QUANTITATIVE METHODS IN ECONOMICS
Vol. XV, No. 2, 2014, pp. 242 – 251

DISTRIBUTIONS OF THE NUMBER OF CLAIMS BY AGE GROUPS OF INSURED IN CIVIL LIABILITY MOTOR INSURANCE PORTFOLIO

Anna Szymańska
Department of Statistical Methods, University of Lodz
e-mail: szymanska@uni.lodz.pl

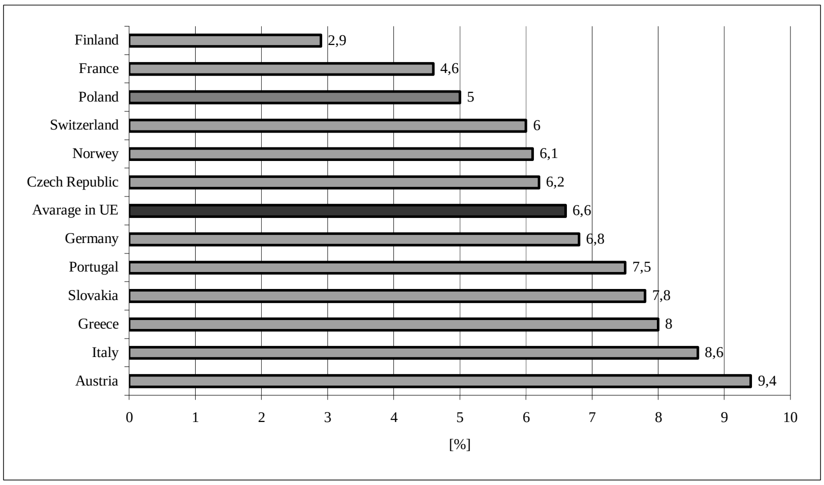
Abstract: In the process of calculation of the insurance premium in civil liability motor insurance the knowledge of distribution of the number and value of paid claims is required. The paper presents the methods of assessing the degree of the fit of theoretical distributions of the number of claims 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

INTRODUCTION

Civil liability motor insurance of vehicle owners are most often concluded insurance in our country. Polish motor insurance market, in comparison to other European countries, has one of the lowest frequency of claims (cf. Figure 1).

Figure 1. Frequency of the damage in chosen European countries in 2007



Source: Europejski Rynek Ubezpieczeń Komunikacyjnych, PIU, Warszawa 2010

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. In civil liability motor insurance ratemaking is a two-step process [Antonio et al. 2012]. In the first stage - called *a piori* - the base premium based on known risk factors is determined. Then in the base contribution discounts and increases, mainly resulting from the mileage of claims are included in the previous insurance period (bonus-malus system) and of such factors like e.g. age of insured, the period of holding the driving license, or the age of the car. This stage is called *a posteriori* ratemaking, and its' result is an assigned premium

Both tariffs, *a priori*, and *a posteriori* require from the actuary the determination of distributions of theoretical random variables describing the number and value of paid claims. 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 – Kołmogorow [Panjer et al. 1992, Domański 1990]. 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, index of distribution similarity, ratio of the maximum difference of relative frequencies, ratio of the maximum difference of cummulative distribution functions [Kordos 1973].

The study analyzed the distribution of the number of claims in civil liability motor insurance portfolio of passenger cars of individuals of one of the insurance companies operating on the Polish market in 2006. In the audited insurance company base premium is determined based on two factors: region of registration of the vehicle and engine capacity. One of the factors of a posteriori tariff is the age of the driver, the insured persons under the age of 25 years are charged by increase of the basic premium of 100% to 200%. Insured from 25 to 28 years of age have increase in the amount of 30% to 70% of the basic premium. Height of the increase depends on many factors, such as whether the insurance contract is concluded with the company for the first time, or is a continuation of previous insurance, if the insured has another vehicle insured in this company and whether he continues insurance without claims.

The aim of the study is to assess the legitimacy of used in studied insurance company premium increases by virtue of age of the insured based on an analysis of distributions of the number of claims in particular age groups of drivers. In the paper 5000 attempt drawn from the portfolio of the civil liability motor insurance of one of insurance companies was disposed.

It should be highlighted that in Poland access to individual data is very difficult and insurers are reluctant to provide them. Therefore, in the study the name of the insurance company was not given.

THE CHOICE OF THE DISTRIBUTION OF THE NUMBER OF CLAIMS IN THE CIVIL LIABILITY MOTOR INSURANCE

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].

Definition: Let p_k be the probability function of a discrete random variable. It is a member of the (*a*,*b*,0) class of distributions, provided that there exists constants *a* and *b* such that:

$$\frac{p_k}{p_{k-1}} = a + \frac{b}{k}. \tag{1}$$

Lemma: The only members of the class of discrete distributions on the non-negative integers satisfying (1) are the Poisson, binomial, negative binomial and geometric distributions.

Let the random variable *X* represent the number of claims from individual policy or a policy portfolio. According to the paper [Panjer et al. 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:

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

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: $\overline{X} = M_1$; $S^2 = M_2 - M_1^2$; $K = M_3 - 3M_2M_1 + 2M_1^3$.

For class (a, b,0) and since we might expect $\frac{N_k}{n}$ to be close p_k, it follows that:

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

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 [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.

