A DISTRIBUTION FREE METHOD FOR GENERAL RISK PROBLEMS

H. BÜHLMANN
Zurich, Switzerland

1. INTRODUCTION

In practical applications of the collective theory of risk one is very often confronted with the problem of making some kind of assumptions about the form of the distribution functions underlying the frequency as well as the severity of claims. Lundberg's [6] and Cramér's [3] approach are essentially based upon the hypothesis that the number of claims occurring in a certain period obey the Poisson distribution whereas for the conditional distribution of the amount claimed upon occurrence of such a claim the exponential distribution is very often used. Of course, by weighting the Poisson distributions (as e.g. done by Ammeter [1]) one enlarges the class of "frequency of claims" distributions considerably but nevertheless there remains an uneasy feeling about artificial assumptions, which are just made for mathematical convenience but are not necessarily related to the practical problems to which the theory of risk is applied.

It seems to me that, before applying the general model of the theory of risk, one should always ask the question: "How much information do we want from the mathematical model which describes the risk process?" The answer will be that in many practical cases it is sufficient to determine the mean and the variance of this process. Let me only mention the rate making, the experience control, the refund problems and the detection of secular trends in a certain risk category. In all these cases the practical solutions seem to be sufficiently determined by mean and variance.

Let us therefore attack the problem of determining mean and variance of the risk process while trying to make as few assumptions as possible about the type of the underlying probability distributions. This approach is not original. De Finetti [5] has already proposed an approach to risk theory only based upon the know-

ledge of mean and variance. It is along his lines of thought, although in different mathematical form, that I wish to proceed.

2. THE DISTRIBUTION OF CLAIMS

Even if our method shall not depend on the particular distribution functions characterizing the risk process we still have to refer to these functions. "Distribution free" means then, of course, that we are not at liberty to make assumptions as to the type to which the functions in question should belong.

With this in mind we define two random variables

a) X standing for the frequency of claims and computed as

$$X\ (N) = \frac{k}{N}$$

where k = number of claims in a given period

N = number of risks exposed in the same period.

b) Y standing for the average amount per claim occurred or

$$Y\ (k) = \frac{S}{k}$$

where S = total sum of claims paid in a certain period

k = as under a).

What assumptions—it is of course necessary to make some—are feasible for the distribution functions of X and Y?

I should like to discuss the following two hypotheses:

I. *Homogeneity*: Any k risks chosen at random from the total number of N risks have identical occurrence and claim amount distributions; in mathematical language:

k is the sum of exchangeable 0—1 random variables

$F_{Y\,(k)}\ (y)$ depends only on the number k but not on the choice of the k risks which have produced claims;

II. *Independence*: The occurrence of any one particular claim does not influence the occurrence of any other claim and any claim amount has no bearing on the amount of later claims occurring;

in mathematical language:

k is a sum of N independent 0—1 trials

S is the sum of k independent random variables.

3. DISCUSSION OF THE HYPOTHESES

Hypothesis I: Any mathematical model of practical importance seems impossible without this hypothesis. Even if in theory one might try to evade the assumption of homogeneity by means of conditional distributions this amounts practically speaking only to subdividing the total collective of risks into subcollectives where the postulate of homogeneity holds.

But is hypothesis I justified from a practical point of view? Yes, if we bear the following considerations in mind:

a) If we apply collective risk methods to any collective which has been judged to be homogeneous by standard underwriting practices, we have no reason to doubt hypothesis I.

b) If our collective under consideration is heterogeneous by underwriting standards there is still one way of reasoning which allows us to work with hypothesis I. This is the fact that as k (the number of claims) becomes large we can count on obtaining a fairly representative sample of risks within which the members of the heterogeneous population occur almost in the "correct" proportion.

Hypothesis II: In many cases the assumption of independent individual risks does certainly not hold. Nevertheless collective risk models are mostly constructed under this assumption. (In fact the independence hypothesis is so common to be made that many authors forget to mention it). But again are there other reasons than those of mathematical convenience which can be quoted in support of such a working hypothesis? What are they?

a) First of all, if we speak of independent individual risks, we should say what we mean by individual risks. Independence can very often be achieved by considering groups of insured objects or of individual people as one individual risk. Unfortunately in defining individual risks this way we certainly are bound to in-

crease the degree of heterogeneity within a given collective. One might therefore argue that there is no sense in improving on the applicability of one basic hypothesis if another equally important one tends to lose its justification. Nevertheless I am convinced that in many practical circumstances there is a middle road between Scylla and Charybdis.

b) In addition there is the fact that individual risks are indeed often independent for a great proportion of the whole collective. If this proportion is overwhelming the influence of interdependent risks can be neglected. As justified later, the influence of dependence leads to overestimation of the variance for smaller samples and to underestimation for larger ones. For many practical purposes this seems to put the insurer rather on the safe side.

