Distribution of the number of claims in age groups of insured in civil liability motor insurance portfolio

Anna Szymańska[1]

Abstract

In the process of calculation of the insurance premium in civil liability motor insurance the knowledge of distribution of the number and value of paid claims is required. The paper presents the methods of assessing the degree of the fit of theoretical distributions of the number of claims to empirical distributions in civil liability motor insurance in the example of data from one of the insurance companies in Poland. The distributions of the number of claims in separate age groups of drivers were also analyzed and their compatibility to the theoretical distribution of the number of claims in the portfolio was assessed.

Keywords*: distribution of the number of claims, civil liability motor insurance of vehicle owners, driver's age*
JEL Classification: C12, C13, G220

1. Introduction

At the moment of establishing the insurance premium insurance company does not know the future costs of compensation, but it can estimate them on the basis of historical data of the number and value of paid claims. Estimation of the value of the expected distributions of random variables of the number and value of claims requires the determination of theoretical distributions of these random variables.

In the actuarial literature, the tests which are usually used to evaluate the relevance of the theoretical distribution to empirical data, are: goodness-of-fit test χ^2 and test statistics based on λ – Kolmogorow (Panjer and Willmot, 1992). However, in the case of the 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. Moreover, 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, resulting in the chi-squared test generally rejecting the null hypothesis even though empirical data match theoretical distribution closely. In such cases, measures assessing the degree of the 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,

[1] University of Łódź, Department of Statistical Methods, ul. Rewolucji 1905 r. nr 41, 90-214 Łódź, Poland szymanska@uni.lodz.pl

index of distribution similarity, ratio of the maximum difference of relative frequencies, ratio of the maximum difference of cummulative distribution functions (Kordos, 1973).

In the study the distribution of the number of claims in the portfolio of civil liability motor insurance of individuals' cars in 2006, of one of the insurance companies operating on the Polish market, is analyzed. In civil liability motor insurance ratemaking is a two-step process (Antonio and Valdez, 2012; Dionne and Vanasse, 1989; Szymańska, 2013). In the first stage - called *a piori* – the base premium based on known risk factors is determined. Then, in the base premium the discounts and increases, mainly resulting from the claims experience during the previous period of insurance are taken into account. This stage is called *a posteriori* ratemaking. Because in the examined insurance company the base premium is determined on the basis of two factors: the vehicle registration region and engine capacity, and the age of the driver is taken into account in the premium by means of discounts and increases, the aim of this work is to analyze the distributions of the number of claims in the age groups of the insured and in the whole civil liability motor insurance portfolio.

2. The choice of the distribution of the number of claims in the civil liability motor insurance

Let the random variable X represent the number of claims from individual policy or a policy portfolio. The choice of the distribution of the number of claims in civil motor liability insurance depends on the relationship between the sample expected value and variance (Heilmann, 1988). Three distributions are considered: binomial, Poisson and negative binomial, which belong to the class (*a,b,0*) (Klugman et al., 2004).

According to the paper (Panjer and Willmot, 1992) the 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:

$$M_r = \frac{1}{n}\sum_{k=0}^{\infty} k^r N_k, \quad r = 1,2,... \tag{1}$$

where N_k is the number X_i for which $X_i = k$, (k = 0,1,2, ...), $n = \sum_{k=0}^{\infty} N_k$. The first three central moments of the sample are: $\bar{X} = M_1$; $S^2 = M_2 - M_1^2$; $K = M_3 - 3M_2M_1 + 2M_1^3$.

Frequency coefficients are described by the following equation:

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

Let:

$$T(k) = (a+b) + ak, \quad k = 0,1,2,..., \tag{3}$$

be a function. When the function given by equation (3) 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 (3) grows faster than linearly, the skewness of the distribution should be taken into account.

Let us denote: $W = 3S^2 - 2\overline{X} + 2\frac{(S^2 - \overline{X})^2}{\overline{X}}$. If the equation: $K = W$ holds, the negative binomial distribution should model the number of claims well. If inequality $K < W$ 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 (Nadarajah and Kotz, 2006; Tremblay, 1992). If the inequality $K > W$ 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 measures of fit of the empirical and theoretical distributions

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

$$S_r = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(\gamma_i - \hat{\gamma}_i)^2}, \tag{4}$$

where *k* - the number of classes, γ_i - empirical frequencies, $\hat{\gamma}_i$ - theoretical frequencies. The measure is equal to zero in case of full compliance of the empirical and theoretical distribution. Practice shows that the value $S_r \leq 0{,}005$ is an evidence of high compliance of schedules, if $0{,}005 \leq S_r < 0{,}01$ the compatibility of tested distributions is satisfactory and $S_r \geq 0{,}01$ shows significant deviations between the studied distributions.

