STATISTICS IN TRANSITION new series, Autumn 2013
Vol. 14, No. 3, pp. 507–516

THE DISTRIBUTION OF THE NUMBER OF CLAIMS IN THE THIRD PARTY'S MOTOR LIABILITY INSURANCE

Anna Szymańska[1]

ABSTRACT

In the automobile insurance tarification consists of two stages. The first step is to determine the net premiums on the basis of known risk factors, called *a priori* ratemaking. The second stage, called *a posteriori* ratemaking is to take into account the driver's claims history in the premium. Each step usually requires the actuary's selection of the theoretical distribution of the number of claims in the portfolio. The paper presents methods of consistency evaluation of the empirical and theoretical distributions used in motor insurance, illustrated with an example of data from different European markets.

Key words: distribution of the number of claims, civil liability motor insurance of vehicle owners.

1. Introduction

Calculation of the premium in property insurance is a complex process. The insurer at the time of setting the premium does not know the future costs of compensation, but they can be estimated on the base of the expected number and the amount of claims (Śliwiński, 2002, p.82). Estimation of the expected values of the random variable distributions representing the amount and number of claims requires the determination of theoretical distributions of the random variables.

The aim of the paper is to present methods of evaluation of the goodness-of-fit of theoretical distributions to empirical distributions of the number of claims in motor liability insurance. The data on the number of claims from the Polish market (from one of the insurance companies from Lodz, from the Lodz Region for the years 2000, 2001, 2002) were analyzed. Empirical distributions were chosen deliberately so as to assess functioning in the literature reselection methods of theoretical distribution of the number of claims.

[1] Department of Statistical Methods, University of Lodz. E-mail: szymanska@uni.lodz.pl.

2. Distributions of the number of claims used in car liability insurance

Let the random variable X represent the number of claims from individual policy or a policy portfolio. In motor liability insurance different theoretical distributions may be used to model the number of claims (Lemaire, 1995). In the following most commonly used distributions of random variables are presented.

Bernouli binominal distribution is described with the probability distribution function:

$$P(X=k)=\binom{n}{k}p^k q^{n-k}, where\ \ k=0,1,...,n\ and\ \ \binom{n}{k}=\frac{n!}{k!(n-k)!}. \tag{1}$$

Poisson distribution is a distribution with the function of the probability defined by the formula:

$$P(X=k)=\exp(-\lambda)\frac{\lambda^k}{k!},\ k=0,1,.... \tag{2}$$

The random variable X has a negative binomial distribution (Polya) when its probability distribution function has the form:

$$P(X=k)=\frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{k}. \tag{3}$$

If the random variable X has a Poisson distribution with parameter λ and the parameter λ has the inverse normal distribution, then the random variable X has a *Poisson-inverse normal* distribution (Denuit, Marechal, Pitrebois, Walhin, 2007, p.31). The probability function of Poisson-inverse normal distribution is given by:

$$P(X=k)=\sqrt{\frac{2\alpha}{\pi}}\exp\left(\alpha\sqrt{1-\theta}\right)\frac{(\alpha\theta/2)^k}{k!}K_{k-1/2}(\alpha)\quad k=0,1,...\ , \tag{4}$$

where $K_{k-1/2}(\alpha)$ is a modified third kind Bessel function (for positive and real arguments) in the form of:

$$K_{k-1/2}(\alpha)=\sqrt{\frac{\pi}{2\alpha}}\exp(-\alpha)\left(\sum_{i=0}^{k-1}\frac{(k-1+i)!}{(k-1-i)!i!}(2\alpha)^{-i}\right), k=1,2.... \tag{5}$$

Probability distribution function of the *Poisson-Poisson* distribution (*Neyman type A*) is given by:

$$P(X=k)=\exp(-\lambda_1)\frac{\lambda_2{}^k}{k!}\sum_{n=0}^{\infty}\frac{n^k}{n!}(\lambda_1\exp(-\lambda_2))^n,\quad k=0,1,2,... \tag{6}$$

Generalized Poisson-Pascal distribution is described by the moment generating function, which has the form:

$$M_X(t)=\exp\left\{\lambda\left[\frac{[1-\beta(t-1)]^{-\alpha}-(1+\beta)^{-\alpha}}{1-(1+\beta)^{-\alpha}}-1\right]\right\},\quad \alpha>-1,\lambda>0,\beta>0. \tag{7}$$

For $\alpha > 0$ the above distribution is called the *Poisson-Pascal* distribution. For $\alpha = 1$ the distribution is called *Polya-Aeppli* distribution. For $\alpha = -0{,}5$ it is called Poisson-inverse normal distribution.

