Modelling of claim counts as a base for premium system determination

Ivana Simeunovic[1], Ivana Domazet[2], Hasan Hanic[3]

[1] Belgrade Banking Academy, Union University, Belgrade, Serbia,
[2] Institute of Economic Sciences, Belgrade, Serbia,
[3] Belgrade Banking Academy, Union University, Belgrade, Serbia.

Abstract

The aim of this paper is the analysis of the problem of modelling of claim counts in insurance that implies the study of variations of their occurance through finding out the distribution which fits the observed data most adequately. In that sense, the most important aspects in the process of choosing the probability of claim numbers have been studied on a chosen sample from a Serbian insurance company and it has been found that appropriate sample analysis that was based upon the study of the previous experience of the insured was one of the key elements from the point of view of determining adequate premium systems.

Key words: Claim Frequency, Probability Distribution Functions, Determining Premium, Poisson-Gamma Distribution

1. Introduction

In order to decide on covering certain risks, insurance companies carry out appropriate analysis of all risk factors. The classification of risk factors according to their importance brings us to the second phase that is prior to the final calculation of insurance premium – the phase of determining the size of selected risks - the process of risk quantification. Statistical estimation of risks is based upon the analysis of two key measurements – frequency and the amount of expected claims. Modelling claim counts and claim severities are some of the most important pre-conditions for adequate premium system determination [1].

In this paper the problem of modelling of number of claims will be analysed, as well as the problems of implementation of the mentioned process in practice. Serbian insurer's data on claims of the 95.800 insured portfolio of automobile third-party liability insurance have been used as the base of analysis. The probability distribution functions which are commonly used in this procedure have been described, whereby the overall analysis on a selected sample has been performed. Specific attention is given to the model known as a good risk/ bad risk model. The last part of this paper represents the most important conclusions.

2. Actuarial modelling of claim counts

Losses in insurance are happening by chance, which is why it is not possible to anticipate the exact time of their occurance, as well as their total number and the amount [2]. The process of determining premium rates means previously completed analysis of observed claims as well as finding distribution that can define frequency and amount of claims. Longer periods of risk analysis present a base for properly chosen distribution function and that is one of the biggest problems that insurance companies are facing with. Since there is not enough information from previous period, the analysis need to be based on the observance of homogenic risks or extrapolation of smaller risks. Also, since the most common case in practice is that distribution of claims is not known in advance, the initial assupmtion in solving this problem is that the unknown distribution is a member of a family. In this way, the task becomes to estimate the parameters of the chosen family using the data on frequency and amounts of claims. When the estimation of unknown parameters of distribution is done, then we test the goodness of fit of realised frequencies to the distribution assumed. The most widely used statistical test to evaluate the goodness of a fit is the χ^2 test [3].

Three basic approaches used for the modelling of the claim counts are: empirical, analytical and the moments based methods. In the case of the existence of bases containing the large number of data to run smooth and accurate assessment of the

cumulative distribution function it is possible to use empirical method. The moments based method comprises evaluation of the moments of distribution, usually the mean and variance. The most widely accepted in the actuarial literature and in practice is the analytical method that involves finding an appropriate analytical expression that can describe the observed data [4]. In the following, certain distributions from the aspect of their proper selection of modelling claim frequencies will be analyzed.

Poisson distribution occupies a central possition in discrete distribution theory that is used for describing events that occur randomly and independatly in space or time, i.e. the number of claims. Its application is justified for mass events with very low probability. Unlike binomial distribution, the random value of Poisson type can take an unlimited number of values. The random variable $N(t, t+\Delta t)$ that describes the number of claims in a given time interval can be presented as Poisson random variable, whose probability mass function is:

$$p_k = P(N = k) = e^{-\theta}\frac{\theta^k}{k!},\ \theta > 0$$

Parameter θ of Poisson distribution presents both mean and a variance of distribution at the same time, where θ is equal to the average number of occurrences of a certain event (claim) in a unit of time.

$$E(N) = Var(N) = \theta$$

One of the important features of Poisson distribution is the assumption on homogeneity of the population which is the subject of a certain analysis. However, it is not a real assumption when it comes to modelling of certain variables in the field of insurance. This fact is well recognized in the case of modelling claims couts in automobile third-party liability insurance, where we have especially emphasized heterogeneity of portfolio observed [5], ie. difference in the behaviour of the drivers who are insured. In these circumstances, the value of the variance of the observed variable of Poisson type will be higher than its mean value and it will be necessary to use other, especially compound distributions.

