estimation of parameters, credibility for trend and regression models, credibility incorporating more than one type of prior estimate, credibility weighting of the prior with a "hyperprior", methods

of incorporating more complex relationships between firms of different sizes, methods of improving estimates for distributions with nonlinear posteriors, and treating parameters that may change over time. Many of these generalizations arise because the real world is more complicated than the original models assume; thus in addition to requiring more theory, they are more practical as well.

Section 2 - Review of Statistical Concepts

An understanding of some basic statistics will be presumed in this chapter. A few of the topics most germane to credibility theory will be briefly reviewed in this section, but reference to statistical texts may be required if some material has been unused recently.

Two concepts that will be called upon frequently are covariance and conditional distributions. To review, for two random variables X and Y, the covariance of X and Y is defined as:

$$Cov(X,Y) = E[(X-EX)(Y-EY)] \qquad (2.1)$$

and often can be calculated more conveniently by:

$$Cov(X,Y) = E(XY)-(EX)(EY) \qquad (2.2)$$

Thus $Cov(X,X)=Var(X)$. The covariance of X and Y divided by the product of their standard deviations yields the correlation coefficient. The covariance is zero when X and Y are independent, but not necessarily vice versa.

Recall that if $f(x,y)$ is the joint density for X and Y, then the marginal density for X is defined as:

$$f_X(x) = \int f(x,y)dy \tag{2.3}$$

The integral is taken over the entire support of Y, and the resulting marginal density is basically the probability density function for X. The same thing can be done for Y. The conditional density of Y given X is defined by:

$$f(y|x) = f(x,y)/f_X(x) \tag{2.4}$$

This is interpreted as the density function for Y given that X takes on the value x.

Substituting $f_Y(y)f(x|y)$ for $f(x,y)$ in (2.3), and substituting the result of that for $f_X(x)$ in (2.4) yields Bayes' rule:

$$f(y|x) = f_Y(y)f(x|y)/\int f_Y(y)f(x|y)dy \tag{2.5}$$

which is used to get from one conditional distribution to another. Once x is fixed, the denominator of (2.5) is the constant needed to make the entire right hand side a probability density, i.e., make it integrate to unity. In many applications this constant can be computed later, or not at all, and so Bayes' rule can be written:

$$f(y|x) \propto f(x|y)f(y) \tag{2.6}$$

where "$\propto$" is read "is proportional to". In (2.6) and hereafter, the subscript on the marginal density is dropped unless it is needed to avoid confusion.

Conditional moments can be defined by using the conditional densities in the usual moment definitions. For instance:

$$E(Y|X=x) = \int y f(y|x)dy \tag{2.7}$$

$$Var(Y|X=x) = \int (y-E(Y|X=x))^2 f(y|x)dy \tag{2.8}$$

Since $EY = \int\int y f(x,y)dydx = \int E(Y|X=x)f_X(x)dx$,

$$EY = E[E(Y|X)] \tag{2.9}$$

Similarly it can be shown that

$$VarY = EVar(Y|X) + VarE(Y|X) \tag{2.10a}$$

$$Cov(Y,Z) = ECov(Y,Z|X) + Cov(E(Y|X),E(Z|X)) \tag{2.10b}$$

In applications of Bayes' rule, some distributions are described as prior, conditional, posterior, or predictive. To introduce and illustrate this terminology, an example using some well known distribution functions is given.

Example 2.1

A population of drivers is insured by XYZ insurance company, and each driver has a Poisson distribution for the number of physical damage claims to be submitted each year. For a given driver, let N_i denote the claim count random variable for year i, and y the driver's Poisson parameter, which is assumed not to vary over time. Then the conditional density is $f(n|y)=e^{-y}y^n/n!$, which has mean and variance both equal to y, and skewness of $y^{-1/2}$. In this example it is supposed that y is not known, but it is a random variable having the gamma distribution with parameters b and c, which has the density $f(y) = y^{c-1}e^{-y/b}/b^c(c-1)!$. Here and throughout the chapter a! will be used to denote $\Gamma(a+1)$, as they

agree on integers and this can be used to define a! at other points.

This distribution is considered the prior distribution for Y, which is now capitalized to signify that it is a random variable. The gamma distribution in b,c has mean bc and variance b^2c, and in general, $EY^j=b^jc(c+1)\cdots(c+j-1)$ when j is a positive integer, and $EY^j=b^j(c+j-1)!/(c-1)!$ for any real $j>-c$. The shape of the distribution is determined by c; b is referred to as the scale parameter.

The unconditional or mixed distribution for N is its marginal distribution with density $f_N(n) = \int f(n,y)dy=\int f(n|y)f(y)dy$. This is the distribution the insurer faces for the driver's claim counts, as it combines the process distribution for N given Y with the parameter distribution for Y. It is sometimes referred to as the mixture of the process distribution by the parameter distribution. Doing the integration finds this to be a negative binomial distribution, with parameters c and p=1/(1+b). The negative binomial density with parameters c and p is $f(n)=(c+n-1)!p^c(1-p)^n/n!(c-1)!$. This has mean c(1-p)/p, variance $c(1-p)/p^2$, and skewness $(2-p)(c-cp)^{-1/2}$.

From the mixed distribution it can be found that $EN_1=cb$, since for p=1/(1+b), (1-p)/p=b. This could have been calculated using $EN_1=EE(N_1|Y)$, because $E(N_1|Y)=Y$, from the Poisson distribution,

and $EY=bc$ from the gamma distribution. Similarly, $VarN_1=cb(b+1)$. From (2.10) this should equal $EVar(N_1|Y)+VarE(N_1|Y)=EY+VarY=cb+cb^2$. These two components of the total variance are sometimes referred to as "expected value of process variance" and "variance of hypothetical means", respectively. The latter terminology considers $E(N_1|Y)$ as hypothetical, since Y is not a known quantity.

The posterior distribution is the density for Y given N_1, as calculated by Bayes' rule, and can be used to update the prior distribution once an observation is available. By Bayes' rule, $f(y|n)\propto f(n|y)f(y)$. The proportionality means that any factors not involving y can be computed later, as the integral of $f(y|n)dy$ must equal 1. Thus $f(y|n) \propto e^{-y}y^n y^{c-1}e^{-y/b} = y^{n+c-1}e^{-y(1+1/b)}$. But from the gamma density above, the gamma distribution in parameters $b/(b+1)$ and $(n+c)$ is proportional to this same quantity, so that must be the posterior distribution of Y.

A measure of the dispersion of a random variable relative to its mean is the coefficient of variation, or CV, which is the ratio of the standard deviation to the mean. For the gamma in b,c this is given by $1/\sqrt{c}$, and so reduces to $1/\sqrt{n+c}$ for the posterior gamma.

Finally, the predictive distribution is the marginal distribution of N_2 resulting from the mixture of the Poisson model by the posterior gamma distribution for Y given N_1. Since a Poisson mixed by a gamma in b,c gives a negative binomial in c,1/(b+1), the Poisson mixed by the posterior gamma in b/(b+1),n+c can be seen to give negative binomial parameters n+c,(b+1)/(2b+1). This is the distribution for N_2 the insurer faces for this driver after observing N_1=n. It has mean (n+c)b/(b+1), which can be written as Zn+(1-Z)bc, with Z=b/(b+1). This can be interpreted as a credibility weighting between the observation n and the previous mean bc.

The usefulness of the predictive distribution goes beyond estimating the subsequent expected value. It gives the probabilities for N_2=j for all values of j, and thus quantifies the possible divergence of actual from expected results.

Exercise

a. Calculate $EE(N_2|Y)$, where the outer expected value uses the posterior gamma above.

b. Calculate Var N_2:

   1. As $EVar(N_2|Y) + VarE(N_2|Y)$.

   2. Directly from the predictive distribution.

When, as in this example, the posterior distribution is of the same type as the prior, just with different parameters, the prior

and conditional distributions are said to be conjugate. Since the posterior of N_1 becomes the prior of N_2, etc., conjugate distributions allow for continued updating of the parameters of a single distribution type as subsequent data becomes available.

Thus the gamma-Poisson combination is a conjugate pair. Another is the inverse gamma-gamma pair, as the next example illustrates.

Example 2.2

In this example, the total workers compensation losses X_i for a certain factory in year i are assumed to be gamma distributed with parameters y,c. Here, however the scale parameter y is not known, but is specified by the prior distribution

$$f(y)=y^{-r-1}e^{-b/y}b^r/(r-1)! \qquad (2.12)$$

This is referred to as the inverse gamma distribution in b,r because Y^{-1} is gamma distributed in b^{-1},r. The moments are given by $E(Y^j) = b^j/(r-1)(r-2)\cdots(r-j)$ for positive integers $j<r$ and $E(Y^j) = b^j(r-j-1)!/(r-1)!$ for any real number $j<r$. If $j\geq r$, the jth moment does not exist. In particular $E(Y) = b/(r-1)$ for $r>1$ and $VarY = b^2/(r-1)^2(r-2)$ for $r>2$. Note that this prior can be specified simply as $f(y)\propto y^{-r-1}e^{-b/y}$, and the conditional by $f(x|y)\propto e^{-x/y}x^{c-1}y^{-c}$.

The posterior can then be calculated as $f(y|x) \propto f(x|y)f(y) \propto e^{-x/y}y^{-c}y^{-r-1}e^{-b/y} \propto e^{-(x+b)/y}y^{-c-r-1}$. But this is the inverse

gamma in (x+b), (c+r). This shows the conjugate nature of the pair.