Heilmann suggests the choice of the distribution of the number of claims in civil motor liability insurance depending on the relationship between the expected value and variance of the sample (Heilmann, 1988, p.46). Three distributions are considered: binomial, Poisson and negative binomial, which belong to the class (*a, b, 0*) (Otto, 2002, p.95). Wherein the family of distributions is called family distributions class *(a, b, 0),* where *a* and *b* are constants such that:

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}, \quad k = 1,2,3,..., \tag{8}$$

where p_k is a function of the probability distribution of the discrete random variable. Treating the pair of parameters *(a, b*) as a point on the coordinates space of the areas corresponding to certain types of distributions can be designated. For $\{(a,b) : a = 0 \wedge b > 0 \wedge b \in R\}$ the distribution (*a, b, 0*) is the Poisson distribution, for $\{(a,b) : a \le 0 \wedge b > -a \wedge a \in R \wedge b = 1,2,3,...\}$ it is the binomial distribution, for $\{(a,b) : a \in (0,1) \wedge b > -a \wedge a,b \in R\}$ – the negative binomial distribution (Otto, 2004).

According to the paper (Panjer, Willmot, 1992, p.292) pre-selection of the theoretical distribution of the number of claims can be based on the calculated moments of the sample and the frequency coefficients.

Let $X_1, X_2,..., X_n$ be an i.i.d. random sample. In case of aggregated data, where we know only the number of policies for the number of claims, simple sample moments usually are:

$$M_r = \frac{1}{n}\sum_{l=1}^{\infty} k^r N_k, \quad r = 1,2,..., \tag{9}$$

where N_k is the number X_i for which $X_i = k$, (k = 0,1,2, ...), $n = \sum_{k=0}^{\infty} N_k$

and $\overline{X} = \frac{1}{n}\sum_{k=0}^{\infty} kN_k$. The first three central moments of the sample are: $\overline{X} = M_1$; $S^2 = M_2 - M_1^2$; $K = M_3 - 3M_2M_1 + 2M_1^3$. Frequency coefficients are described by the following equation:

$$T(k) = (k+1)\frac{N_{k+1}}{N_k}, \quad k = 0,1,2,.... \tag{10}$$

Let:

$$T(k) = (a+b) + ak, \quad k = 0,1,2,... \tag{11}$$

be a function. When the function given by equation (11) is linear, whose slope coefficient:

- is zero and $\overline{X} = S^2$; then to describe the distribution of the number of claims the Poisson distribution is suggested;
- is negative and $\overline{X} > S^2$; then the binomial distribution can be assumed;
- is positive and $\overline{X} < S^2$; then the negative binomial distribution should be chosen.

When the function described by equation (11) grows faster than linearly, the skewness of the distribution should be considered. If the equation $K = 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$ holds, the negative binomial distribution should model the number of claims well. If inequality $K < 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$ holds, the generalized Poisson Pascal distribution, or its special case the Poisson-inverse normal distribution can be used to describe the distribution of the number of claims. If the inequality $K > 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$ holds, the Neyman type A, Polya-Aeppli, Poisson-Pascal or negative binomial distributions are suitable for modeling the distribution of the number of claims.

3. Statistical methods of assessing the fitness of empirical and theoretical distributions

In the actuarial literature the tests which are most commonly used for evaluation of the relevance of the theoretical distribution to empirical data are: goodness-of-fit test χ^2 and test statistics based on λ – Kolmogorow (Domański, 1990, p.61). However, in case of distribution of the number of claims in car automobile insurance the number of classes is often not larger than four, which means that the number of degrees of freedom of the chi-squared test is too small. Additionally, most policies in the insurance portfolios are concentrated in the number zero class, which results in the distortion of the distribution. Portfolios are usually large, with the consequence that chi-squared test generally rejects the null hypothesis even though empirical data closely match theoretical distribution. In such cases, measures assessing the degree of fit of the theoretical distribution to empirical data may be found in statistical literature, such as the standard deviation of the differences in relative frequencies, the index of structures similarity, index of distribution similarity, ratio of the maximum difference of relative frequencies, ratio of the maximum difference of cumulative distribution functions (Kordos, 1973, p.115 -118).