In order to prove the stated fact, we are going to analyse a Serbian insurance company and its chosen portfolio of automobile third-party liability insurance for which we will use the assumption of homogeneity, as well as the assumption that the number of claims of each insured of the observed homogeneous sample is a random value that can be approximated by Poisson distribution whose unknown parameter equals θ. On the basis of the given portfolio we will estimate an unknown parameter of distribution and then test whether empirical data fit Poisson (assumed) distribution. Table 1 shows the distribution of the number of claims in the automobile third-party liability portfolio which contains n=95.800 observations.

Table 1. Observed distibution of number of claims in a portfolio

Number of claims (X)	Number of the insured (f_i)
0	88.035
1	7.117
2	591
3	52
4	5
≥ 5	0
$\sum$	**95.800**

In the following part, it will be tested whether the given data adjust to the assumed, i.e. Poisson distribution. In other words, we are going to test the following statistical hypotheses:

H_0: the number of claims per automobile third-party liability insurance policy is adjusted to Poisson distribution.

H_1: the observed distribution of number of claims does not adjust to the assumed distribution.

The mean of the sample is equal to $\bar{x} = \frac{1}{n}\cdot\sum_{i=1}^{n} X_i \cdot f_i = 0.088465$, while the variance of the observed sample is equal to $s^2 = \frac{1}{n-1}\cdot\sum_{i=1}^{k} f_i \cdot (X_i - \bar{x})^2 = 0.096862$. An unknown parameter of distribution is estimated by the moments method and it is equal to $\hat{\theta} = 0.088466$.

Hereafter, we will carry out the testing by the implementation of χ^2 goodnes-of-fit test. Appli-

cation of this test is based on the rule that all expected frequencies belong to five of more grouping procedures, with the level of significance α = 0.05.

In order to determine expected frequencies, first it is necessary to calculate the probabilities of Poisson type for each of the values of a random variable marked by: X - number of claims. Then, we will multiply the probabilities by the sample size and get fitted frequencies presented in the following table. Thus, we have:

$$f_i' = P\left(X = i \middle| H_0\right) \cdot n = \frac{\hat{\theta}^i}{i!} \cdot e^{-\hat{\theta}} \cdot n,$$
$$i = 0,1,2,3,4,5$$

From the χ^2 distribution, $\chi_{3;0,95}{}^2 = 7.815$, according to $\chi^2 = 508.58 > 7.815 = \chi^2_{3;0..95}$, we can conclude that null hypothesis was rejected, so the given claim frequency in automobile third-party liability insurance portfolio cannot adjust to Poisson distribution. That is especially true for right tail of the observed portfolio, which is why the data should be modeled by the distribution whose variance exceeds its mean.

The hypothesis on homogeneity of the observed sample of the insured is not in accordance with the statistical analysis. The rejection is interpreted as the sign that the portfolio is heterogeneous. This means that it is not justified to use the same tariff system for all the insured of the observed portfolio. The analysis should include the data on „behaviour" of the insured in the past [6]. The mentioned example shows the fact that behaviour of the insured in a portfolio differs, and shows heterogeneity of the group which is why it is needed to find a model that will express the heterogeneity. This will consequently produce different tariff system.