The index of *structures similarity* is given by:

$$w_p = \sum_{i=1}^{k} \min(\gamma_i, \hat{\gamma}_i). \tag{5}$$

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:

$$W_p = 1 - \frac{1}{2} \sum_{i=1}^{k} |\gamma_i - \hat{\gamma}_i|. \tag{6}$$

Distribution similarity index is equal to 100% for fully compatible distribution. The distributions show high compatibility when $W_p \geq 0{,}97$. If $W_p < 0{,}95$ distributions show significant differences.

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

$$r_{\max} = \max_i |\gamma_i - \hat{\gamma}_i|. \tag{7}$$

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:

$$D_{\max} = \max_i \left| F_i - \hat{F}_i \right|, \tag{8}$$

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.

4. Empirical examples

In this part of the study investigated the distribution of the number of paid claims in the portfolio and in the separated age groups of drivers of the analyzed civil liability motor insurance portfolio of the insurance company offering property insurance in Poland in 2006. Figure 1 shows the structure of according to the age of the insured of the civil liability motor insurance portfolio of the analyzed insurance company.

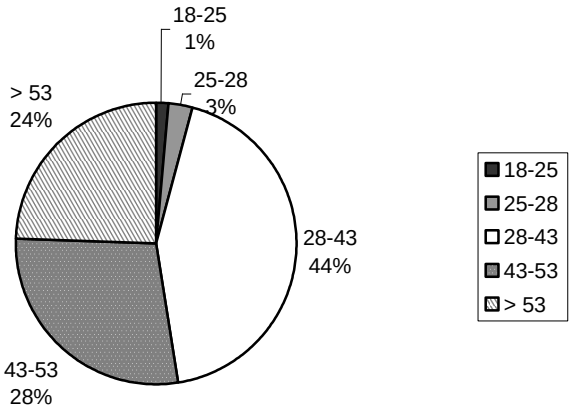

Fig.1. Portfolio structure by age of the insured in 2006

On the basis of empirical data about the number of paid claims the numerical characteristics of distributions in particular age groups of the insured and in the portfolio were calculated (see Table 1).

Numerical characteristics of distributions	**portfel**	**Age of insured [years]**				
		18-25	**25-28**	**28-43**	**43-53**	**>53**
$\overline{X}$	0.0441	0.0743	0.0477	0.0425	0.0442	0.0411
S^2	0.0457	0.0783	0.0490	0.0443	0.0455	0.0426
a	0.0308	0.0383	0.0356	0.0386	0.0189	0.0298
K	0.0492	0.0875	0.0514	0.0482	0.0485	0.0459
W	0.0490	0.0869	0.0518	0.0481	0.0481	0.0457

Table 1 Numerical characteristics of the empirical distributions of the number of claims in the age groups of the insured in civil liability motor insurance portfolio insurance company in 2006.

In the next stage the compatibility of the empirical distribution in the portfolio and in the particular age groups of the insured with selected theoretical distributions was examined (results are shown in Tables 2-7). In the selection of theoretical distributions the linearity of the function frequency and the relationships between the parameters of distributions from the sample were taken into account. The frequency functions for particular age groups of the

insured are not linear (see Fig. 2), which suggests consideration of the skewness of distributions. For each of the considered distributions, in addition to the age group of 25-28 years, the following relations hold: $a > 0, \overline{X} < S^2$ and $K > W$ (see Table 1). In further analyzes the following theoretical distributions were considered: *Poisson* (Poi), negative-binomial (*NB*), the Poisson-inverse normal (*PGI*) and Neyman A (*NA*). Generalized Poisson-Pascal distribution in this case could not be considered due to the assumptions about the parameters of this distribution not fulfilled by the empirical distribution. The parameters of distributions for each age group and for the portfolio estimated by the maximum likelihood method in the case of the Poisson distributions, by the method of moments in the case of the negative binomial distributions and by means of recursive formulas for the Poisson-inverse normal and Neyman A distributions.

