DISTRIBUTION FREE APPROXIMATIONS IN APPLIED RISK THEORY

GUNNAR ANDREASSON
Stockholm

Most large insurance companies have today electronic computers that enable not only efficient actuarial statistics, but also research in applied risk theory. An important task in this latter field is the developing of an information system for control of the business as to the statistical balance between premiums and claims. The entire system can be separated into two parts, one descriptive and one analytic part. The descriptive part, that also may be called statistics production, is the base of the whole system and must be constructed in a general way to make it possible to apply mathematical tools in risk analysis. For the analytic part and its applications for computers there is a growing interest among actuaries as can be noticed from the reports in actuarial journals. The classical models of collective risk theory have recently been extensively illustrated by numerical calculations performed by the convolution committee in Sweden.

When starting to construct the analytical part of the information system one finds that in spite of programming for the computer there is firstly a hard work to find realistic mathematical models, especially mathematical expressions for claim distributions, and secondly to estimate their parameters. According to the mathematical theory of risk, or the risk process, it is necessary to assume and test specific mathematical forms of two distributions, the number of claims and claims amount. When dealing with practical problems in motor car insurance the Polya-process

$$P_x(t) = \binom{x+h-1}{x} \left(\frac{h}{t+h}\right)^h \left(\frac{t}{t+h}\right)^x \qquad (1)$$

seems to give a very good approximation of the number of claims x for one policy during the time interval $(0, t)$ measured on the

operational scale. If we assume stochastic independence among the N policies in a risk group, and define a random variable n as the sum of all claims in the group

$$n = x_1 + x_2 + \ldots\ldots + x_N \tag{2}$$

we can get the probability distribution of n, as the N:th convolution of the distribution (1). The Polya process has been investigated by O. Lundberg, H. Ammeter and others. Here shall be mentioned a basic feature of the process that has great importance for the applications.

The characteristic function of the distribution (1) is

$$\varphi_{x_i}(u) = \left[1 - \frac{t}{h}(e^{iu} - 1)\right]^{-h} \tag{3}$$

and the characteristic function of its N:th convolution

$$\varphi_n(u) = \left[1 - \frac{t}{h}(e^{iu} - 1)\right]^{-Nh} = \left[1 - \frac{tN}{hN}(e^{iu} - 1)\right]^{-Nh} \tag{4}$$

This implies that the total number of claims n has the following distribution

$$P_n(Nt) = \binom{n + Nh - 1}{n} \left(\frac{Nh}{Nt + Nh}\right)^{Nh} \left(\frac{Nt}{Nt + Nh}\right)^{n} \tag{5}$$

Applications to risk problems show that in the Polya case mathematical convenience and practical usefulness coincide.

When the question comes to accumulated claims amount and risk premiums the problems are harder to solve. The classical risk theory starts with the distribution of the accumulated claims amount as a sum of weighted convolutions. Thus the accumulated claims amount

$$Y = Y_1 + Y_2 + \ldots\ldots + Y_n \tag{6}$$

has the following distribution

$$F(y) = \sum_n p_n V^{n*}(y) \tag{7}$$

where $V(y)$ is the distribution function of any Y_i.

We obtain well-known models if we chose as weights probabilities from the Poisson or negative binomial distributions, or in more

general terms their corresponding stochastic processes. It can be proved (Ammeter 1948) that under the assumptions made above in the Polya case the normalized variable

$$z = \frac{Y - E\,Y}{(\text{Var } Y)^{1/2}} \tag{8}$$

has a distribution that for large N can be represented by the first two terms in the Charlier expansion

$$F(z) = \Phi\,(z) - \frac{\gamma \Phi^{\text{III}}(z)}{3!} \tag{9}$$

Where γ is the coefficient of skewness

$$\gamma = \frac{\mu_3}{(\text{Var } Y)^{3/2}} \tag{10}$$

and μ_3 is the third central moment of the variable Y. It seams very reasonable that this expansion should be valid even if the moments appearing in the relations (8) and (9) are not exactly from a Polya (or Poisson) population in the general convolution (7). If N is fairly large we can take the empirical moments of the risk group and in the case of small groups, we can use estimates from similar larger groups. The main problem in this way of approaching the distribution of classical risk theory is to estimate the moments of the variable Y. This can be done in an indirect way without any assumptions on the mathematical form of the distributions p_n and $dV(y)$ as will be shown below. The procedures are well adapted for programming and thus integration in an acturial control system applied for a computer.