With that aim in mind, we assume that the frequency of claims of every single insured of automobile third-party liability portfolio can be approximated by Poisson distribution, while the parameter θ of Poisson distribution takes different values. Thus, each insured is characterized by a certain value of the parameter, which means that the behaviour of each insured is presented by the realised value θ of a random variable Θ. In that way, we get the expression which will represent the distribution of the total number of claims of the observed portfolio:

$$p_k = \int_0^\infty e^{-\theta} \cdot \frac{\theta^k}{k!} g(\theta) d\theta, \quad k = 0,1,2,...$$

where $g(\theta)$ is the density function of a random variable Θ. Previous expression is also called *mixed Poisson distribution*. We can further assume that the parameter Θ of Poisson distribution follows Gamma distribution whose parameters are a and τ - $\Theta : \Gamma(a,\tau)$:

$$g(\theta) = \frac{\tau^a e^{-\tau \cdot \theta} \theta^{a-1}}{\Gamma(a)}, \quad a, \tau > 0$$

The resulting distribution of the number of claims in the portfolio known as a *negative binomial distribution* then have the following form [7]:

$$p_k = \binom{k+a-1}{k} \left(\frac{\tau}{1+\tau}\right)^a \left(\frac{1}{1+\tau}\right)^k = \binom{k+a-1}{k} p^a q^k$$

Table 2. Observed and fitted distribution of number of claims – Poisson distribution, the method of moments

Number of claims: X	Empirical frequencies (f_i)	Expected frequencies (f_i')	$\frac{(f_i - f_i')}{f_i'}$
0	88.035	87.689.02	1.365080681
1	7.117	7.757.50	52.8825403
2	591	343.14	179.0416857
3	52	10.12	173.3476413
4	5	0.22	101.9359555
≥5	0	0.00	0.003959551
Σ	**95.800**	**95.800**	**508.576863**

and whose parameters are:

$$E(X)=\frac{a}{\tau} \quad \text{and} \quad \sigma^2=\frac{a}{\tau}\left(1+\frac{1}{\tau}\right)$$

Hereupon, we can conclude that the value of a variance of the random variable which is adjusted to this distribution is higher than its mean value. The stated characteristic is especially important in the case of the sample analysis which contain the units coming from heterogeneous population [8], as it is the case in automobile third-party liability insurance where insured individuals show a constant, but each different tendency (ie. probability) that they will suffer a claim. It is why this distribution could be an adequate choice for analysed random value modelling. Therefore, we will carry out a testing of the adjustment of the observed data to Poisson-Gamma (negative binomial) distribution using the χ^2 goodnes-of-fit test.

Finally, using the critical value of the χ^2 distribution we have that $\chi_{2;0,95}{}^2 = 5.991$, wherefore from $\chi^2 = 0.57 < 5.991 = \chi^2{}_{2;0..95}$ we can conclude that null hypothesis can be sustained with the level of significance $\alpha = 0{,}05$. In other words, the Poisson-Gamma model can be applied to this automobile portfolio.

Binomial distribution presents another type of distribution for the modelling of claim counts. This is especially true in cases in which the stated random value number of claims has the mean value which is higher than its variance. As we could see, it is not in accordance with any of described distributions so far. The resulting number of values that binomial random variable can take stands as another important feature of this distribution. It can be especially useful when, for instance, the number of traffic accidents per an insured per year is modelled. In the mentioned case it is fully justified to set the upper limit of the modality of random variable defined as the number of claims, since it is physically impossible to associate positive probabilities to individual values beyond the previously set limit. For instance, the mentioned upper limit in this example can be 12, so the probability of the frequency of claims beyond this limit would be considered insignificant. Normal and Poisson distribution stand out as the two most important approximations of binomial distribution. In the first case, when the sample size of the sample is large enough and when observed frequency does not show important skewness binomial distribution is well approximated by normal distribution. On the other hand, when the number of observations is very large and probability of success extremely low, then the approximation that matches binomial is Poisson distribution. From this reason Poisson distribution is often called the law of small numbers [9], which can be understood easily through the occurance of claim in automobile third-party liability insurance: the probability of its occurance is huge, due to the great number of traffic participants, while, on the other hand, the probability of traffic accidents is still low.

Finally, we will mention some other compound (mixed) distributions which are gaining more and more important role in contemporary literature on claim modelling in automobile insurance.