The mixed distribution is $f(x) = \int f(x|y)f(y)dy$, and turns out to be $f(x) = b^r x^{a-1}(c+r-1)!/(b+x)^{c+r}(c-1)!(r-1)!$ This is a generalization of both the F-distribution and the shifted Pareto, and has been called different names. Here it will be referred to as the Beta2 in b,c,r, following McDonald ( ). The moments are given by:

$$E(X^j) = b^j c(c+1)\cdots(c+j-1)/(r-1)(r-2)\cdots(r-j) \qquad (2.11a)$$

for positive integers $j<r$ and

$$E(X^j) = b^j(c+j-1)!(r-j-1)!/(c-1)!(r-1)! \qquad (2.11b)$$

for any real j, $-c<j<r$.

In particular $E(X_1) = cb/(r-1)$ and $Var(X_1) = b^2c(c+r-1)/(r-1)$.

Exercise

Calculate EX and VarX via (2.9) and (2.10).

The predictive distribution for X_2 given $X_1=x$ is the conditional gamma mixed by the posterior inverse gamma and is thus the Beta2 in $(x+b),c,(c+r)$. For $r>1$, this has mean $E(X_2|X_1=x) = (x+b)c/(c+r-1)$. Letting $Z = c/(c+r-1)$, $1-Z = (r-1)/(c+r-1)$, and then the predictive mean can be expressed as $E(X_2|X_1=x) = Zx + (1-Z)E(X_1)$. It is also possible to write $Z = 1/(1+K)$, by letting $K = (r-1)/c$. Thus again a credibility formula arises for the

predictive mean. As will be seen below, this does not always happen, but it does for an important class of distributions.

Diffuse Priors

In the above example the prior distribution could have come from information about the distribution of risks within the class. Lacking such information a prior can be developed by actuarial judgement. When information and judgement lack precision, it is often felt best to make the prior as nonspecific as possible. One method that has been developed to do this is to use so called diffuse priors. One class of diffuse priors for a positive parameter y is specified by $f(y) \propto y^p$. There is no value of p for which the integral of y^p over the positive reals is finite; thus no constant can be calculated to make f(y) a proper density function. Nonetheless, $f(x|y)y^p$ may have a finite integral, and if so, a posterior distribution can be calculated. A more detailed discussion of diffuse priors may be found in Berger ( ).

For instance, p=0 specifies a uniform prior on the positive reals. For this p, the integral from 0 to M is finite, while that from M to infinity is not, for any number M, no matter how large. This may seem to give too much weight to large possible values of y. For example, the likelihood of y being between 1 and 2 is the same as for it being between 1,000,000,000,001 and 1,000,000,000,002.

The infinite part of the integral of y^p is from M to infinity for $p>-1$, and from 0 to ε for $p<-1$. In the latter case the weight is on values of y near zero. For p=-1 neither the interval 0 to ε

nor the interval M to infinity has a finite integral. Thus for p=-1, even though the probability is concentrated in unlikely places (near zero and infinity), there is no clearcut pull by the prior to higher or lower values of y.

Example 2.3

In Example 2.2, suppose the prior had been specified as $f(y) \propto y^p$. Then $f(y|x) \propto e^{-x/y} y^{p-c}$. As long as $p<c-1$ this is an inverse gamma posterior, with parameters x and c-p-1. The predictive distribution will thus be the Beta2, with parameters x, c, and c-p-1. Thus the predictive mean is $cx/(c-p-2)$. For p=-2 this is equal to the observation x, which is an appealing result in that it takes the observation at face value. As mentioned above, p=-1 seems to make more sense as a prior; for the predictive mean this increases the observation by a factor of $c/(c-1)$, as long as $c>1$. This also has a logical interpretation, in that $c/(c-1)$ is the ratio of the conditional mean to conditional mode, which is the most likely observation. For p=-1, the posterior inverse gamma has parameters x_1,c and the predictive Beta2 is in x_1,c,2c. Repeated application after n observations yields a predictive Beta2 in $\sum_{i=1}^{n} x_i$,c,nc. If $c<1$, this will eventually have a finite predictive mean when $nc>1$.

Example 2.4

In Example 2.1, taking $f(y) \propto y^p$ yields the posterior $f(y|n) \propto e^{-y} y^{n+p}$. This is a gamma distribution in 1, n+p+1 as long as p>-n-1. The predictive mean is n+p+1, which for p=-1 yields the observation n.

Note that in both of these examples the posterior and conditional distributions are conjugate, and so can then be used to begin the Bayesian updating process as more observations become available.

Aggregate Claims Distributions

The application of credibility to insurance problems often involves a decomposition of the total losses into frequency and severity components. This part of the statistical preliminaries will be the calculation of the moments and percentiles of aggregate claims from those for frequency and severity.

The definition of the aggregate claims T for a given period is:

$$T = X_1 + \cdots + X_N \qquad (2.12)$$

where N is the number of claims in the period and X_i is the amount of the ith claim. It is usually assumed that the X_i are independent of each other and of N, and that all the claims follow a common severity distribution. Thus, the subscripts can be dropped when referring to the severity random variable X.

The moments of T are given by:

$$E(T)=E(N)E(X) \tag{2.13a}$$

$$Var(T)=E(N)Var(X)+E(X)^2Var(N) \tag{2.13b}$$

$$Skw(T)=(Skw(X)CV^3+3n_2CV^2+n_3)/\sqrt{E(N)(CV^2+n_2)^3} \tag{2.13c}$$

Here CV denotes the severity coefficient of variation, and $n_i=E(N-EN)^i/EN$. (2.13a) is proven from (2.9) by noting $E(T)=EE(T|N)=E(NE(X))=E(N)E(X)$. (2.13b) follows similarly from (2.10) since:

$$\begin{aligned} Var(T) &= EVar(T|N)+VarE(T|N) \\ &= E(NVar(X))+Var(NE(X)) \\ &= E(N)Var(X)+E(X)^2Var(N) \end{aligned}$$

This could alternatively be computed by evaluating $E(T^j|N)$ via (2.9), which is what is used to derive (2.13c).

One method of estimating the percentiles of T is to assume a particular distributional form, e.g., T is normal or gamma distributed. If the moments of X and N are given, the distribution for T can then be estimated from its moments, which are computed via (2.13). The normal distributional assumption incorporates a skewness of zero. The gamma has a skewness of twice its coefficient of variation. This is probably more realistic for property-casualty lines, but neither distributional form is likely to be correct.

Several approaches to improved estimation of percentiles of aggregate claims have been developed. One is to incorporate a third parameter so that the first three moments can be matched. For instance for the normal distribution the so called normal power approximation (NP) incorporates a skewness correction as follows. Let σ_T denote the standard deviation of T, and t_p the pth percentile. Then the normal approximation estimates t_p by $ET+\sigma_T y_p$, where y_p is the pth percentile of the standard normal. The NP approximation for t_p is:

$$t_p=ET+\sigma_T y_p+\sigma_T SkwT(y_p^2-1)/6 \qquad (2.14)$$

This NP formula is derived using a power series expansion for t_p. Pentikainen ( ) recommends its accuracy only for SkwT<1, after which the NP tends to exaggerate the difference between the percentiles t_p and their normal approximation estimates.

Another approximation for aggregate claims is offered by Seal (====), who adds a third parameter to the gamma that shifts the origin to the left or right. The percentiles are calculated using a fairly simple modification to the gamma distribution function. Pentikainen ( ) finds the accuracy of this approximation comparable to that of the NP.

Another way of adding a third parameter to the gamma is to use a power transform, i.e., to assume T^a is gamma distributed for some real number a. If a=-1 this gives the inverse gamma distribution

used above. Applications of this method can be found in Venter (1984).

It is also possible to compute the aggregate distribution function without making a distributional assumption for the aggregate claims. However this usually requires knowledge of the density functions for frequency and severity, not just their moments. One such method is simulation. A possible number of claims n is generated according to the frequency distribution, then n possible claim sizes are drawn from the severity distribution. This gives one possible realization of T. This process can be repeated many thousands of times to estimate the distribution function of T. While conceptually simple, this process is often expensive and time consuming.

Another method is to build up the aggregate distribution function recursively, i.e., the probability that T<t is computed from the probabilities that T is less than t-1, t-2, etc. Panjer ( ) shows a fairly efficient way to do this for a discrete severity distribution and a Poisson, negative binomial, or binomial frequency. For the Poisson frequency, dePril ( ) finds an even more efficient algorithm for a piecewise linear distribution function.

Finally, a method of calculating aggregate claim probabilities based characteristic functions is becoming widely used. The

characteristic function is a complex analog of the moment generating function, and can be computed for aggregate claims from the moment generating function of frequency and the characteristic function of severity. The distribution function for aggregate claims can be recovered from its characteristic function via numerical integration. The calculation is thus somewhat intricate, but once programmed it is fairly efficient. One difficulty is calculating the severity characteristic function, as it is not usually of closed form. This method was pioneered by Mong ( ), who used a gamma severity. Heckman and Meyers ( ) extended it to a step function probability density, and Venter ( ) generalized this to a piecewise linear density. The latter two severity functions can be used to approximate other distributions, thus making this method of quite general application.

Section 3 - Limited Fluctuation Credibility

The limited fluctuation credibility estimator can be expressed as:

$$C=(1-Z)M+ZT \qquad (3.1)$$

where T is the observation and M is a previous estimate. M is generally supposed to be the estimate one would use if the observation T were not available, and it could come from previous experience and/or related data. Typically T will be the loss ratio, pure premium, frequency, or severity for a class, state,

or risk for a certain time period, and C estimates its value for another, usually future, period. Here, to be specific, T will be the aggregate losses for a one year period, thus $T=X_1+\cdots+X_N$ as above, with the usual independence assumptions.