Deviation of the differences in relative frequencies is a measure given by:

$$S_r = \sqrt{\frac{1}{k}\sum_{i=1}^{k}\left(\gamma_i - \hat{\gamma}_i\right)^2}, \tag{12}$$

where: k - the number of classes, γ_i - empirical frequencies, $\hat{\gamma}_i$ - theoretical frequencies.

The measure is equal to zero in the case of full compliance of the empirical and theoretical distribution. The practice shows that the value $S_r \le 0.005$ is an evidence of high compliance of schedules, if $0.005 \le S_r < 0.01$ the compatibility of tested distributions is satisfactory and $S_r \ge 0.01$ shows significant deviations between the studied distributions.

The index of *structures similarity* is given by:

$$w_p = \sum_{i=1}^{k}\min\left(\gamma_i, \hat{\gamma}_i\right). \tag{13}$$

The index value is in the range [0,1]. The closer the value is to the unity, the more similar the structures of the studied distributions are.

Index of distribution similarity is determined by the equation:

$$W_p = 1 - \frac{1}{2}\sum_{i=1}^{k}\left|\gamma_i - \hat{\gamma}_i\right|. \tag{14}$$

Distribution similarity index is equal to 100% for fully compatible distribution. The distributions show high compatibility when $W_p \ge 0.97$. If $W_p < 0.95$ distributions show significant differences.

Ratio of the maximum difference of relative frequencies is given by the formula:

$$r_{\max} = \max_i\left|\gamma_i - \hat{\gamma}_i\right|. \tag{15}$$

This ratio is equal to zero for distributions fully compatible. If $r_{\max} < 0.02$, it is believed that the distributions are quite compatible.

Ratio of the maximum difference of cummulative distribution functions is given by the equation:

$$D_{\max} = \max_i\left|F_i - \hat{F}_i\right|. \tag{16}$$

where: $F_i = \sum_{j=1}^{i}\gamma_j$ - value of the empircial cummulative distribution function,

$\hat{F}_i = \sum_{j=1}^{i}\hat{\gamma}_i$ - value of the theoretical cummulative distribution function. This ratio is equal to zero for fully consistent distributions.

In the analyses of the consistency of distributions of the number of claims with theoretical distributions, comparisons of distribution parameters such as mean, median, first and third quartile as well as measurement variation distributions can be used, in addition to the data indicators formulas (12) - (16). It is assumed that if the relative differences in ratings of all parameters between theoretical and empirical distribution do not exceed 5%, the distributions are fairly consistent.

4. Empirical example

This part of the paper presents the evaluation of the distribution of the number of claims in motor insurance for the actual sample data sets.

Table 1 presents data from the Polish market respectively in the years: 2000 (P1-option 1), 2001 (P2-option 2), 2002 (P3-option 3).

Table 1. Distribution of the number of claims

Number of claims	Number of policies			T_k		
	P1	P2	P3	P1	P2	P3
0	21 570	21 922	22 451	0.0313	0.0332	0.0284
1	676	730	638	0.0947	0.0712	0.0689
2	32	26	22	0.1875	0.2308	0.2727
3	2	2	2			
4	2	0	0			
sum	22 282	22 680	23 113			

Table 2 presents estimated parameters of the distributions presented in table 3.

Table 2. Parameters of the distribution of the number of claims

Parameters of the distribution	Distribution		
	P1	P2	P3
a	0.063334732	0.037932995	0.040548075
$\overline{X}$	0.033838973	0.034744268	0.029766798
S^2	0.037181825	0.036358973	0.031303615
K	0.04618218	0.039907237	0.034732823
$3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$	0.044527988	0.039738468	0.034535935

For each of the considered distributions the relationships $a > 0, \overline{X} < S^2$ and $K > 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$ hold. Further analysis included the following theoretical distributions: negative-binomial, Poisson-inverse normal and Neyman type A.