Summarizing this discussion it seems to me that hypothesis I can very often be accepted. Hypothesis II can well be justified as a first approximation. However, one might wish to adjust results obtained on account of Hypothesis II by the use of correlation coefficients.

4. CONSEQUENCES OF THE TWO HYPOTHESES

From Hypothesis I:

a) A very immediate consequence is the mathematical fact that the expected value of X and the conditional expectation of Y given k do not depend on the parameters N and k

in mathematical form

$$E\ [X\ (N)] = \mu \text{ (independent of } N)$$

$$E\ [Y(k)/_k] = \nu \text{ (independent of } k)$$

where $E\ [./k]$ = conditional expectation given k

b) A more elaborate discussion of this hypothesis can be mathematically phrased as follows:

$X\ (N)$ is a sum of exchangeable random variables

$Y\ (k)$ is, for any given k, a sum of exchangeable random variables

Referring the reader to papers by De Finetti [4] and the author [2] on the theory of exchangeable random variables, I should

merely like to point out one consequence of exchangeability which seems important within the scope of this paper:

All collectives of exchangeable random variables which can be enlarged without loss of the exchangeability property, have necessarily positive correlation coefficients.

In practice this means, provided a homogeneous collective of risks can always be imbedded within a bigger still homogeneous collective, correlations between individual risks will be positive. Or, more intuitively, if there is interdependence among a homogeneous collective of risks, the following holds

the occurrence of a claim will never decrease the chance for the occurrence of any other claim but could have the opposite effect.

the fact, that a high claim amount has occurred will never decrease the chances of a high claim amount for any other risk but may rather increase it.

From Hypotheses I and II

a) If in addition to hypothesis I we also accept II we find that the variance of X (N) and the conditional variance of Y (k) are linearly decreasing functions of N and k respectively

in mathematical form

$$\sigma^2[X\ (N)] = \sigma^2/N \quad \sigma^2 = \text{constant}$$
$$\sigma^2[Y\ (k)/_k] = \tau^2/k \quad \tau^2 = \text{constant}$$

where $\sigma^2[./k] =$ conditional variance given k.

b) Our discussion of hypothesis I indicates that without hypothesis II we would most likely expect the above two variances to decrease *more slowly* than in the case of independence.

Again our intuition may help us in understanding this mathematical formalism if we bear in mind, that a) indicates how big a sample we have to take to apply the law of larger numbers. The discussion under b) would then tell us that in the case of interdependent risks we would need a bigger sample to achieve the same averaging effect of large numbers.

5. THE MATHEMATICAL MODEL

Based upon our definitions, hypotheses and considerations as

made in the previous paragraphs the following formal set up is indicated

$X\ (N)$ = random variable representing the frequency of claims

$Y\ \ (k)$ = random variable representing the average amount per claim occurred.

Hypotheses I and II lead us to postulate

a) $X\ (N)$ has mean μ (independent of N)
variance σ^2/N (σ^2 = constant)

b) given k $Y(k)$ has mean ν (independent of k)
variance τ^2/k (τ^2 = constant)

6. MEAN AND VARIANCE OF THE RISK PROCESS

Considering that we are interested in the totality of claims originating from the risk process our attention will be focused on the random variable

$$Z = X.Y = \frac{\text{number of claims}}{\text{number of risks}} \cdot \frac{\text{total of claims}}{\text{number of claims}} = \frac{\text{total of claims}}{\text{number of risks}}$$

Our goal is to find $E\ (Z)$ and $\sigma^2\ (Z)$

a) $$E\ (Z) = E\ (X.Y) = E\ [X.E\ (Y/k)]$$
$$= E\ [X.\nu] = \mu.\nu$$

Hence

(1) $$E\ (Z) = \mu.\nu$$

b) $$\sigma^2\ (Z) = E\ (Z^2) - E^2\ (Z)$$
$$E\ (Z^2) = E\ [X^2 E\ (Y^2/X)]$$
$$= E\left[\frac{k^2}{N^2}\, E(Y^2/k)\right]$$
$$= E\left[\frac{k^2}{N^2}\left(\nu^2 + \frac{\tau^2}{k}\right)\right]$$
$$= \nu^2\left(\mu^2 + \frac{\sigma^2}{N}\right) + \frac{1}{N}\,\tau^2\mu$$
$$= \nu^2\,\mu^2 + \frac{\nu^2\sigma^2 + \mu\tau^2}{N}$$

Hence

$$\sigma^2(Z) = \frac{\nu^2\sigma^2 + \mu\tau^2}{N} \tag{2}$$

Formulae (1) and (2) allow us to determine completely the mean and the variance of the totality of claims. All that remains to be done is to estimate the four parameters μ, σ, ν, τ from the actual observations.

