COLLECTIVE RISK MODEL

DOMEN GAL ŠKERLEP

mentor: prof. Bojana Dvoržak
Gimnazija Bežigrad

One of the main tasks of actuaries is to make the best possible predictions for the sum of insurance claim amounts in a specific time interval. There are several possible approaches that all to a certain extent give accurate values. One of them is the collective risk model, which is defined in this article. Using Wald's Identity and linear regression the expected sum of insurance claims for helicopter rescuing in one year is calculated.

KOLEKTIVNI MODEL TVEGANJA

Ena izmed temeljnih nalog aktuarjev je čim boljša ocena škode zavarovalnice iz naslova zahtevkov za izplačilo zavarovalnine v določenem časovnem intervalu. V ta namen se med drugim uporablja tudi kolektivni model tveganja, ki je definiran v članku. Z uporabo Waldove identitete in linearne regresije je iz ocenjenih porazdelitev izračunana pričakovana vrednost vsote zahtevkov za izplačilo stroškov helikopterskih prevozov v enem letu.

1. Introduction

After having a close frustrating experience of seeing injured skiers being transported to the hospital from the ski slopes with a helicopter, I started to question, who covers the costs of such emergency treatment, as those costs are quite high. In Slovenia the government covers the payments of all helicopter rescues. However, when I researched further and wanted to check how the neighbouring countries, where we were skiing, manage this solution, I came across the involvement of insurance companies. As I explored this topic further, I was suddenly exposed to many different packages offered by them. These differed in the amount of monthly payments as well as the values of possible insurance claims. I knew the basic policy behind insurance contracts, yet I did not know how they manage to have enough money that they are able to pay all the claims in case of rapid growth of demands. As stated above, my main focus was how insurance companies manage to maintain solvency (i.e. the ability of a company to meet its long-term debts and financial obligations [1]) and what measures do they take to make a profit. This was the question that none of my family members knew the answer to, therefore I was thrilled to discover the mathematical background of insurance contracts. I expected that I will have to broadly expand my knowledge of probability and statistics, however, the wish to understand this puzzle kept me going forward and discovering topics greatly beyond the mathematical syllabus.

Insurance companies are obliged to quantify and minimize risk resulting from their activities (see regulation of Solvency II regulation from European Union, which came into effect on 1 January 2016) [2]. To analyse this topic I needed to collect some real-life data. Since the number and values of insurance claims are not openly published for our specific case, I wrote to my insurance company, if they can provide me with such data. They agreed to send me their data statistics under the condition, that this data will be used only for the purpose of this research and their name will not be explicitly mentioned in the article due to protection of personal data of the clients and company's privacy policy. The only thing that I am allowed to mention is that they are one of the three biggest Slovenian insurance companies, offering health insurance which includes helicopter transportation. Furthermore, I needed to establish a narrow focus on one single type of contract in order to be able to create objective outcomes. In connection with my initial interest in helicopter rescues, I decided to look at the health insurance claims that include helicopter transportation and were made in the year 2019.

After obtaining the data, I started with research on how to predict the money that insurance company needs at the beginning of the year to remain solvent throughout the year. In order to make the best possible approximation for the expected value of the total amount of claims in the year 2020, we need to apply the concept of the collective risk model, which is used by insurance companies to predict the total amount of insurance claims from their clients in the short term (typically in the time frame of one year). A non-life insurance company aims to determine the probability of damage and the average amount of damage as accurately and credibly as possible as a basis for the best possible assessment of total (aggregate) claims relating to a particular risk group over a period of one year. In this article, the collective risk model will be mathematically formulated and applied in solving the real-world problem using the data from the Slovenian insurance company.

Therefore, the aim of this article is to predict how much money an insurance company needs so that the probability of insolvency in the following year is small. The prediction will be based on the expected value of an individual claim amount and the expected value of the number of claims made in one year.

2. Theoretical model

Before we begin, let us recall key definitions in probability theory. A random variable is a quantity whose possible values are numerical outcomes of a random phenomenon. It can be either discrete or continuous. A discrete random variable can only take a countable number of values, while a continuous one can take any real number within the range of realistic possible values [3, 4].