Table 3. Measures of the degree of fit of theoretical distributions to distribution P1

Measure	Theoretical distribution		
	Negative-binominal	Poisson-inverse normal	Neyman A type
S_r	0.00029986	0.00047542	0.01311831
w_p	0.99940006	0.99941457	0.96913855
W_p	0.99999978	0.99999943	0.99956977
r_{max}	0.00000027	0.00000091	0.00085846
D_{max}	0.00032057	0.00054293	0.03061508

Table 4. Measures of the degree of fit of theoretical distributions to distribution P2

Measure	Theoretical distribution		
	Negative-binominal	Poisson-inverse normal	Neyman A type
S_r	0.00006487	0.00031944	0.01390219
w_p	0.99986817	0.99949218	0.96772665
W_p	0.99999999	0.99999974	0.99951682
r_{max}	0.00000001	0.00000031	0.00096509
D_{max}	0.00006366	0.00048811	0.03223130

Table 5. Measures of the degree of fit of theoretical distributions to distribution P3

Measure	Theoretical distribution		
	Negative-binominal	Poisson-inverse normal	Neyman A type
S_r	0.00007375	0.00026573	0.01198220
w_p	0.99985053	0.99962895	0.97220702
W_p	0.99999999	0.99999982	0.99964107
r_{max}	0.00000001	0.00000026	0.00071699
D_{max}	0.00007263	0.00034735	0.02774574

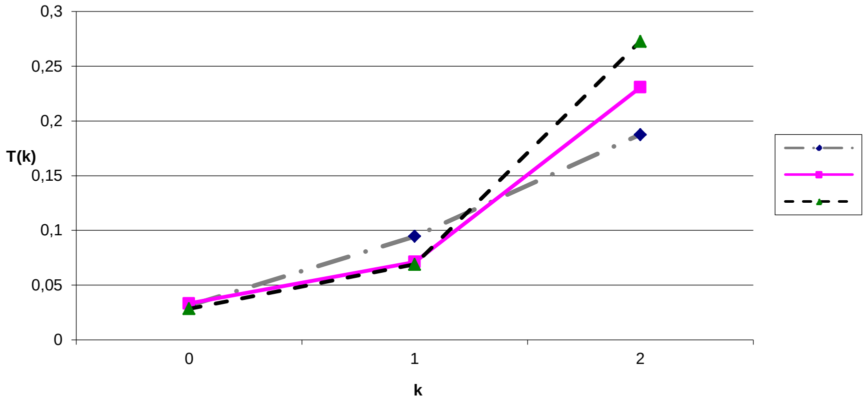

Figure 1. *T*(*k*) function of analyzed distributions

All empirical distributions received on the basis of data from the Polish market comply with the negative binomial distribution. This distribution gives the best fit to the empirical distribution of P2 (see tables 3-5), the worst to the distribution P1. The pre-fit empirical distribution of the negative binomial distribution can be assessed by the graph of *T*(*k*) function. The closer the graph of the function *T*(*k*) is to a straight line, the weaker fit it gives.

5. Conclusions

The conducted study shows that the methods of the selection of theoretical distribution of the number of claims proposed in the actuarial literature usually show the theoretical distribution that gives the best fit to the empirical data. Further studies are needed to confirm that for a linear *T*(*k*) function with a positive slope coefficient and for $\overline{X} < S^2$ the negative binomial distribution should be chosen to describe the number of claims. In the case of the distributions considered, the results of the fit of the empirical distribution to the negative binomial distribution were worst the closer the function *T*(*k*) was to a linear function.

In assessing the consistency of distributions, in most cases, the chi-square test cannot be used due to the nature of the data on the number of claims in motor liability insurance. Measures proposed in the paper offer a possibility to assess the goodness-of-fit of empirical and theoretical distributions.

It is not possible to unequivocally specify the type of theoretical distribution of the number of claims in motor liability insurance, although the distribution that gives the best fit is the negative binomial distribution. However, for each insurance market the distribution of the number of claims can be consistent with different theoretical distributions.

REFERENCES

DENUIT M., MARECHAL X., PITERBOIS S., WALHIN J., (2007). Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems, John Wiley & Sons, England.

DOMAŃSKI CZ., (1990). Testy statystyczne, PWE, Warszawa.

HEILMANN W. R., (1988). Fundamentals of Risk Theory, Verlag Versiecherungswirtschaft, Karlsruhe.

KORDOS J., (1973). Metody analizy i prognozowania rozkładów płac i dochodów ludności, PWE, Warszawa.

LEMAIRE J., (1995). Bonus-Malus Systems in Automobile Insurance, Kluwer, Boston.

OTTO W., (2002). Matematyka w ubezpieczeniach. Ubezpieczenia majątkowe, WNT, Warszawa.

PANJER H. H., WILLMOT G. E., (1992). Insurance risk models, Society of Actuaries, Schaumburg.

ŚLIWIŃSKI A., (2002). Ryzyko ubezpieczeniowe, taryfy - budowa i optymalizacja, poltext, Warszawa.