For limited fluctuation theory, (3.1) can be rewritten as:

$$C=(1-Z)M+ZET+Z(T-ET) \qquad (3.2)$$

only the last term is considered random, and the goal of the theory is to keep its contribution within specific bounds. In particular, k and p are selected and then Z is sought so that $Pr(Z(T-ET)<kET)=p$. For example, $p=.95$ and $k=.05$ are typical choices and result in requiring that the random component of (3.2) be less than 5% of the expected value ET with 95% probability.

Actually this requirement is only an upper bound on Z(T-ET), but in applications T is always assumed to be symmetric or slightly skewed to the right, so that this upper bound requirement guarantees that $Pr(|Z(T-ET)|<kET) > 1-2(1-p)$. Thus for $p=.95$, $k=.05$, the credibility requirement provides that the random component has 90% probability of being less than 5% of ET in absolute value.

The criterion can be restated as $Pr(T<ET+kET/Z)=p$, or $t_p=ET + kET/Z$, where t_p again is the pth percentile of T. To find Z, different methods of computing t_p can be invoked. Under the

normal approximation, $t_p=ET+y_p\sqrt{VarT}$, and so $Z=kET/y_p\sqrt{VarT}$. In terms of frequency and severity.

$$Z^2=(k/y_p)^2(ENEX)^2/\ (ENVarX+(EX)^2VarN) \tag{3.3a}$$

$$=(k/y_p)^2(EN)^2/(ENCV^2+VarN) \tag{3.3b}$$

$$=(k/y_p)^2EN/(CV^2+n_2) \tag{3.3c}$$

Where again CV is the severity coefficient of variation and n_2 is the frequency ratio of variance to mean. $Z=1$ when $EN = (CV^2+n_2)(y_p/k)^2$. This value of EN is called the full credibility value, denoted as n_F. The value of EN that produces credibility Z, n_Z, can be seen to follow $n_Z=Z^2n_F$, or $Z=\sqrt{n_Z/n_F}$.

This "square root rule" holds only for the normal approximation. For the NP,

$$t_p=ET+\sqrt{VarT}(y_p+SkwT(y_p^2-1)/6) \tag{3.4}$$

and so $kET/Z=\sqrt{VarT}(y_p+SkwT(y_p^2-1)/6$. This can be solved for Z in term of frequency and severity moments using (2.13) to yield:

$$Z=k/[y_p\sqrt{m_2/EN}+(m_3/m_2)(y_p^2-1)/6EN] \tag{3.5}$$

where m_2 and m_3 are aggregate claim shape descriptors defined by:

$$m_2=n_2+CV^2$$

$$m_3=CV^3SkwX+3n_2CV^2+n_3$$

The normal approximation formula (3.3c) can then be seen to be the special case $m_3=0$, i.e., $SkwT=0$, which the normal approximation assumes, but which is unlikely in practice. The square root rule does not apply for the NP credibilities; rather they must be

calculated from (3.5) directly. It is possible to solve (3.5) for EN by considering it a quadratic in $\sqrt{EN}$. This produces a formula for n_Z, the value of EN needed for credibility Z:

$$n_Z=(Z^2/4k^2)(y_p\sqrt{m_2}+\sqrt{y_p^2m_2+2(k/Z)(y_p^2-1)m_3/3m_2})^2 \qquad (3.6)$$

Both (3.3) and (3.5) have an important invariance property: the calculation of Z from EN is affected neither by simple monetary inflation nor the addition of independent identical distributed exposure units. In fact, without the latter invariance, Z could not really be regarded as a function of EN. The former allows credibility standards to remain constant until the shape of the severity distribution changes.

The invariance under simple monetary inflation results because the severity coefficient of variation and skewness do not depend on scale. The latter invariance follows because the frequency mean, variance and third central moment are all additive functions; that is, the additional units will increase these moments all by the same factor. Thus n_2 and n_3 will not be affected. (Anyone who thinks this is because these three moments are all cumulants is probably correct.)

An example may help clarify these concepts.

Example 3.1

Commercial fire losses for a state are assumed to have a Poisson frequency distribution and a lognormal severity, with CV=7. For the Poisson, n_2 and n_3 both equal 1, and so with this CV, m_2=50. The skewness of the lognormal is given by $SkwX=CV^3+3CV$, and so for this example SkwX=343+21=364. Thus $m_3=364\cdot 343+3\cdot 49+1=125,000$. The credibility requirements are specified by p=.95 and k=.05, which gives y_p=1.645 from a normal table.

The normal approximation n_F is given by $n_F=m_2(y_p/k)^2$, and thus in this case is $50(1.645/.05)^2=54,120$. For the NP, n_F can be calculated via (3.6) to be 80,030. Thus considering skewness has a substantial impact in this case, basically because the severity distribution is highly skewed. The assumption of a CV of 7 for commercial fire is consistent with the findings of Simon (1969). The skewness of aggregate claims may be calculated as $SkwT = m_3/m_2^{1.5}\sqrt{EN}$, which in this case is 1.25. This is somewhat above Pentikainen's recommendation for the boundary of the accuracy of the NP, and thus the NP n_F estimate may be somewhat too high.

Instead of the lognormal severity, it is interesting to consider a constant severity. This could arise, for example, in a group of life insurance policies all with the same benefit. In this case, CV=0, and $m_2=m_3=1$. For the normal approximation, n_F then becomes 1082, which has been a widely used credibility standard.

For the NP approximation, n_F is 1094, via (3.6). Thus for the Poisson alone, the skewness correction is not substantial.

The negative binomial frequency could have been used instead of the Poisson. With parameters c and p, $n_2=1/p$ and $n_3=(2-p)/p^2$. In a study of automobile claims, Dropkin (1959) found $n_2=1.184$. This implies $p=.8446$, and so $n_3=1.620$. For the constant severity case, $m_2=n_2$ and $m_3=n_3$; the normal approximation then yields $n_F=1282$ and the NP gives $n_F=1297$. For the lognormal above, n_F increases to 54,320 under the normal approximation, and to 80,150 with the NP. Thus the negative binomial assumption with these parameters seems to have some impact in the frequency only case, but little when a highly skewed severity has already been included.

Exercise

Verify the calculations in the paragraph above.

Meyers and Schenker (1983) discuss the possibility that the negative binomial n_2 may be substantially larger than 1 for individual large commercial risks. In their model, exposure units are not independent, so some of the above reasoning does not apply. However it is instructional to explore the implications of a large n_2. Thus suppose a negative binomial distribution is given with $n_2=51$. Then $p=1/51$, and $n_3=5151$. For the

above severity, m_2=100, and m_3=137,500. Then the normal and NP n_F's are 108,200 and 123,400 respectively.

Exercise

Verify these n_F's. What would they be for frequency only? How many claims would be neede for 50% credibility under the normal and NP approximations?

The limited fluctuation Z depends only on the distribution of T, and treats the previous estimate M as a constant. Thus Z does not depend on how good this estimate may be or where it comes from, although such matters could influence the selection of p and k, on which Z depends. If T is the aggregate losses for a state, M could be the previous year's estimate. If T represents only a single class or territory, M could be the statewide estimate for the same year. In general, M is supposed to be the best estimate available without the particular observation T, and in fact may be formed as a combination of other estimators.

The nondependence of Z on the properties of the previous estimator is both a strength and a weakness. It provides flexibility and a simple algorithm for routine application, and does not require the estimation of additional parameters. However it may ignore or only judgementally consider elements that can be quantified with some additional research. The least squares methodology, to be reviewed next, takes such an approach.

Section 4 - Least Squares Credibility

In the least squares theory, the previous estimator applied to the complement of credibility is specified much more explicitly. Consequently, more details of the estimation problem need to be modelled. This requires some notation. To have a particular problem to work with, it will be supposed that the losses for N risks are observed for a period of n years. The pure premium for the ith risk in year u is denoted as X_{iu}. Pure premium is loss divided by exposure; for now all risks are assumed to have the same number of exposure units, which is constant over time. In Section 6, application of credibility theory to risks of different sizes will be made.

The pure premium for a future time period, time 0, is to be estimated for the gth risk. This will end up being estimated as a credibility weighting of the average observed pure premium for risk g over the n years, denoted as $X_{g.}$, with the grand average of all the risks for those years, denoted as $X_{..}$. In formulas, $X_{g.} = \sum_u X_{gu}/n$, and $X_{..} = \sum_g X_{g.}/N$.

The credibility given to the risk experience will depend in part on the stability of that experience, as in limited fluctuation theory, but it will also depend on the relevance of the grand mean to the individual risk, which is quantified by the variance across risks of the individual risk means. The greater this variance, the more diverse are the risks, and thus the grand mean

provides less relevant information about an individual risk. This will in turn lead to greater credibility assigned to the risk's own experience, and less to the grand mean. The explicit consideration of the relevancy of the estimator applied against the complement of credibility is one of the distinctive features of least squares credibility.

The least squares credibility estimator could be dervied by finding the weight Z that minimizes $E[X_{g0}-(ZX_{g.}+(1-Z)X_{..})]^2$, and this approach will in fact be followed in Section 7. However, the same estimator also arises as a result of a more general estimation problem as follows. X_{g0} is estimated as <u>any</u> linear combination of all the observations X_{iu}, not just a weighted average of $X_{g.}$ with $X_{..}$, with the expected squared error to be minimized. The general linear combination of the observations can be expressed as $a_0+\Sigma_{i,u}a_{iu}X_{iu}$, so the credibility criterion will be to find the weights (a's) that minimize:

$$E[X_{g0}-(a_0+\Sigma_{i,u}a_{iu}X_{iu})]^2 \qquad (4.1)$$

It will turn out that the resulting weights can be combined into a simple credibility formula, which gives further justification for such a formula.

