ON THE CALCULATION OF VARIANCES AND CREDIBILITIES BY EXPERIENCE RATING

K. LOIMARANTA
Finland

1. INTRODUCTION

By experience rating the main problem is to estimate the credibilities. We have for the credibility α_k the famous formula *)

$$\alpha_k = \frac{\sigma_0^2}{\sigma_0^2 + \sigma_k^2}$$

but it is often troublesome to find suitable estimates for the variances σ_0^2 and σ_k^2. In the present paper a general method to estimate them from the actual statistics is given.

A disadvantage of the method is that good estimates require relatively extensive statistical material. If one of the variances is known, the method can be easily modified to give the other variance from statistics of moderate size.

The method is based on the Maximum Likelihood principle and leads to a system of non-linear equations. The equations can be solved by an iterative process, easily programmable for computers.

The mathematical model underlying the experience rating problem differs in our case lightly from the usual one.

2. FORMULATION OF THE PROBLEM

We consider a portfolio, which is divided into N classes. In each class we have observed a claim amount per risk unit. Our assumption is that the relative claim amount y_k in the class k ($k = 1, 2, \ldots, N$) has a definite but unknown meanvalue m_k and a variance σ_k^2, which is inverse proportional to some known measure t_k of the size of that class, e.g. the number of risk units in the class. We can thus write

$$\text{Mean:} \quad E(y_k) = m_k$$
$$\text{Variance:} \quad V(y_k) = \sigma_k^2 = h/t_k.$$

* Bühlman, H.: Mathematical Methods in Risk Theory.

https://doi.org/10.1017/S0515036100011508 Published online by Cambridge University Press

As a second step we assume, that the quantities m_k are random variables with a common probability distribution. Let this distribution have the meanvalue m_0 and the variance σ^2:

$$\text{Mean:} \quad E(m_k) = m_0$$
$$\text{Variance:} \quad V(m_k) = \sigma_0^2.$$

Assuming that both steps are independent on each other (or at least uncorrelated) we have for the *compound* random variable y_k

$$\text{Mean:} \quad E_c(y_k) = m_0$$
$$\text{Variance:} \quad V_c(y_k) = \sigma_0^2 + h/t_k.$$

The result might be better known from the theory of compound Poisson processes.

To calculate the credibilities

$$\begin{aligned} \alpha_k &= \frac{\sigma_0^2}{\sigma_0^2 + \sigma_k^2} = \frac{\sigma_0^2}{\sigma_0^2 + h/t_k} \\ 1 - \alpha_k &= \frac{\sigma_k^2}{\sigma_0^2 + \sigma_k^2} = \frac{h/t_k}{\sigma_0^2 + h/t_k} \end{aligned} \qquad \text{(A)}$$

we must have estimates for the variance σ_0^2 and the constant h, which determines the variances σ_k^2.

3. The Maximum Likelihood Solution

We suppose from now on, that the random variables y_k are with required accuracy normally distributed, i.e. y_k has the distribution function

$$f(y_k) = \frac{1}{\sqrt{2\pi\,(\sigma_0^2 + h/t_k)}}\, e^{-\frac{(y_k - m_0)^2}{2(\sigma_0^2 + h/t_k)}}$$

We use the Maximum Likelihood method *) to estimate the parameters m_0, σ_0^2 and h in the distribution function of y_k.

For the logarithm of the Likelihood function L we have the expression

$$\log L = -\sum_k \left[\frac{(y_k - m_0)^2}{2(\sigma_0^2 + h/t_k)} + \tfrac{1}{2}\log(\sigma_0^2 + h/t_k)\right] + \text{const.}$$

* E g. Cramér, H., Mathematical Methods of Statistics.

https://doi.org/10.1017/S0515036100011508 Published online by Cambridge University Press

Its maximum value is a solution of the equations

$$\frac{\partial \log L}{\partial m_0} = \sum_k \frac{y_k - m_0}{\sigma_0^2 + h/t_k} = 0$$

$$\frac{\partial \log L}{\partial (\sigma_0^2)} = \sum_k \left[\frac{(y_k - m_0)^2}{2(\sigma_0^2 + h/t_k)^2} - \frac{1}{2(\sigma_0^2 + h/t_k)}\right] = 0$$

$$\frac{\partial \log L}{\partial h} = \sum_k \left[\frac{(y_k - m_0)^2}{2(\sigma_0^2 + h/t_k)} \cdot \frac{1}{t_k} - \frac{1}{2(\sigma_0^2 + h/t_k)} \cdot \frac{1}{t_k}\right] = 0$$