The formulas for this distribution free procedure are deduced for the risk premium R defined by

$$R = \frac{Y}{N} \tag{11}$$

From this relation we get immediately

$$\begin{aligned} E\,R &= \frac{1}{N}\,E\,Y \\ \text{Var } R &= \frac{1}{N^2}\,\text{Var } Y \end{aligned} \tag{12}$$

To estimate the moments of R we proceed in an indirect way. From the implicit moment relations that can be expressed by the series expansion of the characteristic function of the variable R, or easier the variable Y, we get the corresponding expansion of the variables n and Y in a functional relation that can be solved successively in respect to the moments of R, after taking in this case the first, second and third derivative of the characteristic function

$$\varphi_Y(u) = E\, e^{iuY} = \int_0^\infty e^{iuY} \Sigma p_n \, dV^{n*}(y) = \Sigma\, p_n [\varphi_{Y_i}(u)]^n. \qquad (13)$$

and setting the variable u equal to zero in each step. From the assumption of independence among the variables x_i in the sum (2) we get

$$\text{Var } n = N \text{ Var } x_i \qquad (14)$$

Let us denote the moments of the variable x_i by α_k and the moments of Y_i by β_k. For the variable n the following relations can be deduced

$$\begin{aligned} \Sigma n p_n &= En = N\alpha_1 \\ \Sigma n^2 p_n &= En^2 = N\alpha_2 + N(N-1)\alpha_1^2 \\ \Sigma n^3 p_n &= En^3 = N\alpha_3 - 3N(N-1)\alpha_1\alpha_2 + N(N-1)(N-2)\alpha_1^3 \end{aligned} \qquad (15)$$

Using the standards formulas for the central moments μ_k

$$\begin{aligned} \mu_2 &= E\, R^2 - E^2 R \\ \mu_3 &= E\, R^3 - 3\, E\, R^2 \,.\, E\, R + 2\, E^3 R \end{aligned} \qquad (16)$$

and the relations (16) and (17) we obtain after simple reductions

$$\begin{aligned} E\, R &= \alpha_1 \beta_1 \\ \mu_2 = \text{Var } R &= \frac{1}{N}[(\beta_2 - \beta_1^2)\alpha_1 + \beta_1^2(\alpha_2 - \alpha_1^2)] \\ \mu_3 &= \frac{1}{N_3}[(\beta_3 - 3\beta_1\beta_2 + 2\beta_1)\, En + (3\beta_1\beta_2 - 3\beta_1^3)\, En^2 + \beta_1^3\, En^3 \\ &\quad - (3\beta_1\beta_2 - 3\beta_1^3)\, E^2 n - \beta_1^3\, En\, En^2 + 2\beta_1^3\, E^3 n] \end{aligned} \qquad (17)$$

The variance above was computed by Bühlmann 1964 from another point of view. We now have the mean, variance and third central moment of the risk premium expressed in the moments of the distributions $P_i(x)$ and $dV(y)$, and we can estimate these latter moments from the material simply by the corresponding

empirical moments. If the investigation concerns a small risk group, the moments β_k can be estimated from a similar larger group to avoid the uncertainty in computing moments of a very skew distribution from a small material. The moments α_k generally become meaningful already in a riskgroup of 1,000 policies observed during a calendar year. This can be theoretically verified if the Polya distributions (1) and (5) are assumed, since the hypothesis of asymptotic normality here holds, and can be used to get confidence intervals of functions of moments. Here can be referred to Ammeter 1948 and Cramér 1945.

For the tabulation of (9) it is necessary to use a computer, if several risk groups are investigated. The procedures are then easy to handle especially if a modern problem-oriented language is used. Here will be given an example of the distribution free method compared to a standard mathematical model of the general convolution (7), in which p_n is assumed to have the form (5) and an exponential polynomial as approximation for the distribution $dV(y)$ of individual claims.

$$dV(y) = \Sigma A_i B_i e^{-B_i y} \tag{18}$$

The numerical evaluation of the distribution of R in the mathematical model is done by the Esscher method for the Polya case. None of these details will be given here. The reader is referred to Ammeter 1948 and to Esscher-Bohman 1963. The Esscher method has also been systematized by the present author in a general computer program, and can thus be used for large scale investigations. The statistical material given below is drawn from Swedish third party motor car insurance.