There are Nn+1 weights a_{iu} to find, and this is approached by setting the partials of (4.1) with respect to these variables to zero. Doing this, with some algebraic manipulation, produces the following system of Nn+1 equations:

$$EX_{g0} = a_0 + \Sigma_{i,u} a_{iu} EX_{iu} \qquad (4.2a)$$

$$Cov(X_{g0}, X_{jv}) = \Sigma_{i,u} a_{iu} Cov(X_{iu}, X_{jv}) \qquad (4.2b)$$

There are Nn equations expressed by (4.2b), one for each j,v combination.

Exercise

Derive (4.2). Hint: The partial with respect to a_0 will give (4.2a). Set the partial with respect to a_{jv} to zero and subtract (4.2a) multiplied by EX_{jv} from this equation.

In order to solve this system for the a's, more model assumptions are needed, so that the covariances can be evaluated. As an example, a fairly simple model will be investigated first. It will be assumed that the risk i loss ratio for time u can be decomposed as follows:

$$X_{iu} = m + R_i + Q_{iu} \qquad (4.3)$$

Here m is the overall average, R_i is a risk effect that does not vary over time, and Q_{iu} is a random fluctuation. The R's and Q's are treated as random variables, as their values are not known. The average over all risks of the R_i's is assumed zero, i.e., $ER_i=0$. Also it is assumed that $EQ_{iu}=0$, and so $EX_{iu}=m$. This is an overall expected value; $E(X_{iu}|R_i)=m+R_i$ is the conditional expected value for the ith risk. Finally, it is assumed that different Q's and R's are independent random variables with $VarR_i=t^2$ and $VarQ_{iu}=s^2$.

To compute $Cov(X_{iu},X_{jv})$ under these assumptions, it will be convenient to introduce the following notation: $\delta_{ij}=1$ if $i=j$; otherwise $\delta_{ij}=0$.

With this in hand, note that $E(R_iR_j)=\delta_{ij}t^2$: since different R's are independent, if $i\neq j$, $E(R_iR_j)=ER_iER_j=0$; also, $ER_i^2=VarR_i+(ER_i)^2=t^2$. Similarly, $E(Q_{iu}Q_{jv})=\delta_{ij}\delta_{uv}s^2$.

Now, by definition of covariance,

$$\begin{aligned}Cov(X_{iu},X_{jv}) &= E[(X_{iu}-EX_{iu})(X_{jv}-EX_{jv})]\\ &= E[(X_{iu}-m)(X_{jv}-m)]\\ &= E[(R_i+Q_{iu})(R_j+Q_{jv})]\\ &= E[R_iR_j] + E[Q_{iu}Q_{jv}] \text{ (by independence of R's and Q's)}\end{aligned}$$

And thus,

$$Cov(X_{iu},X_{jv}) = \delta_{ij}t^2 + \delta_{ij}\delta_{uv}s^2 \qquad (4.4)$$

The notation says this covariance is zero unless $i=j$, in which case it is t^2, unless also $u=v$, in which case it is t^2+s^2. This means that $VarX_{iu}=s^2+t^2$, which can also be expressed as $VarX_{iu}=EVar(X_{iu}|R_i)+VarE(X_{iu}|R_i)$ - the expected process variance plus the variance of the hypothetical means.

Exercise

Show that $EVar(X_{iu}|R_i)=s^2$ and $VarE(X_{iu}|R_i)=t^2$.

Because so many of the covariances are zero, plugging (4.4) back

into (4.2b) will make many terms drop out, and in fact produces the equation:

$$\delta_{gj}t^2 = \Sigma_u a_{ju}t^2 + a_{jv}s^2 \qquad (4.5)$$

There is still one such equation for every j,v combination; for fixed j summing all the v-equations (n of them) produces:

$$n\delta_{gj}t^2 = n\Sigma_u a_{ju}t^2 + \Sigma_u a_{ju}s^2 \qquad (4.6)$$

and so,

$$\Sigma_u a_{ju} = n\delta_{gj}t^2/(s^2+nt^2) \qquad (4.7)$$

Plugging this into (4.5) will yield, after some algebra:

$$a_{jv} = \delta_{gj}t^2/(s^2+nt^2) \qquad (4.8)$$

This says the weight is zero unless j=g, and then it is:

$$a_{gv} = t^2/(s^2+nt^2) \qquad (4.9)$$

To find a_0, substitute

$$EX_{iu} = m \qquad (4.10)$$

into (4.2a) to yeild $m=a_0+\Sigma_{i,u}a_{iu}m$, and so $a_0=m(1-\Sigma_{i,u}a_{iu})=m(1-\Sigma_u a_{gu})=ms^2/(s^2+nt^2)$. Finally, since the estimator of X_{g0} is $a_0+\Sigma_{i,u}a_{iu}X_{iu}$, which simplifies to $a_0+\Sigma_u a_{gu}X_{gu}$, the credibility estimator can be written as:

$$\tilde{X}_{g0} = ms^2/(s^2+nt^2) + \Sigma_u X_{gu}t^2/(s^2+nt^2) \qquad (4.11)$$

Now $\Sigma_u X_{gu}$ may be written as $nX_{g.}$; defining $Z=nt^2/(s^2+nt^2)$ produces $\tilde{X}_{g0} = (1-Z)m + ZX_{g.}$; here a natural estimate for m would be $X_{..}$, and in fact this is the minimum variance unbiased estimate of m (see ISO (1983)). Substituting this estimate gives:

$$\tilde{X}_{g0} = (1-Z)X_{..} + ZX_{g.} \qquad (4.12)$$

Thus the best linear estimate of X_{g0} turns out to be a credibility formula. This formula can alternatively be derived as the least squares linear estimate having $a_0=0$ but constrained to be unbiased (see ISO (1983)).

From the definition of Z it can be seen that if t^2 is higher, so is the credibility given to the risk experience. Since t^2 measures the dispersion of individual risk conditional means around the grand mean m, it can be seen that greater dispersion leads to greater credibility; the more different a risk is likely to be from the average, the greater credence will be placed on a risk's own experience. On the other hand, Z reduces as s^2 increases; higher s^2 means that the risks are less stable over time, and thus less reliance can be placed on their individual results. This was also seen in limited fluctuation credibility.

By defining $K=s^2/t^2$, Z can be written as $Z=n/(n+K)$, which is basically Whitney's 1918 formula.

The credibility formulas illustrated by this simple model will be found to hold in more general situations. In fact, if (4.4) and (4.10) are satisfied, the rest of the development will be the same, ending up with (4.12) with the same definition of Z.

This example is typical in one respect, which is in the division of the uncertainty about X_{iu} into two components: a time invar-

iant risk specific component (here R_i), and a random fluctuation in each time period (Q_{iu}). Some observers may feel that this distinction is somewhat artificial, because neither component is ever observed in isolation; however it is an intuitively reasonable distinction, and leads to a model that seems to have practical value.

A More General Model

In the simple model above, each risk had one parameter R_i which described the risk, and then a random fluctuation. More generally it is now assumed that each risk has a vector of parameters, denoted by R_i, that describe the risk, which still nonetheless is subject to random fluctuation. For example, a risk with a negative binomial frequency distribution and an inverse gamma severity would have four parameters describing these distributions, and random fluctuation from year to year as provided by those distributions. Letting R denote the vector $\langle R_1, R_2, \ldots, R_N \rangle$, it is assumed that X_{iu} and X_{jv} are conditionally independent given R. Each risk has its own conditional mean and variance, which may be denoted by $E(X_{iu}|R)=E(X_{iu}|R_i)=m_i$ and $Var(X_{iu}|R)=Var(X_{iu}|R_i)=s_i^2$

It is assumed that for different i the R_i are independent identically distributed random vectors with $Em_i=m$, $Varm_i=t^2$, and $Es_i^2=s^2$. This implies that $EX_{iu}=m$ and $VarX_{iu}=s^2+t^2$ (why?).

Here again, s^2 is the expected process variance and t^2 is the variance of the hypothetical means.

In order to apply (4.2), it is necessary to compute $Cov(X_{iu},X_{jv})$ from these assumptions. By (2.10),

$$Cov(X_{iu},X_{jv})=ECov(X_{iu},X_{jv}|R)+Cov[E(X_{iu}|R),E(X_{jv}|R)] \quad (4.13)$$

Now by the conditional independence of the X's, the first term is zero unless i=j and u=v, in which case it is $EVar(X_{iu}|R)=Es_i^2=s^2$. The second term is also $Cov[E(X_{iu}|R_i),E(X_{jv}|R_j)]$, which by the independence of R_i and R_j is zero unless i=j, in which case it is $Varm_i=t^2$. Thus:

$$Cov(X_{iu},X_{jv}) = \delta_{ij}\delta_{uv}EVar(X_{iu}|R) + \delta_{ij}VarE(X_{iu}|R) \quad (4.14)$$

$=\delta_{ij}\delta_{uv}s^2+\delta_{ij}t^2$, which is (4.4). Plugging this back into (4.2) will then yield (4.12) by the same reasoning used for the original simplified model above.

