ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 175, 2004

*Anna Szymańska**

SELECTED STATISTICAL METHODS OF INSURANCE RISK ASSESSMENT

Abstract. Effective management of an insurance company calls for diverse types of quantitative information. Different data sets are needed to fix premiums and different ones for loss handling purposes. The mass character of insurance contracts causes that we deal with a huge amount of data. Hence, the necessity of applying statistical methods in examining regularities governing insurance processes. The goal of this work is to present statistical methods most frequently used for insurance risk assessment, that is for examining distributions of random variables of the number and size of claims in the portfolio.

Key works: insurance risk, the number of claims, approximating methods.

1. INTRODUCTION

The objects of insurance statistics research are the insureds set and the insurance accidents set. The statistical unit is the object of insurance. For example, vehicles are statistical units in motor insurance.

A set of statistical units of a given type is called the insurance portfolio.

Analysing a statistical unit from the insurance risk assessment point, we are interested primarily in such characteristics as: number and frequency of occurred losses and value of claims.

Due to the fact that statistical estimates are made on the basis of historical data, statistical methods in the case of new insurance products suffer from some constraints, which are aggravated by a short history of the Polish insurance market.

Statistical methods of insurance risk assessment aim at determining distributions of random variables of the number and size of claims and their main parameters. Three basic groups of statistical methods used for insurance risk assessment can be distinguished:

* PhD, Department of Statistical Methods, University of Łódź.

– descriptive statistics methods used for estimating the empirical distribution function and its main parameters,

– analytical methods of estimating the random variable distribution function fitted to real data,

– method of estimating only main characteristics of the random variable.

2. PROBABILITY DISTRIBUTION OF THE NUMBER OF CLAIMS

If the period of time, during which losses occur is fixed, then the number of claims caused by a given entity or the portfolio of risks is usually a discrete random variable. In the actuarial practice there is most frequently used the empirical distribution function, with probabilities being estimated by means of observed frequencies, with which the values of the random variable have been taken (Bowers *et. al.* 1986). However, the past is not always representative for the future. In such cases the distribution of the random variable of the number of claims is sought. In the case when the number of losses is, moreover, a function of time, we can speak about a discrete random process.

In practice we can meet portfolios composed of a big number of individual risks characterised by small probabilities of loss occurrence. Then the process of $N(t)$ losses in the time period from 0 to t is Poisson distributed on condition the following assumptions are fulfilled:

– numbers of losses occurring in any two disjoint time intervals are independent;

– no more than one claim can arise from the same event;

– probability, that a loss will occur at a definite time point is equal to zero.

Denoting the claim number random variable by N, probability of exactly n claims occurrence in a given period of time amounts to:

$$P(N = n) = e^{-\lambda}\frac{\lambda^n}{n!} \quad \text{for} \quad n = 0,\ 1,\ 2,\ \ldots \tag{1}$$

with $\lambda(t) = E[N]$.

Properties of Poisson distribution indicate that the number of independent Poisson distributed random variables $N_1, N_2, \ldots, N_m$ is a random variable having Poisson distribution with parameter $\lambda = \lambda_1 + \lambda_2 + \ldots + \lambda_m$, being the sum of parameters of respective random variables $N_1, N_2, \ldots, N_m$. Let G denote the joint distribution function of the random variable N.

Recursion formula for calculating distribution of the random variable of the number of claims.

One of methods for estimating probabilities is using the following recursion formula:

$$\mathrm{P}(N = n) = \frac{\lambda}{n} \cdot \mathrm{P}(N = n - 1) \tag{2}$$

with initial value

$$\mathrm{P}(N = 0) = \mathrm{e}^{-\lambda} \tag{3}$$

Normal approximation of probability distribution of the claim number random variable.