Poisson-Inverse Gaussian distribution - $X : IGau(\mu, \beta)$ if its probability density function could be presented by the following expression [10]:

Table 3. Observed and expected distribution of the number of claims: Poisson-Gamma distribution, the method of moments

Number of claims: X	Empirical frequencies (f_i)	Expected frequencies (f_i')	$\frac{(f_i - f_i')}{f_i'}$
0	88.035	88.036.12	0.000014196
1	7.117	7.113.07	0.002169728
2	591	595.65	0.036225995
3	52	50.46	0.04679922
4	5	4.30	0.113951054
≥ 5	0	0.37	0.367671729
Σ	**95.800**	**95.799.97**	**0.566831922**

$$f(x)=\frac{\mu}{\sqrt{2\pi\beta x^3}}\mathrm{e}^{\left(-\frac{1}{2\beta x}(x-\mu)^2\right)},\ x>0$$

The expected value and variance of a random variable of the mentioned type are equal to:

$$E[X]=\mu \qquad Var(X)=\mu\beta$$

With introduced assumption that an unknown parameter of Poisson distribution Θ follows inverse Gaussian distribution, we will assume that $E(\Theta)=1$, since we want to find the average claim frequency in a portfolio. Thus, from $\Theta : IGau(1,\beta)$:

$$f_\Theta(\theta)=\frac{1}{\sqrt{2\pi\beta\theta^3}}\mathrm{e}^{\left(-\frac{1}{2\beta\theta}(\theta-1)^2\right)}$$

we get the resulting distribution of the number of claims in the porfolio:

$$p_k=\int_0^\infty e^{-\theta}\cdot\frac{\theta^k}{k!}\frac{1}{\sqrt{2\pi\beta\theta^3}}e^{\left(-\frac{1}{2\beta\theta}(\theta-1)^2\right)}d\theta,$$

$$k=0,1,2,...$$

Inverse Gaussian distribution is an excellent choice in modelling of values which take exclusively positive values of those that show right skewness, which is characteristics of the claim frequency in automobile third-party liability insurance.

Poisson-LogNormal distribution - $X : lNor(\mu,\sigma^2)$, if the variable $\ln X$ is normally distributed - $\ln X : Nor(\mu,\sigma^2)$, and whose probability density function is given by the following expression [11]:

$$f(x)=\frac{1}{\sqrt{2\pi}x\sigma}\mathrm{e}^{\left(-\frac{1}{2\sigma^2}(\ln x-\mu)^2\right)},\ x>0$$

The mean and the variance of a random variable which is lognormally distributed are equal to:

$$E[X]=e^{\left(\mu+\frac{\sigma^2}{2}\right)}$$

$$Var(X)=\exp(2\mu+\sigma^2)\left(\exp(\sigma^2)-1\right)$$

Finally, if we put $\mu=-\frac{\sigma^2}{2}$ in order to provide $E(\Theta)=1$, we get the resulting distibution of the number of claims in the portfolio:

$$P(N=k)=p_k=\frac{1}{\sigma\sqrt{2\pi}}\int_0^\infty\frac{\theta^{k-1}}{k!}e^{-\theta}\cdot e^{\left(-\frac{\left(\ln\theta+\sigma^2/2\right)^2}{2\sigma^2}\right)}d\theta$$

The described mixed distribution can be successfully implemented in such cases where analysed data show exquisite skewness.

3. Good risk/bad risk model: the main assumptions and implementation

For the modelling of claim counts in automobile third-party liability insurance, among the models that have been derived from the elements of Poisson processes which are successfully implemented in certain cases, there is a model well-known under the name: *Good risk/bad risk (good driver/bad driver) model.*

This simple model [8] is based on the assumption that all drivers (insured) of a portfolio in automobile third-party liability insurance can be divided in two categories: "good drivers" and "bad drivers". Each of the mentioned categories of drivers experiences different number of claims that can be approximated by Poisson distribution. If we mark a parameter of Poisson distribution - a random value of the number of claims for good drivers with θ_1 and the matching parameter for bad drivers with θ_2, then the resulting distribution of the number of claims in the portfolio can be presented by the following expression:

$$p_k=a\cdot\theta_1{}^k\frac{e^{-\theta_1}}{k!}+b\cdot\theta_2{}^k\frac{e^{-\theta_2}}{k!}$$

where a and b are relative frequencies of good and bad insured drivers in a portfolio, respectively, which is why it is clear that: $b=1-a$.