Example 4.1

Suppose severity is constant at one unit, frequency is Poisson in R_i, and exposure is one. Then the pure premium X_{iu} is the number of claims for risk i in time u; by the Poisson hypothesis, $m_i=R_i$, and $s_i^2=R_i$ as well. If R_i is gamma distributed in b.c. then $t^2=Varm_i=VarR_i=b^2c$, and $s^2=Es_i^2=ER_i=bc$. Thus $K=s^2/t^2=1/b$, and $Z=n/(n+K)=nb/(nb+1)$. For n=1 this gives the predictive mean computed in Example 2.1.

Example 4.2

X_{iu} is assumed to be gamma distributed in $R_i=<Y_i,c>$. Thus $m_i=Y_ic$, and $s_i^2=Y_i^2c$. Y_i is assumed inverse gamma in b,r; so $t^2=Varm_i=c^2VarY_i=c^2b^2/(r-1)^2(r-2)$, and $s^2=Es_i^2=cEY_i^2=$

$cb^2/(r-1)(r-2)$; then $K=s^2/t^2=(r-1)/c$, and $Z=n/(n+K)= nc/(nc+r-1)$. Thus $\tilde{X}_{g0}$ agrees with the predictive mean from Example 2.2.

Section 5 - Estimation of K

Up until now, s^2 and t^2 were treated as known constants, but in practice they usually have to be estimated. One approach is to estimate s^2 based on observed deviations of risk annual results from risk means, and t^2 from observed deviations of risk means from the grand mean. Sometimes it is more convenient to estimate the total variance $VarX_{iu}=s^2+t^2$ from the deviations of individual risk observations from the grand mean, and then get s^2 or t^2 by subtraction. This is simplified when the conditional distribution is Poisson, because then the conditional mean and variance are equal, so $s^2=EVar(X_{iu}|R_i)=ER_i=m$, the grand mean.

Example 5.1

A group of 300 car owners in a high crime area submit the following number of theft claims in a one year period:

Number of Claims:	0	1	2	3	4	5
Number of Owners:	123	97	49	21	8	2

Each owner is assumed to have a Poisson distribution for X_{i1}, the number of thefts, but the mean number may vary from one owner to another. A credibility estimate is desired for X_{i0}, the number of claims for each driver for the next year.

The average number of claims per driver can be calculated to be 1.0. By the Poisson assumption, this is also s^2. The average value of X_{i1}^2 can be found to be 2.2, so $s^2+t^2=VarX_{i1}$ can be

estimated to be $2.2-1.0^2=1.2$. This implies $t^2=0.2$, and so $K=5$, and $Z=1/6$. The credibility estimate for X_{i0} is thus $5/6+X_{i1}/6$. In general, estimating s^2 and t^2 separately can be approached by calculating the statistics $S_i=\sum_u(X_{iu}-X_{i.})^2/(n-1)$, $S=\sum_i S_i/N$, and $T=\sum_i(X_{i.}-X_{..})^2/(N-1)$. The expected value of these statistics can be calculated (laboriously) from (4.4) and (4.10). As a hint of how that might proceed, multiplying out the squares in S and T result in a whole lot of terms of the form $X_{iu}X_{jv}$, whose expected values then need to be evaluated. This is done using (4.4) and (4.10), which together imply that $E(X_{iu}X_{jv})=m^2+\delta_{ij}(t^2+\delta_{uv}s^2)$. The answers are: $E(S_i|R_i)=s_i^2$; $ES_i=EE(S_i|R_i)=Es_i^2=s^2$; $ES=s^2$; $ET=t^2+s^2/n$.

The formula for S_i looks like a fairly usual statistical result. T looks like it should be something like t^2, but probably a little bit higher, because some extra fluctuation is added from the use of the estimated means rather than m_i and m. Thus the formula for T looks about right also. From these formulas, S is an unbiased estimator of s^2, nT is an unbiased estimator of s^2+nt^2, and $T-S/n$ is an unbiased estimator of t^2. Since $1-Z=s^2/(s^2+nt^2)$, it could be estimated by S/nT, as both numerator and denominator are unbiased. Such an approach may be satisfactory in many cases, and is supported by the independence of S and T, as shown by Klugman (1985).

Example 5.2

Table 5.1 displays the pure premium experience for 9 risks, all with the same constant number of exposure units, for a 6 year period. $X_{i.}$ and S_i are calculated from this experience, as shown, and $X_{..}$=.563, S=.357, and T=.066 can then be computed from the formulas above. These yield S/nT=.899, which can be used as the estimate of 1-Z, and so Z is estimated to be .101.

Table 5.1

Risk	Year 1	Year 2	Year 3	Year 4	Year 5	Year 6	Xi.	Si
1	.430	.375	2.341	.175	1.016	.466	.801	.649
2	.247	1.587	1.939	.712	.054	.261	.800	.615
3	.661	.237	.063	.250	.602	.700	.419	.072
4	.182	.351	.011	.022	.019	.252	.139	.021
5	.311	.664	1.002	.038	.370	2.502	.815	.792
6	.301	.253	.044	.109	2.105	.891	.617	.622
7	.219	1.186	.431	1.405	.241	.804	.714	.251
8	.002	.058	.235	.018	.713	.208	.206	.071
9	.796	.260	.932	.857	.129	.349	.554	.121
							.563	.357

An important issue in credibility theory is the accuracy of this estimate of Z. For this example, Table 5.1 was generated by taking random draws from assumed gamma distributions for each risk. The parameters of these gamma distributions are shown below, along with the risk conditional means and variances.

Risk	b	c	Mean	Variance
1	.6159	1.0476	.6452	.3974
2	.8001	0.9063	.7251	.5802
3	.6098	0.9654	.5887	.3590
4	.2391	0.9219	.2204	.0527
5	.5206	1.0184	.5302	.2760
6	.6768	1.0937	.7402	.5010
7	.9575	1.1395	1.0911	1.0447
8	.1999	1.0153	.2030	.0406
9	.5083	0.9320	.4737	.2408
			.5797	.3880

Thus m=.5797, which is not too different from $X_{..}$ and s^2=.388, which again is fairly close to S. The variance of the above

conditional means can be found to be $t^2=.0664$, and thus $t^2+s^2/n=.1311$, which is fairly different from T. Thus the "population" value of 1-Z of $s^2/(s^2+nt^2)=.493$ and Z=.507 is quite a bit different from that estimated by the data.

This experiment was repeated twice more, that is, six years of data were simulated two more times, with the following results:

Experiment	S	T	1-Z	Z
1	.357	.066	.899	.101
2	.274	.103	.443	.557
3	.219	.172	.211	.789

Note:Calculations based on unrounded values

Thus this method does not seem to be able to produce a close estimate of Z with this quantity of data when the process is this unstable. However the average of the three estimates, .482, is just slightly below the underlying value of Z, .507, which gives hope that with just somewhat more data good estimates are possible. Estimating the variance of the estimated Z would help provide an understanding of the accuracy of the calculation, and this is discussed further in Section 7.

Empirical Bayesian Approaches - the (N-1)/(N-3) Correction

Estimating 1-Z by S/nT has a drawback in that while the numerator and denominator are both unbiased, 1-Z is not. This is a typical problem for quotients of unbiased estimators. In this case it arises because $E(1/T)>1/ET$ (see exercise below). This implies that $E(1/nT)>1/s^2+nt^2$, and thus $E(S/nT) > s^2/(s^2+nt^2)$, i.e., S/nT overstates 1-Z, and thus understates Z, on the average.

Exercise

Show that $E(1/T) \geq 1/ET$. Hint: Schwartz' inequality says that $[\int g(t)h(t)dt]^2 \leq \int g(t)^2 \int h(t)^2$. Take g^2 and h^2 to be $tf(t)$ and $f(t)/t$. Equality occurs only in degenerate cases.

The excess of $E(1/T)$ over $1/ET$ varies from one distribution to another, so it is not possible to find a general correction for the bias in Z. This excess is greater for heavy tailed distributions, however, so an approximate lower bound could be found by computing its value in the normal distribution case.

The calculation of 1-Z when both the conditional and prior distributions are normal has been the focus of a field known as parametric Empirical Bayes statistics. A classic article in this field is Efron and Morris (====). Following Morris (====), Klugman (1985) shows that S and T are independent random variables, with S gamma distributed in $2s^2/N(n-1)$, $N(n-1)/2$ and T gamma in $2(t^2-s^2/n)/(N-1)$, $(N-1)/2$. By the gamma moment formula, if Y is gamma in b, c then $E(1/Y)=1/b(c-1)$ as long as $c>1$, and is non-existent otherwise. This is greater than $1/EY$ by a factor of $c/(c-1)$. For T the value of c is $(N-1)/2$, so $ET^{-1}=ET(N-1)/(N-3)$, as long as $N>3$. Thus $E(S(N-1)/(N-3)nT)=1-Z$, and so $S(N-1)/(N-3)nT$ is an unbiased estimator of 1-Z. This is the above credibility estimator of 1-Z adjusted by the factor $(N-1)/(N-3)$. For $N<4$, credibility weighting would not be

recommended by this school, as then E(S/nT) would not be finite, and so no correction factor could make S/nT unbiased.

Evaluating the (N-1)/(N-3) Correction

For other conditional and prior distributions, the excess of E(1/T) over 1/ET is likely to be greater than for the normal, so the (N-1)/(N-3) correction factor is probably a lower bound. However, there is a potential problem with this correction factor which could cause it to actually overcorrect for bias, namely that it does not take into account the usual pratice of capping 1-Z at 1. The true value of 1-Z, $s^2/(s^2+nt^2)$, must be in the range [0,1]. In practice, however, the calculated value of nT may be less than S, which would make S/nT>1. Typically 1-Z would be capped at 1 in this case, giving Z=0. However by this practice the estimator of 1-Z has effectively become min[1,S/nT], which has a lower expected value than S/nT. That is, the capped estimator has lower bias than S/nT, and may even be unbiased or be biased in the other direction.