On the basis of the central limit theorem (D o m a ń s k i, P r u s k a 2000) for large m, the random variable of claims number N having Poisson distribution with parameter $\lambda = m$ is the sum of m independent random variables identically Poisson distributed with parameter 1, and it can be approximated by means of the normal distribution:

$$G(n) = \mathbf{N}\left(\frac{n - \lambda}{\sqrt{\lambda}}\right) \tag{4}$$

Anscombe approximation of probability distribution of the random variable of the number of claims:

$$G(n) \approx \mathbf{N}\left(\frac{3}{2}\left(n + \frac{5}{8}\right)^{2/3} \cdot \lambda^{-1/6} - \frac{3}{2}\sqrt{\lambda} + \frac{1}{24\sqrt{\lambda}}\right) \tag{5}$$

Peizer and Pratt probability distribution of the claim number variable:

$$G(n) \approx \mathbf{N}\left(\left[\frac{n - \lambda}{\sqrt{\lambda}} + \frac{1}{\sqrt{\lambda}}\left(\frac{2}{3} + \frac{0.22}{n + 1}\right)\right] \cdot \sqrt{1 + T(z)}\right) \tag{6}$$

where

$$z = \frac{n + 0.5}{\lambda} \tag{7}$$

$$T(z) = \frac{1 - z^2 + 2z \ln(z)}{1 - z^2} \tag{8}$$

and $T(1) = 0$.

Approximation formulas have been compared with one another in the Tab. 1.

Table 1

Comparison of the value of random variable of the number of claims estimated by different methods

λ	n	Values of distribution function $G(n)$ for $n \leqslant \lambda$ and $1-G(n)$ for $n>\lambda$ of Poisson distributed random variable			
		exact values	normal approximation	anscombe approximation	Peizer and Pratt approximation
10	0	0.00045	0.000783	0.000034	0.000044
	2	0.002769	0.005706	0.002672	0.002763
	8	0.332820	0.263545	0.332775	0.332833
	10	0.583040	0.500000	0.582704	0.583059
	12	0.208444	0.263545	0.208786	0.208432
	18	0.007187	0.000252	0.007137	0.007187
	23	0.000120	0.000020	0.000115	0.000120
100	80	0.022649	0.022750	0.022643	0.022649
	90	0.171385	0.158655	0.171405	0.171386
	100	0.526562	0.500000	0.526551	0.526563
	110	0.147137	0.158655	0.147161	0.147137
	120	0.022669	0.022750	0.022665	0.022669
	130	0.001707	0.001350	0.001703	0.001707
	140	0.000064	0.000032	0.000064	0.000064
	145	0.000010	0.000003	0.000010	0.000010
1 000	905	0.001215	0.001332	0.001214	0.001215
	937	0.023172	0.023173	0.023172	0.023172
	968	0.159596	0.155786	0.159599	0.159596
	1 032	0.152095	0.155786	0.152097	0.152095
	1 063	0.023155	0.023173	0.023155	0.023155
	1 095	0.001446	0.001332	0.001446	0.001446

Source: Daykin *et al.* 1994.

Results of numerical tests made by C. D. Daykin showed that the maximal error in Anscombe approximation is lower than 10^{-4} for $\lambda/35$. A constraint for Peizer and Pratt method is $\lambda/6$. Normal approximation yields good results for $\lambda/1000$. An advantage of Peizer and Pratt approximation formula is its correctness for small λ values. Anscombe approximation is more convenient than Peizer and Pratt approximation for numerical reasons.

Changes in intensity of losses in the portfolio caused by external factors such as weather, economic conditions, and so on are frequently observed in practice.

If changes in intensity of losses are of random character, then the random variable of the number of claims can have Poisson mixed distribution. Mixed Poisson claim number variable $Q > 0$ fulfils condition $E[Q] = 1$, which means that the intensity of losses over a certain time period can be at a definite level. If Q assumes values bigger than 1, then the intensity of losses is higher than expected, if it assumes values in the interval from 0 to 1, then the intensity of losses is lower than expected.

If random variable Q accepts value q, then the conditional claim number d.f. $P(N/Q = q)$ is Poisson distribution with λq parameter.

The problem is to estimate distribution of the mixing random variable Q. Too few data are usually available to construct analytically the form of the mixing random variable distribution function. At such time the method of moments (D o m a ń s k i 2001) is used and only the main characteristics are estimated without seeking the distribution function form. If the number of data is big enough frequency series are formed and the distribution function form is estimated on their basis. The mixing random variable most frequently has gamma distribution (B o w e r s *et al.* 1986).

Equally interesting problem is a search for the distribution function of the number of claims coming from individual insurance policies. This is significant due to fluctuations in occurrence of losses from particular policies in the portfolio, which has a direct impact on the level of premiums.

Each nth policy has the claim frequency parameter k_i described by the formula:

$$k_i = kh_i \tag{9}$$

where k is the average number of losses, and h_i the deviation coefficient per unit from k. Distribution of risk in the portfolio is characterised by the distribution function H of random variable h_i. Function H is called the risk structure function in the portfolio.