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:

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

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:

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

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:

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

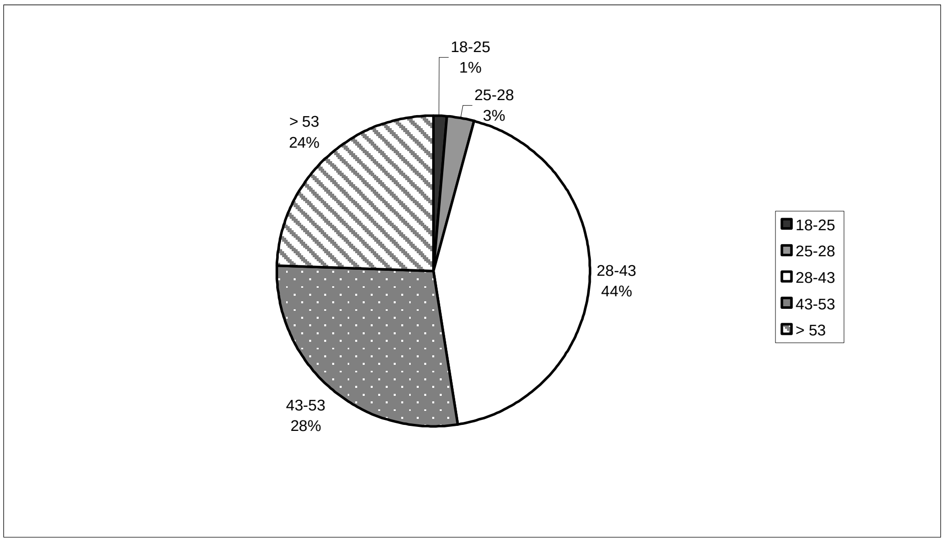
where: $F_i = \sum_{j=1}^{i} \gamma_j$ - value of the empircial cummulative distribution function, $\hat{F}_i = \sum_{j=1}^{i} \hat{\gamma}_i$ - value of the theoretical cummulative distribution function. This ratio is equal to zero for fully consistent distributions.

EMPIRICAL EXAMPLES

In this part of the study the distribution of the number of claims paid in civil liability motor insurance portfolio and in separate age groups of drivers of analyzed insurance company in 2006 was investigated.

Figure 2 shows the structure of according to the age of the insured of the civil liability motor insurance portfolio of the analyzed insurance company.

Figure 2. Portfolio structure by age of the insured in 2006



Source: own calculations

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).

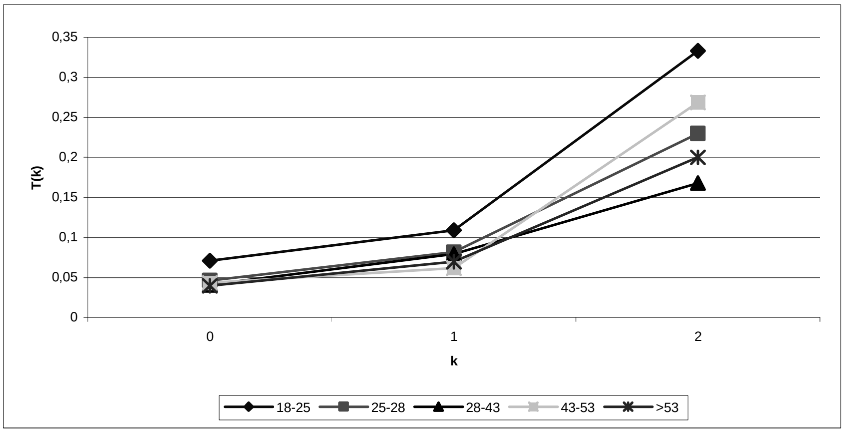
Table 1. Numerical characteristics of the empirical distributions of the number of claims in the age groups of the insured in civil liability motor insurance portfolio insurance company in 2006

Numerical characteristics of distributions	portfolio	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

Source: own calculations

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 Table 2). 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 Figure 3), 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.

Figure 3. The frequency functions in the age groups of insured



Source: own calculations

Table 2. Measures of the degree of distributions' fitting

portfolio				
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
age group:18-25				
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
age group: 25-28				
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
age group: 28-43				
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
age group: 43-53				
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 2. cont.

age group: >53				
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

Source: own calculations

Analyzing the values of the measures of fit from Table 2 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[1] with parameters $\alpha = 1.1227; \beta = 27.5032$ was rated (see Table 3).

Table 3. 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

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

Source: own calculations

Analyzing the results presented in Table 3, 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. Fit of distributions of is unsatisfactory in the group of insured who are under the age of

[1] negative binomial distribution: $P(X=k) = \frac{\Gamma(\alpha+k)}{\Gamma(\alpha)k!}\left(\frac{\beta}{1+\beta}\right)^{\alpha}\left(\frac{1}{1+\beta}\right)^{k}$

25 years. Perhaps it is the result of a small number of policies in this age group (50 policies). Analysis should be performed on larger sample.

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. Distribution of the number of claims in all groups of insured civil liability motor insurance portfolio of the studied insurance company is negative binomial.

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. Distribution of the number of claims in a group of drivers from 25 to 28 years old does not confirm the need to use such large increases of premiums to them. 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.

REFERENCES

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

Domański Cz. (1990) Testy statystyczne, PWE, Warszawa.

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

Klugman S. A., Panjer H.H., Willmot G.E. (2004) Loss Models. From Data to Decisions, J. Wiley &Sons, New York.

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

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

Tremblay L. (1992) Using the Poisson Inverse Gaussian in Bonus-Malus Systems, ASTIN Bulletin 22(1), str. 97-106.