Even knowing the distributions for S and T, as in the normal-normal case, Emin[1,S/nT] does not have a closed form expression. It can be calculated numerically by:

$$\text{Emin}[1,S/nT] = \int\{F_T(s/n) + (s/n)\int_n^\infty t^{-1} f_T(t)dt\}f_S(s)ds$$

A practical problem with this expression is that the distribution functions of S and T depend on s^2 and t^2, and these cannot be brought out explicitly; however in one special case - when s^2 is

a known constant - Emin[1,S/nT], the expected value of the uncorrected estimate of 1-Z, can be calculated as a function of 1-Z. Some results are shown in the table below:

		Emin[1,S/nT]		
N	1-Z:	.800	.500	.200
3		.799	.673	.426
18		.809	.557	.227

Thus when 1-Z is large, the capped estimator does not seem to be upwardly biased, although this is not true for smaller factors. Even for N=3, the bias is finite, and thus credibility weighting is a useful possibility even in that case.

Exercise

A population of risks with X_{iu} gamma in b_i, c_i is determined by independent draws of the b's from a uniform distribution on [0,1], and the c's from a uniform distribution on [.85,1.15]. What is K? (Hint: the uniform with width a has variance a/12; use $E(b^2)=Var(b)+(Eb)^2$ and independence of b and c to find $E(b^2c)$; compute Var(bc) via $E(b^2c^2)-(EbEc)^2$.)

Example 5.3

The answer to the previous exercise is K=3.88. As a test of the correction factor in a non-normal distribution case, 100 risks were drawn from such a population, and the average values of b_ic_i and $b_i^2c_i$ were found to be .500 and .330. These compare to the expected values from the uniform distribution of .500 and .333.

The variance of the $b_i c_i$'s (hypothetical means) was .085, compared to a theoretical value of .086. For each risk, 6 years of data were simulated, as in Example 5.2, and this process was repeated for five different experiments. For n=6, 1-Z=.393 from the uniform prior, and 1-Z=.392 from the b's and c's actually drawn, so this is the target value to which the estimate values of 1-Z can be compared. For the five experiments, 1-Z was estimated using the (N-3)/(N-1) factor, as the impact of the capping by 1 did not seem large. The following results occurred:

Experiment	X..	S	T	1-Z
1	.504	.434	.164	.432
2	.482	.302	.109	.452
3	.455	.303	.122	.406
4	.510	.362	.141	.419
5	.471	.277	.102	.443

For comparison, the anticipated value of T is .086+.333/6=.142 from the uniform prior, and .085+.330/6=.140 from the 100 risks actually selected. The small but consistent overstatement of 1-Z may be due to $E T^{-1}$ being greater for these distributions than the normal-normal (N-3)/(N-1) correction contemplates. The Bayesian methods discussed below give slightly higher estimates of 1-Z, but they also provide estimates of the potential error, as will be addressed later.

Bayesian Estimates of Z

In the normal-normal case, some work has been done on Bayesian estimates of s^2 and t^2. This could be done to reflect some prior belief, however faint, in where s^2 and t^2 are likely to be: it

also produces an entire posterior distribution for these parameters, rather than just point estimates.

Since S and T are both conditionally gamma distributed given s^2 and t^2, inverse gamma priors, as discussed in Example 2.2, could be postulated. As an example, s^2 will be taken to be inverse gamma distributed in p, 2, and the quantity t^2+s^2/n is given an independent inverse gamma in q, 2. While this approach ends up providing reasonable estimates, it does have a theoretical problem in that some possibility that t^2 is negative is allowed.

With the shape parameter 2, the inverse gamma is an infinite variance distribution with mean equal to the scale parameter. This approach then does not tie down the possible values of s^2 and t^2+s^2/n too precisely, but it does specify an expected value for each.

Following Example 2.2, the posterior distributions are:

$$s^2|S \sim \text{Inverse Gamma in } p+N(n-1)S/2,\ 2+N(n-1)/2 \qquad (5.1a)$$

$$t^2+s^2/n|T \sim \text{Inverse Gamma in } q+(N-1)T/2,\ (N+3)/2 \qquad (5.1b)$$

Thus $E(s^2|S)=[2p+N(n-1)S]/[2+N(n-1)]$ and $E[(t^2+s^2/n)^{-1}|T]=(N+3)/(2q+(N-1)T)$, from the inverse gamma moment formulas. Also, the prior expected value of $(t^2+s^2/n)^{-1}$ is $2/q$, and so the prior expected value of $1-Z$ is $2p/nq$. $E(s^2|S)$ can be seen to be between the prior expectation p and the observation S, and much closer to the latter. Similarly, $E[(t^2+s^2/n)^{-1}|T]$ can be seen to

fall between the prior expectation 2/q and the observation 1/T, and is probably somewhat closer to the latter.

Setting the prior expected value of 1-Z to .5 gives q=4p/n, which is a way of picking q once p has been selected. Alternatively, p=nq/4 could be used to set p after q has been selected. Since S gets greater weight than T, the selected q probably has more bearing on the resulting Z than does p. As an example, suppose q=.2 is selected for Example 5.2. This is in the general area of $t^2+s^2/n=.1311$, but not particularly close. Since n is 6, p can be taken as .3. This gives posterior expected values of $E(s^2|S)=(.6+45S)/47$ and $E[(t^2+s^2/n)^{-1}]= 12/(.4+8T)$. For the three experiments, the following values are then generated, and the process is repeated for p=.6, q=.4:

Experiment	$E(s^2\|S)$	$E[(t^2+s^2/n)^{-1}\|T]$	1-Z
p=.3, q=.2			
1	.355	12.93 (=1/.077)	.765
2	.275	9.80 (=1/.102)	.449
3	.222	6.76 (=1/.148)	.250
p=.6, q=.4			
1	.367	9.04 (=1/.111)	.553
2	.288	7.39 (=1/.135)	.355
3	.235	5.51 (=1/.181)	.216

Either selection of priors seems to improve the estimation, but this test is somewhat unrepresentative, as the population Z is close to .5.

Diffuse priors could also be taken for s^2 and t^2+s^2/n, as in Example 2.3. With parameter p for this prior:

$$s^2|S \sim \text{Inverse Gamma: } N(n-1)S/2,\ -p-1+N(n-1)/2 \qquad (5.2a)$$

$$t^2+s^2/n|T \sim \text{Inverse Gamma: } (N-1)T/2,\ -p-1+(N-1)/2 \qquad (5.2b)$$

As discussed in Example 2.2, p=-1 makes the most sense for a diffuse prior. For comparison, p=-2 is also given below:

	$E[s^2\|S]$	$E[(t^2+s^2/n)^{-1}\|T]$	1-Z
p=-1	$N(n-1)S/(N(n-1)-2)$	$1/T$	$S/nT[1-2/N(n-1)]$
p=-2	S	$(N+1)/(N-1)T$	$S(N+1)/n(N-1)T$

Both are somewhat greater than S/nT. Note that if p=-2 for S and p=0 for T, 1-Z is the unbiased estimator (N-3)S/n(N-1)T. Neither of these estimates take into account the possible capping of 1-Z.

Regression Interpretation

Least squares credibility can be thought of as a least squares regression estimate in which the dependent variable has not yet been observed. The credibility estimate (4.12) can be rewritten as $\tilde{X}_{g0}-X_{..}=Z(X_{g.}-X_{..})$. Since the expected squared error is minimized by Z, this is similar to a no constant regression for $\tilde{X}_{g0}-X_{..}$, with $X_{g.}-X_{..}$ as the independent variable, where there is an observation for each risk g. The regression estimate of Z is computed by minimizing the sum of the actual square errors once X_{g0} is observed. A test of different methods of developing the credibility estimate then would be to compare Z to the regression estimate once the data is in.

Example 5.4

Efron and Morris (1975) computed the arcsin transforms of the batting averages for 18 players for their first 45 at bats in the 1970 season, as shown below, and used credibility methods to

estimate the similar figure for the rest of the season. The reason for the arcsin transform is that it results in an approximately normal distribution with $s^2=1$. Thus only t^2 need be estimated to get Z.

Player	First 45	Rest of Season
Alvarado	-3.26	-4.15
Alvis	-5.10	-4.32
Berry	-2.60	-3.17
Campaneris	-4.32	-2.98
Clemente	-1.35	-2.10
Howard	-1.97	-3.11
Johnstone	-2.28	-3.96
Kessinger	-2.92	-3.32
Munson	-4.70	-2.53
Petrocelli	-3.95	-3.30
Robinson	-1.66	-2.79
Rodriquez	-3.95	-3.89
Santo	-3.60	-3.23
Scott	-3.95	-2.71
Spencer	-2.60	-3.20
Swodoba	-3.60	-3.83
Unser	-3.95	-3.30
Williams	-3.95	-3.43

From the data, $X_{..}=-3.317$, and $T=1.115$. Since $n=1$, $S/nT=.897$, and as the $(N-3)/(N-1)$ factor is 15/17, an unbiased estimate of $1-Z$ is .791, or $Z=.209$. The regression estimate of Z is .186, which appears reasonably close. Relying on capping alone to correct S/nT would give $Z=.103$, which is not as close in this case. The inverse gamma prior for t^2+s^2/n with the prior Z of .5 gives $Z=.221$, which again is not as close as the factor approach.