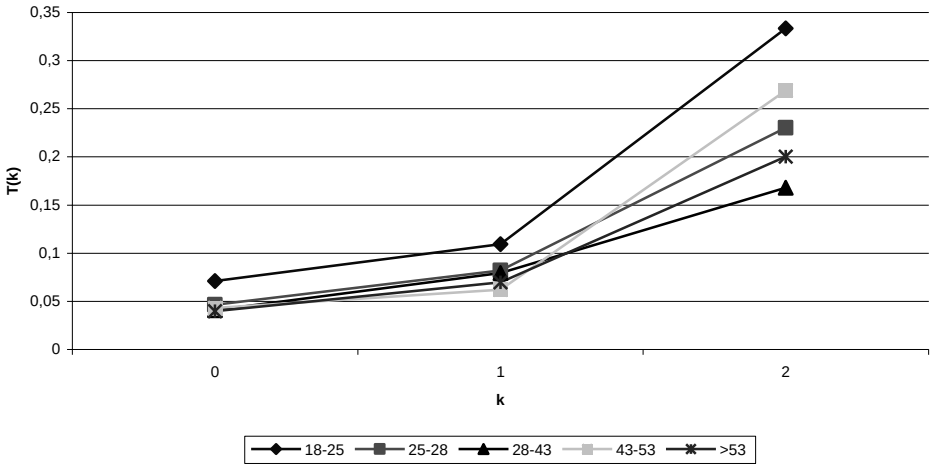

Fig. 2. The frequency functions in the age groups of insured.

Measure	Distribution			
	Poi	***NB***	***PIG***	***NA***
S_r	0.00240807	0.00170703	0.00175441	0.01632648
w_p	0.99597445	0.99716675	0.99658889	0.95944651
W_p	0.99998550	0.99999272	0.99999231	0.99933362
r_{max}	0.00001620	0.00000786	0.00000791	0.00132328
D_{max}	0.00355339	0.00280417	0.00281276	0.04055349

Table 2 Measures of the degree of distributions' fitting claims in the portfolio.

Measure	Distribution			
	Poi	***NB***	***PIG***	***NA***
S_r	0.00164310	0.00026806	0.00098381	0.02734804
w_p	0.99692915	0.99945118	0.99790153	0.93510215
W_p	0.99999325	0.99999982	0.99999758	0.99813021
r_{max}	0.00000942	0.00000017	0.00000325	0.00372750
D_{max}	0.00170156	0.00026007	0.00200691	0.06474332

Table 3 Measures of the degree of distributions' fitting claims in the age group 18-25 years.

Measure	distribution			
	Poi	***NB***	***PIG***	***NA***
S_r	0.00076233	0.00011774	0.00059518	0.01879855
w_p	0.99859679	0.99976459	0.99876536	0.95617589
W_p	0.99999855	0.99999997	0.99999911	0.99911654
r_{max}	0.00000192	0.00000003	0.00000137	0.00176378
D_{max}	0.00070171	0.00011864	0.00103962	0.04382411

Table 4 Measures of the degree of distributions' fitting claims in the age group 25-28 years.

Measure	Distribution			
	Poi	***NB***	***PIG***	***NA***
S_r	0.00086505	0.00002989	0.00044491	0.01676594
w_p	0.99839956	0.99993788	0.99924540	0.96093832
W_p	0.99999813	1.00000000	0.99999951	0.99929726
r_{max}	0.00000256	0.00000000	0.00000056	0.00140318
D_{max}	0.00083739	0.00002819	0.00075343	0.03903725

Table 5 Measures of the degree of distributions' fitting claims in the age group 28-43 years.

Analyzing the values of the measures of fit from Tables 2-7 it follows that in the portfolio, as well as in each considered age group of insured the theoretical distribution, best fit to the empirical data on the number of claims is the negative binomial distribution. The question is: how compatible are the distributions of the number of claims in each age groups of the insured with the distribution of the number of claims in the whole portfolio? In the next stage of the analysis the compatibility of empirical distributions in particular age groups of insured

with the determined for the portfolio negative binomial distribution with parameters $\alpha = 1.1227; \beta = 27.5032$ was rated (see Table 8).