7. ESTIMATION OF THE PARAMETERS

Let us assume that we have the following data available for m observation periods

k_i = number of claims occurred in period i;

N_i = number of risks exposed in period i;

S_i = totality of claim amounts paid in period i;

From these data we derive for each i;

$$X_i = \frac{k_i}{N_i} \text{ mean } \mu; \quad \text{variance } \sigma^2/N_i$$

$$Y_i = \frac{S_i}{k_i} \text{ mean } \nu; \text{ variance } \tau^2/k_i$$

By replacing the sample mean and variance for the population mean and variance respectively the following estimates are obtained

$$\hat{\mu} = \frac{\Sigma X_i N_i}{\Sigma N_i} = \frac{\Sigma k_i}{\Sigma N_i} \tag{3}$$

$$\hat{\sigma}^2 = \frac{\Sigma (X_i - \hat{\mu})^2 N_i}{m} \tag{4}$$

$$\hat{\nu} = \frac{\Sigma (Y_i k_i)}{\Sigma k_i} = \frac{\Sigma S_i}{\Sigma k_i} \tag{5}$$

$$\hat{\tau}^2 = \frac{\Sigma (Y_i - \hat{\nu})^2 k_i}{m} \tag{6}$$

It is interesting to observe that the applied estimation principle does not use any assumption about the underlying probability distributions. However, should they be of the normal type then $\hat{\mu}$, $\hat{\sigma}^2$, $\hat{\nu}$, $\hat{\tau}^2$ are exactly the maximum likelihood estimates.

8. ONE FINAL REMARK

From the estimates for μ, σ^2, ν, τ^2 it is easy, using formulae (1) and (2), to get estimates for $E(Z)$ and $\sigma^2(Z)$.

However one question may cross our minds at this stage: "Why do we have to proceed in such a complicated fashion in order to obtain these estimates?. Would it not be possible to get reasonable estimates directly by the use of the observed values of Z?"

$$Z_i = \frac{\text{Totol claim amounts}}{\text{Total number of risks}}$$

The answer to this question is definitely in the negative. Let me illustrate this by an example; (working with hypothetical figures for X_i and Y_i where for simplicity N is taken constant).

i	X	Y	Z
1	.10	20	2
2	.05	40	2
3	.06	33·333	2
4	.04	50	2
5	.02	100	2

The estimation method proposed in paragraph 7 yields the following results:

$$\hat{\mu} = .054$$
$$\hat{\sigma}^2 = 7.04 \cdot 10^4 \cdot N$$
$$\hat{\nu} = 37.037$$
$$\hat{\tau}^2 = 23.26 \cdot N$$

Hence:

$$\hat{E}(Z) = \hat{\mu} \cdot \hat{\nu} = 2$$
$$\hat{\sigma}^2(Z) = \frac{\hat{\nu}^2\hat{\sigma}^2 + \hat{\mu}\hat{\tau}^2}{N} = 2.22$$
$$\hat{\sigma}(Z) = 1.49$$

We observe that our method gives us a very substantial standard deviation.

On the other hand estimating $E(Z)$ and $\sigma^2(Z)$ from the observed Z_i we obviously find

$$\hat{\hat{E}}(Z) = 2$$
$$\hat{\hat{G}}(Z) = 0$$

This is certainly unreasonable since it suggests that our risk process is deterministic and lacks any random element. The method proposed in paragraph 7, however, avoids such wrong conclusions since it analyses the components of the risk process more carefully.

BIBLIOGRAPHY

[1] AMMETER, H.: Die Elemente der kollektiven Risikotheorie von festen und zufallsartig schwankenden Grundwahrscheinlichkeiten. Mitt. Schw. Vers. Math. 1949.
[2] BÜHLMANN, H.: Austauschbare stochastische Variabeln und ihre Grenzwertsätze. Univ. Cal. Publ. Statist. No. 1, (1960).
[3] CRAMÉR, H.: Collective Risk Theory. Jubilee Volume of Skandia Insurance Company 1955.
[4] DE FINETTI, B.: Classi di numeri aleatori equivalenti. Rend. Lincei Vol. XVIII, 1933.
[5] DE FINETTI, B.: Il problema dei pieni. Giorn. Inst. It. Att. 1940.
[6] LUNDBERG, F.: Über die Wahrscheinlichkeitsfunktion einer Risikenmasse. Skand. Ak. 1930.