Looking at just 3 batters at a time gives a different picture. *Without considering capping, the unbiased estimate would be* $Z=1$. For just 3 players, capping S/nT at 1 may in itself produce an unbiased estimate, however. Six different groups of 3 were

selected from the above table, namely first 3, second 3, etc. For each of these 6 cases, the capped regression estimate of Z is compared to the capped crediblity estimate and the Z from the inverse gamma prior.

Case:	1	2	3	4	5	6
T:	1.679	2.455	3.072	1.748	0.304	0.041
1-Cap S/nT:	.404	.593	.675	.428	0	0
Inv Gamma:	.359	.535	.576	.478	.303	.257
Regression:	.378	.199	0	.351	0	0

It should be noted that the diffuse prior with p=-1 gives the capped estimate in this example. There is not an unambiguous winner between these two estimators of Z; it is not even clear whether the goal should be the regression Z from the 3 points, or the estimate of .186 from the wider population. It is apparent, however, that the unbiased estimate which ignores capping, i.e., Z=1, is not as close as the others.

Section 6 - Incorporating Risk Size

Up to this point, the exposure was assumed to be the same by risk and over time. In many applications (e.g., territory or class ratemaking, commercial lines experience rating), this is not a viable assumption, and it is removed in this section. For instance, in experience rating the formulation Z=E/(E+K) is often used to assign credibility to risks of different sizes, where E is expected losses. Larger risks will receive greater credibility, as their pure premiums, loss ratios, etc. will have lower variances than for smaller risks. The E/(E+K) formula is based on a particular relationship between the variances of risks

of different sizes, namely $Var(X_{iu}|R_i)=s^2/E_{iu}$. That is, the variance is inversely proportional to risk size. With this assumption, it can be shown that $K=s^2/t^2$, where again t^2 is the variance of the hypothetical means.

It will be shown below that the inverse relationship of variance to exposure is a reasonable assumption, but that in fact it does not appear to hold in practice. A few other relationships will be explored to see which best accord with observation. Each of these will lead to different credibility formulas. In order to arrive at these formulas, a general formula will be developed that will hold for any relationship of variance to risk size, then the particular relationship desired can be just plugged in.

For the sake of concreteness, let X_{iu} be the pure premium for risk i time u, with L_{iu} the losses, and P_{iu} the exposure. By changing the definitions of P and/or L, X could just as easily be frequency, severity, loss ratio, etc. E.g., taking P_{iu} as the expected losses E_{iu} gives the experience rating credibility formula above.The general credibility formula is:

$$\tilde{X}_{g0} = (1-Z_g)m_g + Z_g X_{g.} \qquad (6.1a)$$

$$Z_g = P_{g.}/(P_{g.}+K_g) \qquad (6.1b)$$

$$K_g = P_{g.}/t_g^2 \Sigma_u s_{gu}^{-2} \qquad (6.1c)$$

$$Z_{gu} = (1-Z_g)t_g^2/s_{gu}^2 \qquad (6.1d)$$

Here $m_g=EX_{gu}$, $P_{g.}=\Sigma_u P_{gu}$, $X_{g.}=\Sigma_u Z_{gu}X_{gu}$, $t_g^2=VarE(X_{gu}|R)$, and $s_{gu}^2=EVar(X_{gu}|R)$.

To use this general formula, expressions are needed for m_g, t_g^2, and s_{gu}^2. These expressions will come from model assumptions, mainly assumptions about the relationship between variance and risk size.

Relationship of Variance to Risk Size

Since $X_{iu}=L_{iu}/P_{iu}$, the dependence of $Var(X_{iu}|R)$ on risk size will be approached by formulating the conditional variance of L_{iu} under different assumptions. This conditional variance can then be divided by P_{iu}^2 to yield $Var(X_{iu}|R)$.

L_{iu} is assumed to be the sum of the losses from P_{iu} exposure units. Let L_{iau} denote the losses from exposure unit a. If the exposure units are independent, then $Var(L_{iu}|R)=\sum_a Var(L_{iau}|R)$. If these units are conditionally identically distributed given R_i, $Var(L_{iau}|R)$ does not depend on a or u, and so can be denoted as $s(R_i)^2$. Then $Var(L_{iu}|R)=P_{iu}s_i(R)^2$. Thus $Var(X_{iu}|R)=s(R_i)^2/P_{iu}$. Letting $s^2=Es(R_i)^2$ gives $EVar(X_{iu}|R)=s^2/P_{iu}$. Hence, assuming that the risk is a collection of independent identically distributed exposure units yields that the expected conditional variance for a risk decreases in proportion to the exposure.

Hewitt (1967) showed that for a body of risks, the variance did not decrease this fast. The first two columns below derive from that paper.

Average Premium	Estimated Variance			
13,150/Prem	.172+ 9900/Prem	1837 Prem$^{.773}$	12,230+.133Prem 254+Prem	

296	26.3	44.4	33.6	22.7	22.3
628	12.3	20.9	15.9	12.7	14.0
869	10.4	15.1	11.6	9.80	11.0
1,223	7.58	10.7	8.27	7.58	8.39
1,924	5.35	6.83	5.32	5.32	5.73
3,481	3.07	3.78	3.02	3.36	3.40
6,050	2.18	2.17	1.81	2.19	2.07
8,652	1.59	1.52	1.32	1.66	1.50
12,265	1.15	1.07	.980	1.27	1.11
18,944	.749	.694	.695	.906	.769
33,455	.610	.393	.468	.585	.495
68,758	.345	.191	.316	.335	.310
220,786	.163	.060	.217	.136	.188

The variance in this case was not of the pure premium, but of the entry ratio, which is the loss ratio normalized to average to 1. The dollars are at 1958 levels. The other columns are fits of the variance by various functions of premium. The first of these functions specifies that the variance decreases by the inverse of premium. It can be seen that the actual variances are lower than this model would predict for small risks, and higher for large risks.

The difficulty for other functions of premium, however, is finding models that explain them. Such models would have to incorporate exposure units that are not conditionally independent given the risk parameter R_i.

One such model is provided by including the possibility that there are varying conditions that affect the risk, so that the loss probabilities are not the same in every year. For instance the risk parameters R_i could specify a distribution from which another parameter H_{iu} is determined each year. If the exposure

units are conditionally independent given H_{iu}, then given only R_i they are not independent; they have some correlation due to the common parameter H_{iu}. By the above reasoning, $Var(X_{iu}|H_{iu}) = s(H_{iu})^2/P_{iu}$. Then $Var(X_{iu}|R_i) = EVar(X_{iu}|H_{iu}) + VarE(X_{iu}|H_{iu})$, which can be written as $Var(X_{iu}|R_i) = s^2(R_i)/P_{iu} + y^2(R_i)$.

Thus with the inclusion of varying conditions, the conditional variance becomes a linear function of 1/P. The constant term essentially measures how much variance there is over time.

The second fit of the variance shown above uses this linear function. A much better fit to the risk variances is produced, although the smallest and largest risks still do not fit very well. It could be that the large risks are qualitatively different, and that linear functions could be used with different parameters for large and small risks. In a similar application of the linear model, Meyers and Schenker (====) do just that.
The final two columns represent (1) Hewitt's fit to this data based on $Var = s^2/P^C$, and (2) the function $Var = [y^2+s^2/P]/[1+C/P]$.

Neither of these is based on a model decomposing L_{iu} into exposure units, but improved fits are provided. The latter formula approaches a linear function of 1/P for large risks, but is below that line for the small risks. For all the curves, the parameters were selected to minimize squared errors in the log of

the variance, so that percentage errors in the variance would be as small as possible.

To review, then, four formulas relating conditional variance to risk size have been considered. The first two are based on models of the risk process, and the second two are just curves providing better fits. Since the conditional, or "process", variance of X_{gu} is a fuction of the exposure P_{gu}, then the expected value of this variance will be also. That is, $s_{gu}^2 = EVar(X_{gu}|R)$, the expected process variance for the gth risk at time u, is a function of P_{gu}. For the four curves these functions are as follows:

1. $s_{gu}^2 = s^2/P_{gu}$
2. $s_{gu}^2 = y^2 + s^2/P_{gu}$
3. $s_{gu}^2 = s^2/P_{gu}^{.773}$
4. $s_{gu}^2 = [y^2 + s^2/P_{gu}]/[1+C/P_{gu}]$

Each of these can be put into (6.1) to produce a credibility formula. This is done below, after two examples of negative binomial claim frequency distributions corresponding to the first two models.

Example 6.1

In this example, L_{iu} will be the number of claims, so that X_{iu} is claim frequency. The parameter R_i is the ordered pair $\langle V_i, Q_i \rangle$, and L_{iu} is assumed to be negative binomially distributed with

parameters $P_{iu}V_i$ and Q_i. (The sum of the claims for P_{iu} independent exposure units, each negative binomial in V_i, Q_i is itself negative binomial in $P_{iu}V_i$, Q_i.) These assumptions yield:

$$E(L_{iu}|R_i) = P_{iu}V_i(1-Q_i)/Q_i$$
$$E(X_{iu}|R_i) = V_i(1-Q_i)/Q_i$$
$$Var(L_{iu}|R_i) = P_{iu}V_i(1-Q_i)/Q_i^2$$
$$Var(X_{iu}|R_i) = V_i(1-Q_i)/P_{iu}Q_i^2$$

Thus the conditional variance of X_{iu} is inversely proportional to the exposure P_{iu}.