Measure	Distribution			
	Poi	***NB***	***PIG***	***NA***
S_r	0.00053232	0.00012409	0.00037277	0.01766546
w_p	0.99900607	0.99975023	0.99916067	0.95915272
W_p	0.99999929	0.99999996	0.99999965	0.99921983
r_{max}	0.00000099	0.00000004	0.00000045	0.00155876
D_{max}	0.00054717	0.00012324	0.00078425	0.04078068

Table 6 Measures of the degree of distributions' fitting claims in the age group 43-53 years.

Measure	Distribution			
	Poi	***NB***	***PIG***	***NA***
S_r	0.00068258	0.00006351	0.00038041	0.01638094
w_p	0.99873325	0.99987112	0.99928145	0.96200831
W_p	0.99999884	0.99999999	0.99999964	0.99932916
r_{max}	0.00000160	0.00000001	0.00000042	0.00133998
D_{max}	0.00067190	0.00006244	0.00069646	0.03795317

Table 7 Measures of the degree of distributions' fitting claims in the age group above 53 years.

Measure	Age of insured [years]				
	18-25	**25-28**	**28-43**	**43-53**	**>53**
S_r	0.01667476	0.00216927	0.00102821	0.00030703	0.00170703
w_p	0.97252388	0.99643296	0.99834551	0.99940429	0.99716675
W_p	0.99930488	0.99998824	0.99999736	0.99999976	0.99999272
r_{max}	0.00075481	0.00001229	0.00000265	0.00000029	0.00000786
D_{max}	0.02747384	0.00350534	0.00162692	0.00029194	0.00280417

Table 8 Measures of the degree of fitting of the compatibility of empirical distributions of the number of claims in the age groups with negative binomial distribution of the portfolio.

Analyzing the results presented in Table 8, we find that in the case of insured above 25 years old the values of measurements of fitting of distributions show a high compatibility of distributions of the number of claims in these groups with distribution of the number of claims in the portfolio. The highest compatibility was obtained in group of insured aged 43 to 53 years and above 53 years. In the group of insured who are under the age of 25 years the distribution of the number of claims is incompatible with the distribution of the number of claims in the portfolio.

Conclusions

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

In the analyzed civil liability motor insurance portfolio clearly differed in terms of the number of claims was a group of under 25 years old. Average number of compensations paid in 2006 of a single policy in this group was 0.0743, 0.0441 in the portfolio. The average value of compensation paid in 2006 in a group of the drivers aged to 25 years was equal to about 8 thousand zlotys, in the portfolio of approximately 5.5 thousand zlotys. Despite the fact that the insured up to the age of 25 constituted only 1% of the portfolio, the insurance premium for this group of the insured should be estimated separately. Treating the age of the insured as variable in *a priori* tariffication, could reduce the base premiums for the insured who are over the age of 25 years, while increasing the competitiveness of the insurer in the market. Very similar results were obtained for the years 2007-2009.

References

Antonio, K., & Valdez, E. (2012). Statistical concepts of a priopri and a posteriori classification in insurance. *AStA Adv Stat Anal, 96*, 187-224.

Dionne, G., & Vanasse, C. (1989). A generalization of automobile insurance rating models: the negative binomial distribution with a regression component. *ASTIN Bulletin*, *19*(2), 199-212.

Heilmann, W. (1988). *Fundamentals of risk theory*. Karlsruhe: Verlag Versiecherungswirtschaft.

Klugman, S., Panjer, H., & Willmot, G. (2004). *Loss models from data to decisions*. New York: J. Wiley & Sons.

Kordos, J. (1973). *The methods of analysis and forecasting of distribution of population's salaries and income*. Warszawa: PWE.

Nadarajah, S., & Kotz, S. (2006). Compound mixed Poisson distributions I. *Scandinavian Actuarial Journal*, *3*, 141-162.

Nadarajah, S., & Kotz, S. (2006). Compound mixed Poisson distributions II. *Scandinavian Actuarial Journal*, *3*, 163-181.

Panjer, H., & Willmot, G. (1992). *Insurance risk models*. Schaumburg: Society of Actuaries.

Szymańska, A. (2013). *The use of asymmetric loss function for estimating premium rates in motor insurance*. In Papież M. & Śmiech S. (Eds.), *Proceedings of 7th Professor Aleksander Zelias International Conference on Modelling and Forecasting of Socio-Economic Phenomena*. Cracow University of Economics, Poland, 181-189.

Tremblay, L. (1992). Using the Poisson inverse Gaussian in bonus-malus systems. *ASTIN Bulletin*, *22*(1), 97-106.