Definition 1 ([4, 5]). The distribution of a discrete random variable N which takes values $x_1, x_2, \ldots, x_n$, $n \in \mathbb{N}$ is given by the probability mass function

$$P_n(x_i) = P(N = x_i),$$

whereas in the case of a continuous random variable X the probability density function $f_N(t)$ is defined as a non-negative function whose area under the graph is equal to 1.

Definition 2 ([5]). The cumulative distribution function $F(x)$ of a discrete random variable N is a function whose value equals to the probability of N taking a value smaller than $x \in \mathbb{R}$ i.e.

$$F_N(x) = \sum_{x_i \leq x} P(N = x_i).$$

In case of a continuous random variable X the cumulative distribution function is given by

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f_Y(t) \cdot \mathrm{d}t.$$

As already stated in the introduction, our main aim will be to calculate the expected value of a claim amount.

Definition 3 ([6]). For a discrete random variable, N which takes values $x_1, x_2, \ldots, x_n$, $n \in \mathbb{N}$ the expected value is defined as

$$E(N) = \sum_{i=1}^{n} x_i \cdot P\left(N = x_i\right),$$

and for a continuous random variable X as

$$E(X) = \int_{-\infty}^{\infty} x \cdot f_x(t) \cdot \mathrm{d}t.$$

The purpose of the above definitions is to be equipped with the necessary terminology that will be used throughout the article. The notation for random variables used above is general and does not have any correlation with the random variables used in our collective risk model. Now that we are equipped with the necessary terminology, let us begin by stating the formal definition of the model that we said will be our main approach to the initial aim of this article.

Definition 4 ([7]). Assume that N is a discrete random variable with range in $\mathbb{N}_0$, presenting the number of claims by customers. Further let us presume that there exists a sequence of independent, identically distributed (i.i.d.), non-negative continuous random variables $(X_i)_{i\in\mathbb{N}}$, where X_i represents the value of the i-th claim. Let X_i be independent of N for all $i \in \mathbb{N}$. Then we define the total claim amount S as

$$S = X_1 + X_2 + \cdots + X_N = \sum_{i=1}^{N} X_i.$$

In case of an empty sum (i.e. $N = 0$), we set the value of S to 0. For the purpose of this theoretical model, we assume that $(X_i)_{i\in\mathbb{N}}$ can take any real value, so that we can use continuous distributions.

If we knew the distribution, we would be able to determine the expected value of claims in the following year. This would give us the prediction for the total amount of claims, and we could finish at this point. However, we cannot directly know which distribution is applicable to S since the value of S is dependent on two sources of randomness – random variables N and X_i. Now we have come to the main question that actuarial mathematicians face – how to get the distribution of the random variable S. Evaluating the expected value of S is therefore exactly what our aim is.

We have to approach the problem using Wald's Identity, which simplifies evaluating the expected value of the sum of a random number of random quantities [8]. Mathematically, this means that we can calculate the value of $E[S]$ just by knowing the distributions of random variables N and X_i. Knowing the distribution of X_1 is sufficient to determine the distribution for all X_i, $i = 1, 2, \ldots, N$. This comes from the fact that X_i, $i = 1, 2, \ldots, N$ are identically distributed random variables.

Let us now formally define the basic version of Wald's Identity, which will be sufficient for the needs of this article.

3. Wald's Identity

Theorem 1 ([9]). *Assume that N is a random variable with range in $\mathbb{N}_0$ and that there exists a sequence of independent, identically distributed (i.i.d.), non-negative random variables $(X_i)_{i\in\mathbb{N}}$. Further, let X_i be independent of N for all $i \in \mathbb{N}$, let $E[X_i] < \infty$ for all $i \in \mathbb{N}$ and let $E[N] < \infty$(i.e. X_i, $i \in \mathbb{N}$ and N have finite expected values). Then*

$$E[S] = E\left[X_1 + X_2 + \cdots + X_N\right] = E\left[X_i\right] \cdot E[N].$$

Simply, by knowing the distributions of N for X_i for any $i \in \mathbb{N}$, we can compute the desired value. As stated, all random variables X_i have the same distribution, therefore the value of $E[X_i]$ is equal for all $i = 1, 2, \ldots, N$. We will now focus on the distribution of X_i and then move on to evaluating the distribution of N.