Multiplying the first equation by σ_0^2, the second by σ_0^4 and the last one by h^2 and observing the expression (A) for α_k we get the following equations

$$\begin{aligned} &1) \quad \sum_k \alpha_k y_k = m_0 \sum_k \alpha_k \\ &2) \quad \sum_k \alpha_k (y_k - m_0)^2 = \sigma_0^2 \sum_k \alpha_k \qquad \text{(B)} \\ &3) \quad \sum t_k (1 - \alpha_k)^2 (y_k - m_0)^2 = h \cdot \sum (1 - \alpha_k). \end{aligned}$$

From the equations (B) and the expression (A) for α_k the quantities m_0, σ_0^2 and h as well as the credibilities α_k can be calculated by an iterative process. We start with arbitrary values for the quantities α_k (e.g. $\alpha_k = 1/2$) and calculate m_0, σ_0^2 and h from eq. (B). New values for α_k will then be calculated from eq. (A) with the received values of σ_0^2 and h. Subsequently the new values of α_k will be inserted in eq. (B), and so on.

According to our experience about ten steps are required to get the values of α_k with an accuracy of 0.001. The method is cumbersome for manual calculation but suits well for electronic computers.

When the credibilities α_k are determined the premiums net of charges for different classes can be calculated by the normal way

$$P_k = \alpha_k y_k + (1 - \alpha_k) m_0.$$

4. A Possible Generalisation

The method can be generalized to solve more complicated problems. So far we have assumed the quantitities m_k to be drawn from one and the same probability distribution. But we can also think them to be results of a regression analysis

$$m_k = a + \sum_i b_i x_{ik},$$

https://doi.org/10.1017/S0515036100011508 Published online by Cambridge University Press

where x_{ik} is the value of the i:th independent variable in class k. In stead of the simple weighted mean

$$m_0 = \frac{\sum\limits_k \alpha_k y_k}{\sum\limits_k \alpha_k}$$

we have for each step to solve a by α_k weighted regression analysis problem and to put in the equations 2) and 3) the residuals in stead of the quantities $m_k - m_0$.

5. Discussion of the Method

The maximum likelihood method is normally used for observations y_k with equal distributions and gives then under general assumptions asymptotically optimal estimates. The restriction to equal distributions is unessential but nevertheless we have to be careful. An other point to be observed is that we have assumed the quantities y_k to be normally distributed.

The first eq. (B) is same as Hovinen *) has got with a different method in the case σ_0^2 and h are given. The equation gives an unbiased minimum variance estimate for the mean independent of the normality of the quantities y_k. Conserning this equation we are thus on the safe side.

It is interesting to observe that in the credibility theory an other formula is in general used to calculate the gross mean

$$m_0' = \frac{\sum t_k y_k}{\sum t_k} \qquad (1')$$

The first formula in (B) gives the correct estimate for the mean if we choose one class at random, the formula (1') if we choose one risk unit (policy) at random. The differences between m_0 and m_0' can be considerable.

More caution is required by the use of the eq. 2) and 3) in (B). It is not enough that the number of observations (classes) N is sufficiently large. If all quantities t_k are identical the eq. 2) and 3) are linearly dependent and all values α_k = const. are a solution of

* Hovinen, E., On the Estimation of Means and Variances in the Case of Unequal Components, ASTIN Bulletin, Vol. VIII, Part 3.

the equationsystem (A) and (B). An acceptable splitting of the variance

$$\sigma_0^2 + h/t_k$$

into its components requires thus that the variation of t_k is great enough. Is this not the case but one of the components is known, the method can be used to calculate the remaining component and the credibilities α_k simply by omitting the corresponding equation in (B).

It is well-known that Maximum Likelihood method gives for finite samples too low values for variances. The bias is normally of the order $1/N$. An unbiased expression for the variance in the case $\sigma_0^2 = 0$, i.e. $\alpha_k = 0$ is given by Hovinen (ibid.). His formula (26) is by our notations

$$h = \frac{1}{N} \sum_k \frac{t_k}{1 - \frac{t_k}{\Sigma t_k}} (y_k - m_0)^2$$

Our eq. 3) in (B) gives with $\alpha_k = 0$

$$h = 1/N \sum_k t_k (y_k - m_0)^2.$$

Comparing these two formulas we see, that the bias is negligible if all t_k:s are small compared with $\sum_k t_k$.