Table 1
Number of claims

Number of claims	Observed distribution	Polya	Poisson
0	25.356	25.355,7	24.993,6
1	1.521	1.524,1	2.149,1
2	282	276,7	92,6
3	58	61,4	2,7
4	16	14,9	0,0
5	4	3,8	0,0
≧ 6	1	1,4	0,0
	27.238	27.238,0	27.238,0

The fit is very good for the Polya model, but as expected the Poisson model cannot be used even as a rough approximation. As the time scale is operational in the Polya case we have the observed mean number of claims per policy $t^* = 0{,}086$. Because of the asymptotic normality of the mean value, an approximative confidence interval on the 95 percentage level can be computed to (0,08, 0,09). The parameter h is estimated according to the principle of maximum likelihood to $h^* = 0{,}19836$ and $Nh^* = 5402{,}9$. These estimates will be used in the parametric mathematical model.

Table 2
Claims distribution

Interval	Observed number	Exponential polynomial
0- 500	1.229	1.223,12
500- 1.000	578	522,42
1.000- 2.000	334	356,34
2.000- 3.000	83	106,88
3.000- 4.000	36	47,45
4.000- 5.000	25	26,78
5.000- 10.000	35	38,29
10.000- 15.000	12	6,34
15.000- 20.000	4	3,05
20.000- 30.000	3	4,93
30.000- 50.000	4	5,40
50.000- 75.000	3	3,29
75.000-100.000	2	1,64
> 100.000	1	3,05
	2.349	2.349,00

The fit is not exceptionally good, especially not for medium size claims. The estimation has been done by a graphic procedure, where the exponential terms have been estimated successively starting with the largest claims corresponding to the smallest value of B_i. The details in this estimation will not be given here. The mean value in the above distribution of claims is 1.237,23. After the material has been normed to the mean value 1,0 the exponential polynomial has the following values of the parameters

$$A_1 = 0{,}00474 \quad A_2 = 0{,}01035 \quad A_3 = 0{,}20260 \quad A_4 = 0{,}78231$$
$$B_1 = 0{,}01920 \quad B_2 = 0{,}06380 \quad B_3 = 0{,}70800 \quad B_4 = 0{,}56700$$

This material was run in the general purpose programs for the

distribution free method and for the Polya-exponential polynomial model. The following distribution was calculated for the risk premium.

Table 3

Calculated distributions for the risk premium. The function is given at 19 points in the interval (0,84 R^*; 1,20 R^*). $R^* = 106{,}70$. The argument is tabulated as factors of the estimated risk premium.

Argument	D.f. Polya-exp-pol.	D.f. distribution free method
0,84	0,007	0,001
0,86	0,019	0,009
0,88	0,044	0,028
0,90	0,086	0,063
0,92	0,147	0,119
0,94	0,228	0,199
0,96	0,323	0,300
0,98	0,427	0,414
1,00	0,532	0,532
1,02	0,631	0,641
1,04	0,718	0,736
1,06	0,791	0,812
1,08	0,850	0,870
1,10	0,895	0,913
1,12	0,928	0,944
1,14	0,952	0,965
1,16	0,969	0,979
1,18	0,980	0,989
1,20	0,988	0,994
Mean	106,70	106,70
Standard deviation	12,51	11,11
Skewness	0,484	0,474

As expected the variance and skewness is larger in the Polya case. The relative difference between the two distributions is large in the "tails", but the length of confidence intervals will be almost the same for the two distributions. The interval $R^* \pm 0{,}1\ R^*$ contains in the Polya case 81 percent of the probability mass and 85 percent in the distribution free case. For practical purpose when the control of premium rates is concerned, it seems as if the two methods give approximately the same result, if the number of policies is fairly large. In small risk groups we have to make certain assumptions in both methods, and it is here difficult to state which method has the best merits in applied research.

REFERENCES

H. AMMETER, A Generalization of the Collective Theory of Risk in Regard to Fluctuating Basic-Probabilities. SAT 1948.

H. BOHMAN and F. ESSCHER, Studies in Risk Theory with Numerical Illustrations Concerning Distribution Functions and Stop Loss Premiums. Part I SAT 1963.

H. BÜHLMANN, A Distribution Free Method for General Risk Problems. The Astin Bulletin 1964.

H. CRAMÉR, Mathematical Methods of Statistics. Uppsala 1945.

O. LUNDBERG, On Random Processes and their Application to Health and Sickness Statistics. Uppsala 1964.

E. PARZEN, Stochastic Processes. San Francisco 1962.