4. Distribution of X

From data received from the insurance company, we will focus on claim amounts resulting from accidents involving helicopter transportation. In Table 1 there are the 14 entries from the whole set of 200 insurance claims in 2019. The data received were rounded to whole values.

€ 3,067	€ 3,102	€ 3,441	€ 3,448	€ 3,026	€ 3,346	€ 3,133
€ 3,328	€ 3,434	€ 4,325	€ 3,973	€ 4,048	€ 3,291	€ 3,628

Table 1. Sample entries from the set of 200 insurance claims in 2019.

When looking at the whole set of data we can observe that none of the claims are smaller than €3,000. This comes from the fact that the minimum cost of transportation with a helicopter is €3,000. After formatting the data, they were imported to programming software RStudio [10], where all the analyses were done. First, we can observe the histogram of claim amounts, shown in Figure 1, to get a better overview of the received data.

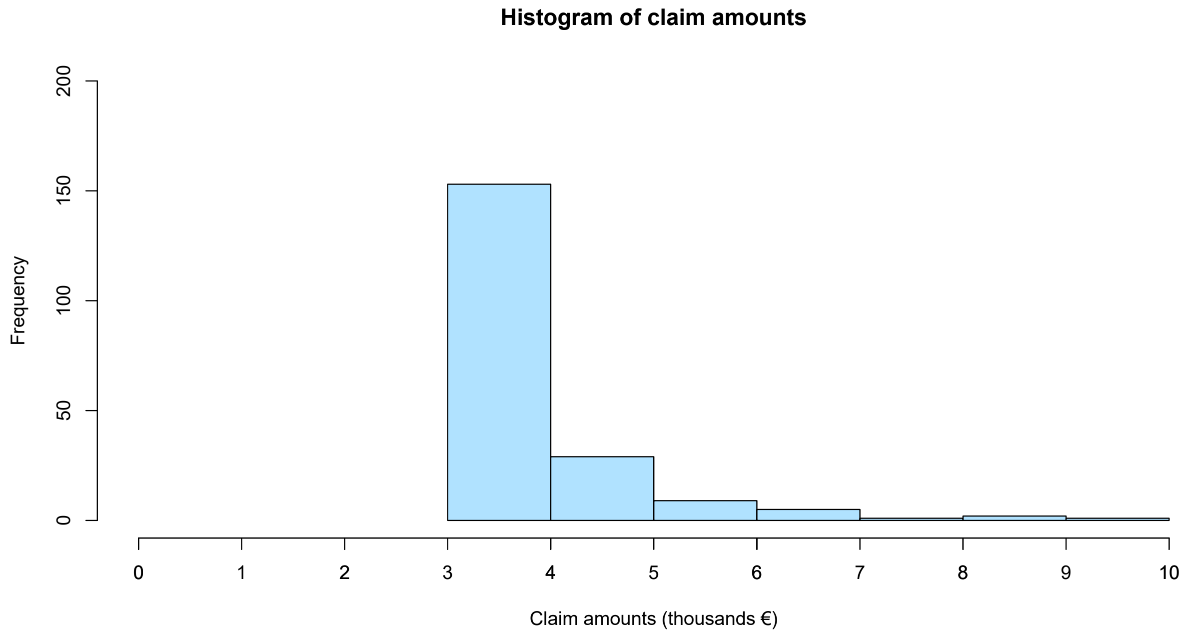

Figure 1. Histogram of claim amounts.

We can observe that most claims lie in the interval between €3,000 and €4,000, while only a few are greater than €7,000. We get an even better picture from looking at the histogram showing us relative frequency (see Figure 2) . Relating it to our case, relative frequency in statistics presents the probability that a claim amount is in a specific interval. It is calculated by dividing the number of claims that lie in the chosen interval by the number of all claims. [11]

For modelling the distribution of a random variable X_i we choose continuous distributions, such as Weibull, Pareto, gamma, exponential, normal, or log-normal distribution [12]. If we study the properties of Pareto distribution, we find that it differs substantially from the other five – one of its parameters represents the minimum value that Pareto distributed random variable can take. When all the data are larger than a specific value x_m the Pareto distribution will be the best approximation for the data. From both histograms we can observe that none of the claim amounts is smaller than €3,000, therefore we will choose the Pareto distribution. Let us look at the formal definition of this distribution.