Example 6.2

The claims for each exposure unit are assumed to be Poisson with parameter H_{iu}, so that L_{iu} is Poisson in $Y_{iu}=P_{iu}H_{iu}$. H_{iu} is in turn gamma distributed in $R_i=<B_i,C_i>$, and so Y_{iu} is gamma in $P_{iu}B_i$, C_i. Thus from Example 2.1, L_{iu} is negative binomial in C_i, $1/(1+P_{iu}B_i)$. Thus:

$$E(L_{iu}|R_i) = P_{iu}B_iC_i$$
$$E(X_{iu}|R_i) = B_iC_i$$
$$Var(L_{iu}|R_i) = P_{iu}B_iC_i(1+P_{iu}B_i)$$
$$Var(X_{iu}|R_i) = B_iC_i/P_{iu} + B_i^2C_i$$

This is then an example of the second variance formula, a linear function of 1/P.

Credibility Formulas Varying By Risk Size

Once an expression relating the variance for different risk sizes has been selected, (6.1) can be used to produce a credibility

formula. If $s_{gu}^2 = s^2/P_{gu}$, as in the first model above, then $\sum_u s_{gu}^{-2} = P_{g.}/s^2$, and so $K_g = s^2/t^2$. Thus K_g is a constant, as in the constant exposure case, and $Z_g = P_{g.}/(P_{g.}+K)$.

For the other models, K_g is more complex. However, a fairly simple expression is possible in the case of just one observed time period. In the second model, $s_{gu}^{-2} = P_{gu}/(P_{gu}y^2+s^2)$ and $P_{gu} = P_{g.}$, so $K_g = (P_{g.}y^2+s^2)/t^2$, which can be written $K_g = P_{g.}A+B$, i.e., a linearly increasing function of the exposure. In this case $Z_g = P_{g.}/((1+A)P_{g.}+B)$.

If $s_{gu}^2 = s^2/P_{gu}^{.773}$, in the case of one exposure period, s^2 is given by $s_{gu}^{-2} = P_{gu}^{.773}/s^2$, so $K_g = P_{g.}^{.227}(s^2/t^2)$, or $K_g = BP_{g.}^{.227}$, again an increasing function of $P_{g.}$. The formula for Z becomes $Z = P_{g.}^{.773}/(P_{g.}^{.773}+B)$.

Finally, if $s_{gu}^2 = [y^2+s^2/P_{gu}]/[1+C/P_{gu}]$, and there is only one exposure period, $t_g^2 s_{gu}^{-2} = [1+C/P_{g.}]/[(y_g^2/t_g^2)+s_g^2/t_g^2 P_{gu}]$, so $K_g = [AP_{g.}+B]/[1+C/P_{g.}]$. With this, $Z = P_{g.}/(P_{g.}+K_g)$ yields, after some algebra, $Z_g = [P_{g.}+C]/[P_{g.}(1+A)+B+C]$. By redefining the constants, this can also be written as $Z_g = [P_{g.}+C]/[AP_{g.}+B]$. An interpretation of this formula based on heterogeneity of exposure units within a risk is given by Mahler (1987).

An important difference between (4.12) and (6.1a) is that the complement of credibility goes to $X_{..}$ in the former and in the

latter to $m=EX_{g0}=EE(X_{g0}|R)$. In (4.12) $X_{..}$ is also the minimum variance unbiased linear estimate of m.

In the unequal exposure case a weighted average of the X_{iu}'s can be used to estimate m. However the usual exposure weighted average is not optimal. At least for the simplest model $s_{gu}^2=s^2/P_{gu}$, it turns out that the minimum variance unbiased linear estimator of m, which will again be denoted $X_{..}$, is $X_{..}=\sum_i Z_i X_{i.}/Z_.$, where $Z_.=\sum_i Z_i$ (see ISO (1983)). This is sometimes referred to as the credibility weighted average of the $X_{i.}$'s. Standard statistical practice advocates weighting observations in inverse proportion to their variances. In this case $Var(X_{i.}) = t^2 + s^2/P_i = t^2/Z_i$, so the credibility is inversely proportional to the variance.

Estimation of Z

To estimate s^2, y^2, t^2, etc., extensions of the methods used in the equal exposure case can be used. First, the model with $s_{iu}^2=s^2/P_{iu}$ will be addressed. Let $S_i=\sum_{u=1}^{n} P_{iu}(X_{iu}-X_{i.})^2/(n-1)$, where here $X_{i.}=\sum_u P_{iu}X_{iu}/P_{i.}$, and let $S=\sum_{i=1}^{N} S_i/N$.

By repeated use of the formula $cov(X_{iu},X_{iv}|R_i)=\delta_{uv}s^2(R_i)/P_{iu}$, enough algebra (Appendix 2) will show that $E(S_i|R_i)= s^2(R_i)$. Thus $E(S_i)=s^2$, and $ES=s^2$ as well. S is a lower variance unbiased estimator of s^2 than is S_i.

Buhlmann and Straub(====) propose the following to estimate t^2. Let $W=\sum_{i,u}P_{iu}(X_{iu}-X)^2/(Nn-1)$, where X is the usual exposure weighted average of the X_{iu}'s. It can be shown that $EW=s^2+qt^2$, where $q=\sum_g P_{g.}(1-P_{g.}/P_{..})/(Nn-1)$. Thus (W-S)/q is an unbiased estimator of t^2. As they point out, this can sometimes be negative, in which case they assign $t^2=0$, and so Z=0.

Klugman (1985) gives an alternative approach, which appears to be more accurate. Let $T=\sum_{i=1}^{N}Z_i(X_{i.}-X_{..})^2/(N-1)$. In Appendix 2 it is shown that, given the Z_i, $ET=t^2$. T cannot be considered an estimator of t^2, because t^2 is needed to compute Z_i in the formulas for $X_{..}$ and T. However if Z_i is initially set to 1, an iterative procedure can be used to compute $X_{..}$ and T, estimate t^2, compute new Z_i's, etc., until the estimate for t^2 stabilizes (usually quickly). DeVylder (1981) uses the term pseudo-estimator for such a T, and suggests another one.

Klugman (1986) details several Bayesian approaches, and shows that these can give dramatically improved credibilities. One of these generalizes the diffuse priors used in (5.2), by specifying that the joint prior density of s^2 and t^2 is proportional to $s^{-2}[\Pi_i(s^2+P_{i.}t^2)]^{-1/N}$. This particular prior is taken after Box and Tiao (1973, p. 426). Introducing the variable $r=t^2/s^2$, and defining $w_i=rP_{i.}/(1+rP_{i.})$ and $w=\sum w_i$, the posterior distribution for r given the observations X_{iu} turns out to be proportional to:

$$f_1(r) = [r\sum_{i,u}P_{iu}(X_{iu}-X_{i.})^2 + \sum_i w_i(X_{i.}-X_{..})^2]^{-(N-1)/2} \times$$
$$\prod_i[(1+P_{i.}r)^{-1/N}(w_i/w)^{.5}].$$

This must be integrated numerically from zero to infinity to find the constant of proportionality. Dividing $f_1(r)$ by this constant gives the conditional density $f(r)$. Then $E(r|\text{the } X_{iu}\text{'s}) = \int rf(r)dr$, which again is done numerically. This gives an estimate for r, and K can then be estimated by $1/r$. Alternatively, $E(1/r)$ could be calculated directly by numerical integration.

To estimate K for the model $s_{gu}^2=y^2+s^2/P_{gu}$, some algebra will show $E(S_i)= s^2+y^2(P_{i.}^2-\sum_{u=1}^{n}P_{iu}^2)/(n-1)P_{i.}$. Thus if a linear regression is done for S_i against $(P_{i.}^2-\sum_u P_{iu}^2)/(n-1)P_{i.}$, the slope and intercept can be used as estimators of s^2 and y^2.

Estimation of t^2 in this case could perhaps be done as follows. Let $w_i=P_{i.}t^2/(P_{i.}t^2+s^2+y^2\sum_u P_{iu}^2/P_{i.})$ and $w=\sum_i w_i$. Define $X_{..}=\sum_i w_i X_{i.}/w$ and $T=\sum_{i=1}^{N} w_i(X_{i.}-X_{..})^2/(N-1)$, where again $X_{i.}$ is the exposure weighted average of the X_{iu}'s. In this case it can be shown that $ET=t^2$, and so T is an unbiased pseudo-estimator of t^2. However, for this and the more complex models, the Bailey-Simon method is often used instead, as discussed below.

Bailey and Simon (1959) presented the idea of estimating Z by seeing which values of Z would have worked best in the past. In their example, each risk had one unit of exposure, namely a single private passenger car. For the risks with no claims, the credibility estimate of X_{g0} is just $(1-Z)X_{..}$. This can be compared to the average experience for these risks in the next year to see what Z should have been. Since only a fixed number of years (usually 1 to 5) are used in automobile experience rating, this value of Z can then be used in the future.

Meyers (1985) uses a similar retrospective approach to estimate A and B in $Z_g=P_{g.}/(AP_{g.}+B)$ in commercial insurance experience rating. Rather than focusing on the zero loss risks, Meyers creates a test statistic for the overall performance of the plan, and optimizes the test statistic. NCCI adopted a similar procedure with a different test statistic to estimate A, B, and C in $Z_g=(P_{g.}+C)/(AP_{g.}+B)$ for workers compensation experience rating.

Section 7 - How Good Is Least Squares Credibility

As discussed earlier, the function of the X_{iu}'s that optimizes the expected squared error in X_{g0} is the conditional expectation $E(X_{g0}|\text{the } X_{iu}\text{'s})$. The best linear function in this sense is the least squares credibility estimate.