Also: $a,b,\theta_1,\theta_2>0$.

The mean and the variance of a random variable that can be approximated by the mentioned distribution of Poisson type are presented by the following expressions:

$$\mu = a \cdot \theta_1 + b \cdot \theta_2$$

$$\sigma^2 = a \cdot \theta_1^2 + a \cdot \theta_1 + b \cdot \theta_2^2 + b \cdot \theta_2$$

Let us apply the model that we described on studied portfolio of the insured in automobile third-party liability insurance, noting that the observed insured portfolio can be classified into two categories - 65% good and 35% bad drivers, using the moments method. The probability that a good driver would report k claims is one of the Poisson type whose parameter of distribution is $\theta_1 = 0.04$, and the probability that a driver from the category „bad drivers" report k claims is of the same type, but the parameter of a distribution is $\theta_2 = 0.13$. For the insured where there is no data on previous claims, the insurer is not able to decide which category they belong to, so the expected total number of claims of this portfolio according to good risk/bad risk model will be:

$$n \cdot \left(0.65 \cdot 0.04 + 0.35 \cdot 0.13\right) = 0.0715n$$

where n is the number of insured drivers of the observed portfolio.

In the following time interval, where the data on reported claims are available, it is possible to determine the probability that certain insured will belong to one or another category of drivers according to his/her number of previous claims. Namely, by Bayes' theorem on conditional probability, we will have:

$$P\left[good \middle| (k\_claims)\right] =$$

$$= \frac{P\left[((k\_claims)) \middle| good\right] \cdot P[good]}{P\left[(k\_claims) \middle| good\right] \cdot P[good] + P\left[(k\_claims) \middle| bad\right] \cdot P[bad]}$$

$$= \frac{e^{-\theta_1} \cdot \frac{\theta_1^k}{k!} \cdot P[good]}{e^{-\theta_1} \cdot \frac{\theta_1^k}{k!} \cdot P[good] + e^{-\theta_2} \cdot \frac{\theta_2^k}{k!} \cdot P[bad]} =$$

$$= \frac{e^{-\theta_1} \cdot \theta_1^k \cdot P[good]}{e^{-\theta_1} \cdot \theta_1^k \cdot P[good] + e^{-\theta_2} \cdot \theta_2^k \cdot P[bad]}$$

The upper line shows the probability that the insured will belong to the category of good drivers under condition that during the first year he/she reported k claims. Obviously, the mentioned probabilities are decreasing with the number of claims reported. In that way, the expression for the expected number of claims for the following (second) year becomes:

$$n \cdot \left(P\left(good \middle| (k\_claims)\right) \cdot \theta_1 + P\left(bad \middle| (k\_claims)\right) \cdot \theta_2\right)$$

Finally, by including $n = 95.800$, the expected number of claims in the following year according to reported claims, calculated by the studied model is given by the table 4.

4. Conclusion

The frequency of claims presents a random variable of a discrete type. Choosing the appropriate probability distribution function in modelling the number of claims can be defined as a statistical problem and its solving is of an extreme benefit

Table 4. The expected number of claims in the following year in the good risk/bad risk model given the number k of claims reported during the first year

The number of claims reported during the first year-k	$P\left[good \middle| (k\_claims)\right]$	$P\left[bad \middle| (k\_claims)\right]$	Expected number of claims in the following year
0	0,6702	0,3289	6.664,33
1	0,3847	0,6153	9.137,12
2	0,1923	0,8077	10.795,99
3	0,0559	0,9441	11.972,03
4	0,0179	0,9821	12.299,67
5	0,0055	0,9945	12.406,58

in determining the premium rate for each of the insurer. The base for carrying it out is the process of modelling of two random variables, which are claim frequency and claim severity. Regarding the complexity of the issue, in this paper we have especially analysed the frequency of claims. Also, the attention was focused on one of the types of non-life insurance – automobile third-party liability insurance which is one of the most frequent type of insurance throughout the world. We used the original data provided from the portfolio of the largest automobile third-party liability insurer in Serbia, which accounts for almost 25% of the total population of the automobile third-party liability in this country and came to the following important conclusions:

- Due to the great number of distribution functions that can be used for modelling of claims couts, first it is necessary to limit their number to several functions.
- In the area of automobile third-party liability insurance that has been especially analysed and for which can be said that it is the typical example of heterogeneous population of the insured, in the process of the modelling of claim counts it is not often appropriate to use Poisson distribution, but certain distributions derived from Poisson, especially Poisson-Gamma, as well as Poisson Inverse Gaussian.
- One of especially appropriate models that can be used for the modelling of claim counts in automobile third-party liability insurance is a good risk/bad risk model. By implementation of described model it is possible to predict the expected number of claims that will occur in the following year, which makes an excellent base point in the process of setting the rates as the most important task of actuarial work.

Finally, we note that the modeling claim frequency may be additionally improved by introducing additional variables into the analysis. Also, on the basis of conducted research, and by using results obtained, the determining of premium rates procedure begins. The described procedure based on the analysis of past experience of the insured is known as Bonus-Malus System and it is an integral part of tariffs of almost all automobile liability insurers in the world. Through its application such tariff system that punishes drivers who are responsible for the occurrence of the insured event is provided, which allows increasing of road safety and ensures principle of fairness.

Due to the fact that the data mentioned could not be implemented into the model presented, the authors of this paper are intending to improve the problem of modelling of claim counts together with tariff process as described above.

Acknowledgements

This paper is a part of research projects number 47009 (European integrations and social and economic changes in Serbian economy on the way to the EU) and 179015 (Challenges and prospects of structural changes in Serbia: Strategic directions for economic development and harmonization with EU requirements), financed by the Ministry of Education and Science of the Republic of Serbia.

References

1. *Simeunović I. Statistical-actuarial basis and solving the problems in the process of premium of third party liability insurance determination. Belgrade Banking Academy - Faculty for Banking, Insurance and Finance, Union University. Belgrade. 2010.*
2. *Beard RE, Pentinainen T, Pesonen E. Risk Theory: The Stohastic Bases of Insurance. Chapman and Hall. London. 1997.*
3. *Besson JL, Partrat C. Trend et systemes de bonus-malus, ASTIN Bulletin 1992; 22: 11-31.*
4. *Klugman SA, Panjer HH, Wilmott GE. Loss Models - From data to decisions. Wiley. New York. 1998.*
5. *Lemaire J. Automobile Insurance: Actuarial Models. Kluwer-Nighoff Publishing. Boston. 1985.*
6. *Pitrebois S, Denuit M, Walhin JF. Multi-Event Bonus-Malus Scales. The Journal of Risk and Insurance 2006; 73(3): 517-528.*
7. *Nadarajah S, Kotz S. Compound mixed Poisson distributions II. Scandinavian Actuarial Journal 2006; (3): 163-181.*
8. *Lemaire J. Bonus-Malus Systems in Automobile Insurance. Kluwer Academic Publishers. Boston. 1995.*

9. *Denuit M, et al. Actuarial Modelling of Claim Count: Risk Classification. Credibility and Bonus-malus systems. Wiley. New York. 2007.*

10. *Hossak IB, Polard JH, Zehenwirt B. Introductory statistics with applications in general insurance. Cambridge University Press. United Kingdom.1999.*

11. *Hogg RV, Klugman SA. Loss Distributions. Wiley. New York. 1984.*

Corresponding author:
Hasan Hanic,
Belgrade Banking Academy,
Faculty for Banking, Insurance and Finance,
Union University,
Belgrade,
Serbia,
E-mail: hasan.hanic@bba.edu.rs