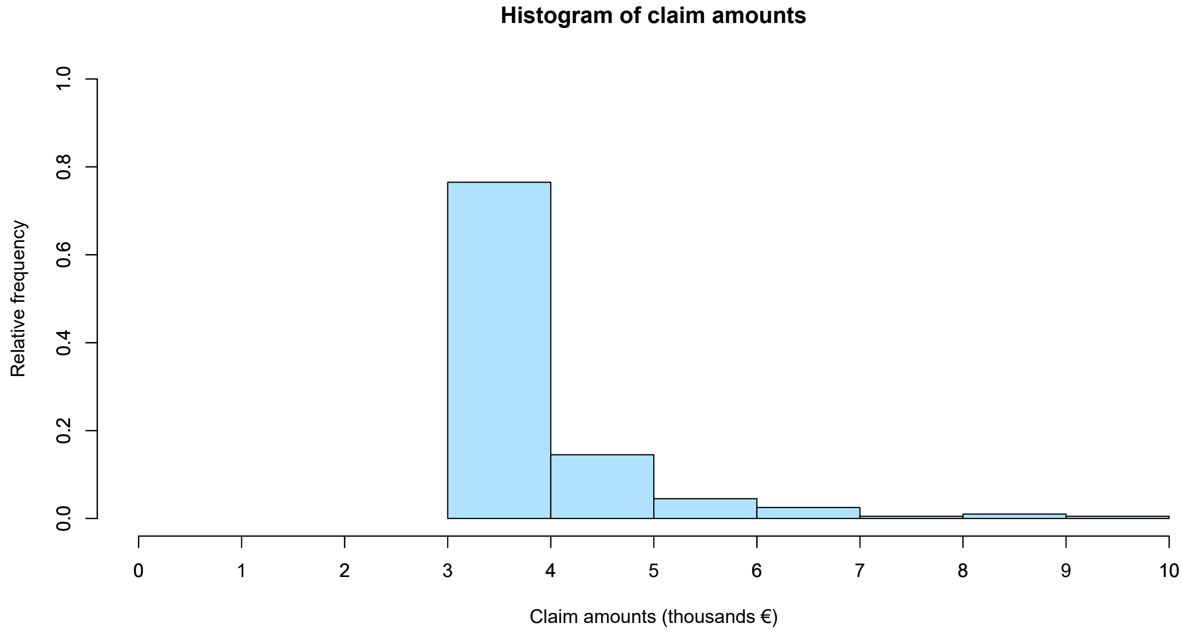

Figure 2. Relative frequency.

4.1 Pareto distribution

Definition 5 ([13, 14]). We say that the random variable X has Pareto distribution (Type 1) when its cumulative distribution function is given by

$$F_X(t) = \begin{cases} 1 - \left(\frac{x_m}{t}\right)^\alpha & ; t \geq x_m \\ 0 & ; t < x_m \end{cases}.$$

Here x_m is the minimum positive value of X and α is a positive parameter. We call x_m the scale parameter and α the shape parameter.

By differentiation of the cumulative distribution function, we get the probability density function in the following process [6].

$$\frac{\mathrm{d}F_X(t)}{\mathrm{d}t} = \begin{cases} \frac{\mathrm{d}}{\mathrm{d}t}\left[1 - \left(\frac{x_m}{t}\right)^\alpha\right] & ; t \geq x_m \\ \frac{\mathrm{d}}{\mathrm{d}t}\, 0 & ; t < x_m \end{cases}$$

Now, we will focus only on the derivative when $t \geq x_m$, as differentiation when $t \leq x_m$ is trivial.

$$\frac{\mathrm{d}}{\mathrm{d}t}\left[1 - \left(\frac{x_m}{t}\right)^\alpha\right] = \frac{\mathrm{d}}{\mathrm{d}t}1 + \frac{\mathrm{d}}{\mathrm{d}t}\left[-\left(\frac{x_m}{t}\right)^\alpha\right] = 0 + \frac{\mathrm{d}}{\mathrm{d}t}\left[-\left(x_m\right)^\alpha \cdot t^{-\alpha}\right] =$$
$$= -\left(x_m\right)^\alpha \cdot (-\alpha) \cdot t^{-\alpha-1} = \alpha \cdot \left(x_m\right)^\alpha \cdot t^{-(\alpha+1)} = \frac{\alpha \cdot \left(x_m\right)^\alpha}{t^{\alpha+1}}$$