3. PROBABILITY DISTRIBUTION OF THE AMOUNT OF CLAIMS

The amount of claims resulting from an individual loss is a random variable of continuous type (B o w e r s et al. 1986).

Three methods are used for estimating the distribution of random variables for an individual loss (D a y k i n *et al.* 1994):

- analytical method,
- empirical distribution function construction method,
- basic parameters estimation method,

These methods will be briefly discussed below.

The analytical method consists in choosing the analytical form of distribution function fitting to the observed data. This method practically means choosing such a distribution among known distributions, which fulfils a definite criterion best.

The most frequently used distributions are: random variable has: gamma, logarithmic or Pareto distributions (Ronka-Chmielowiec 1997).

We choose the distribution, which minimises the value of χ^2 statistics (Domański 2001).

The empirical distribution function construction method consists in building a disjoint series from observed data and constructing the empirical distribution function on this basis.

The tabular method of the empirical distribution function construction $P(x)$ is commonly used here:

$$P(x) = \frac{k \leqslant x}{K} \tag{10}$$

where k is the number of claims with values lower or equal to x and K is the total number of losses. The disjoint series table is built on the basis of data and next the empirical distribution function is constructed (Daykin *et al.* 1994).

It should be remembered that intervals have to be built in an unbiased way and that due to specificity of insurance the length of class intervals should increase along with the number of claims (for instance, geometrically). The tabular method is used in the case of a big data number.

The method of main parameters estimation consists in choosing the distribution function with given parameters determining the distribution. The parameters are estimated by means of the maximum likelihood method or the method of moments. Following estimation of the theoretical distribution function the fit of theoretical and empirical distributions is tested using χ^2 or λ-Kolmogorov tests (Domański 2001).

If for all known family distributions the zero hypothesis (H_0: the sample for which the empirical distribution has been estimated comes from the population with the tested theoretical distribution) is untrue, then the insurance portfolio is heterogeneous. In such case the portfolio should be divided into risk groups so that a uniform distribution of the number of claims P_i can be found in each group. Then the distribution function P will be a convex combination of distribution functions P_i with appropriate weights. However, finding such weight can be impossible. Some authors suggest that the method of limited expected value function should be used in such case (Daykin *et al.* 1994).

All three methods are generally used simultaneously to estimate the claim number distribution in the actuarial practice. Data are verified in such a way that allowances can be made for inflation.

The length of research periods is frequently differentiated according to the loss size. Data about large losses, for example, catastrophes, for which the research period should amount from 10 to 100 years, are most frequently unavailable Despite the fact that these losses exert a significant impact on the distribution function, it is impossible to accept such long research period in the Polish market.

REFERENCES

Bowers N. I., Gerber H. U., Hickman J. C., Jones D. A., Nesbitt C. J. (1986), *Actuarial Mathematics*, The Society of Actuaries, Itasca (Ill).

Daykin C. D., Pentiainen T., Pesonen M. (1994), *Practical Risk Theory for Actuaries*, Chapman & Hall, London.

Domański C. (2001), *Statistical Methods*, University of Łódź Press, Łódź.

Domański C., Pruska K. (2000), *Non-Classical Statistical Methods*, PWE, Warszawa;

Ronka-Chmielowiec W. (1997), *Insurance Risk – Assessment Methods*, Akademia Ekonomiczna, Wrocław.

Anna Szymańska

WYBRANE METODY STATYSTYCZNE OCENY RYZYKA UBEZPIECZENIOWEGO

Prawidłowe zarządzanie towarzystwem ubezpieczeniowym wymaga różnorodnych informacji wartościowych i ilościowych. Inne dane są potrzebne do wyznaczania składek, inne dla potrzeb likwidacji szkód. Masowość ubezpieczeń sprawia, że w przypadku ubezpieczeń mamy do czynienia z olbrzymią liczbą danych. Stąd potrzeba zastosowania metod statystycznych do badania prawidłowości rządzących procesami ubezpieczeniowymi. Celem pracy jest przedstawienie najczęściej stosowanych metod statystycznych, służących do oceny ryzyka ubezpieczeniowego, czyli do badania rozkładów zmiennych losowych liczby roszczeń i wielkości roszczeń w portfelu.