Hence, the probability density function of the Pareto distribution is given by

$$f_X(t) = \begin{cases} \frac{\alpha \cdot (x_m)^\alpha}{t^{\alpha+1}} & ; t \geq x_m \\ 0 & ; t < x_m \end{cases}.$$

4.2 Approximating the parameters of Pareto distribution

We now have to approximate the two parameters, x_m and α of Pareto distribution. Using the Minimum Distance Estimation (Cramer-von Mises method) in programming language R we get a cumulative distribution function, which has the minimum squared difference from the data points of our empirical claim amounts [15]. This gives us the following parameter estimates for the best fit. The parameter estimates were calculated by the computer and taken from the RStudio.

$$\alpha = 4.9712 \qquad x_m = 2.9831$$

If we reflect back to the aim, we would like to get a result rounded to 2 decimal places, since we are operating with money. Therefore, the value approximations to 4 decimal places will give an accurate representation and are enough for the scope of this article. Any further results in this article will be rounded to 4 decimal places as well, except for the final value. This way we avoid making any error due to the rounding of intermediate results in the process of obtaining the final result. Above the value of x_m should be 3.00, since all the data are higher than €3,000. However, due to the numerical error, the approximation given by the computer is 0.56% smaller, which is quite accurate and will have a negligible impact on our final result. Drawing the probability density function of the Pareto distributed variable, we can observe a close match with the empirical distribution (see Figure 3).

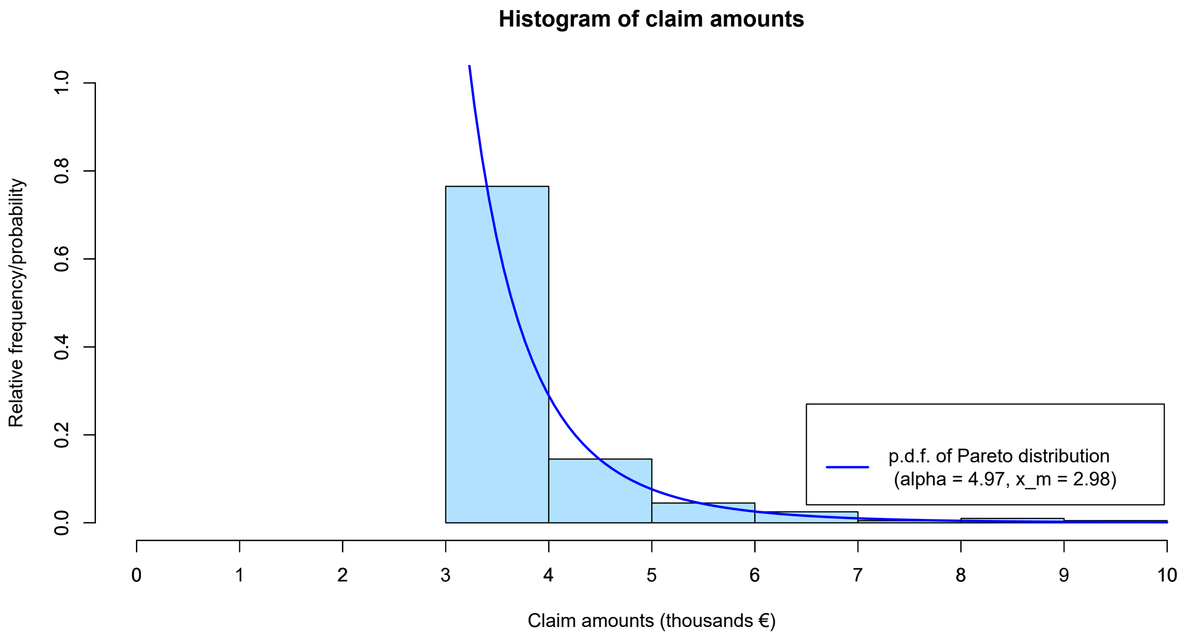

Figure 3. Relative density curve.

An even better match is observable in Figure 4, showing us the cumulative distribution function of empirical data and of the Pareto distribution with the above-mentioned parameters.

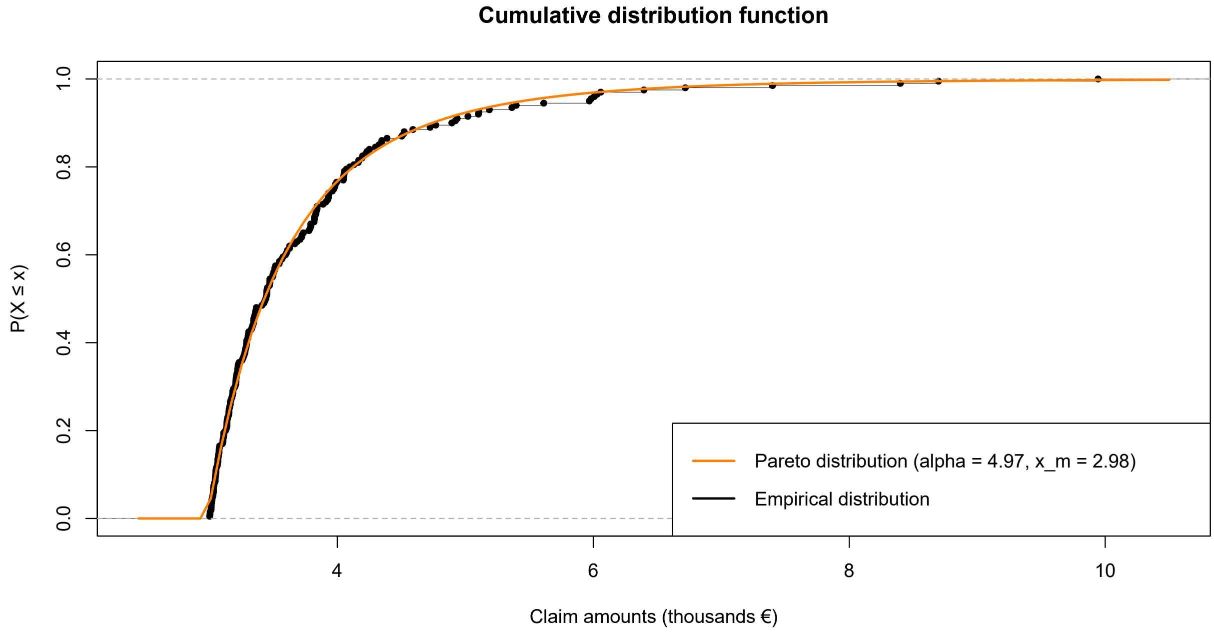

Figure 4. Cumulative distribution function.

In order to use Wald's Identity, we will have to compute the expected value of the Pareto distribution. First, we will derive the general formula and then we will use it to get the value of $E[X_i]$.

$$
\begin{aligned}
E\left[X_i\right] &= \int_{-\infty}^{\infty} t \cdot f_X(t) \cdot \mathrm{d}t = \int_{-\infty}^{x_m} t \cdot f_X(t) \cdot \mathrm{d}t + \int_{x_m}^{\infty} t \cdot f_X(t) \cdot \mathrm{d}t = \\
&= 0 + \int_{x_m}^{\infty} t \cdot f_X(t) \cdot \mathrm{d}t = \int_{x_m}^{\infty} t \cdot \frac{\alpha \cdot (x_m)^{\alpha}}{t^{\alpha+1}} \cdot \mathrm{d}t = \\
&= \alpha \cdot (x_m)^{\alpha} \int_{x_m}^{\infty} \frac{1}{t^{\alpha}} \cdot \mathrm{d}t = \alpha \cdot (x_m)^{\alpha} \cdot \left[\frac{t^{-\alpha+1}}{-\alpha+1}\right]\Bigg|_{x_m}^{\infty} = \\
&= \frac{\alpha \cdot (x_m)^{\alpha}}{-\alpha+1} \cdot \left[\lim_{c \to \infty}\left(\frac{1}{c^{\alpha-1}}\right) - \frac{1}{(x_m)^{\alpha-1}}\right] = \begin{cases} \frac{\alpha \cdot x_m}{\alpha-1} & ; \quad \alpha > 1 \\ \infty & ; 0 < \alpha \leq 1 \end{cases}
\end{aligned}
$$

Above, we begin with the equation for the expected value of a continuous random variable [16]. Then we convert the definite integral into two definite integrals, which differ in the interval of integration. The first integral from $-\infty$ to x_m is equal to zero because Pareto distributed random variable takes no values smaller than x_m meaning that probability density function on the interval $(-\infty, x_m]$ is equal to 0. Then we take constant factors out of the integral and integrate with respect to t. We can observe that the integral is an improper integral, which has to be solved using limits [17]. After resolving the limits and simplifying the results, we derive the formula for the expected value of a random variable X, which is dependent only on the scale and shape parameters of the Pareto distribution.

Using the estimations for these two parameters, we get the expected value of X_i.

$$
E\left[X_i\right] = \frac{4.9712 \cdot 2.9831}{4.9712 - 1} \cong 3.7343
$$

If we convert the upper value to thousands form, we get the expected amount per claim of €3,734.3.

It is common practice that insurance companies perform backchecking and compare estimated average claims to the actual average claims. From the received data we can calculate that the average claim amount in the year 2019 was €3,768.6, which is approximately 0.92% greater that the expacted value of Pareto distributed variable.

5. Distribution of N

Now, that we know the expected value of X, we need to determine it for N as well. N is a discrete random variable that measures the number of insurance company contract holders, who will receive restitution in the following year. In Table 2 we can observe the number of contract holders and the number of claims in years from 1995 to 2019.

YEAR	NUMBER OF CONTRACT HOLDERS	NUMBER OF CLAIMS	PERCENTAGE
1995	885	123	13.9%
1996	893	129	14.4%
1997	896	112	12.5%
1998	903	140	15.5%
1999	927	141	15.2%
2000	934	136	14.6%
2001	944	135	14.3%
2002	972	130	13.4%
2003	977	133	13.6%
2004	999	142	14.2%
2005	1017	134	13.2%
2006	1036	150	14.5%
2007	1005	165	16.4%
2008	942	158	16.8%
2009	935	127	13.6%
2010	964	146	15.1%
2011	986	160	16.2%
2012	1012	143	14.1%
2013	1043	155	14.9%
2014	1087	165	15.2%
2015	1128	185	16.4%
2016	1164	174	14.9%
2017	1221	211	17.3%
2018	1266	198	15.6%
2019	1274	205	16.1%

Table 2. Number of contract holders and number of claims in years from 1995 to 2019.

From the above data, we can see that there is a significant drop in the number of contracts in the year 2008. If we critically evaluate and connect this drop with the global situation in that year, we can account for this observation with the global financial market crisis. It is reasonable to conclude that the crisis also affected insurance companies since some people were not able to afford insurance policies.

For modelling the number of claims N, we will have to choose a discrete distribution. Normally we choose binomial, negative binomial, or Poison distribution [12]. When deciding for the most appropriate one, we have to be aware of certain characteristics of our problem, as well as the conditions that need to be met for a certain distribution. We have only two possible outcomes of an insurance contract, i.e., two complementary events – an insurance claim is made or not, and the i-th claim is made with the same probability p for all $i \in \mathbb{N}$ (because X_i are independent between each other). We know that one of those two events will certainly occur, therefore their probabilities add up to 1. For clarity, we say that $1 - p = q$, which will denote the probability that an individual contract holder does not make a claim. The binomial distribution is a discrete probability distribution which is used for modelling repeated independent binary events, that have a known probability for each of the two outcomes. Specifically, it is used when we want to evaluate the probability that a single event occurs a certain number of times in Bernoulli trials [18]. We can observe, that this type of distribution exactly matches our specific case; therefore, it is reasonable to use the binomial distribution for calculating the expected value of a discrete random variable N.

Definition 6 ([18]). We say that random variable N has a binomial distribution with parameters $n \in \mathbb{N}$ and $p \in [0, 1]$ when the probability of the event where exactly k insurance claims will be made, is

$$\mathrm{P}(N = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

The total number of contract holders is represented by n, the probability of a claim event by p and $1 - p$ is the probability that there will be no claim from the contact holder.

First, we will predict the expected number of insurance contract holders in the following year using linear regression, which we need to calculate the expected value of N. In Figure 5 we can see a drastic fall in 2008. This anomaly refers back to our observations from the data table and is even better seen in this graph.

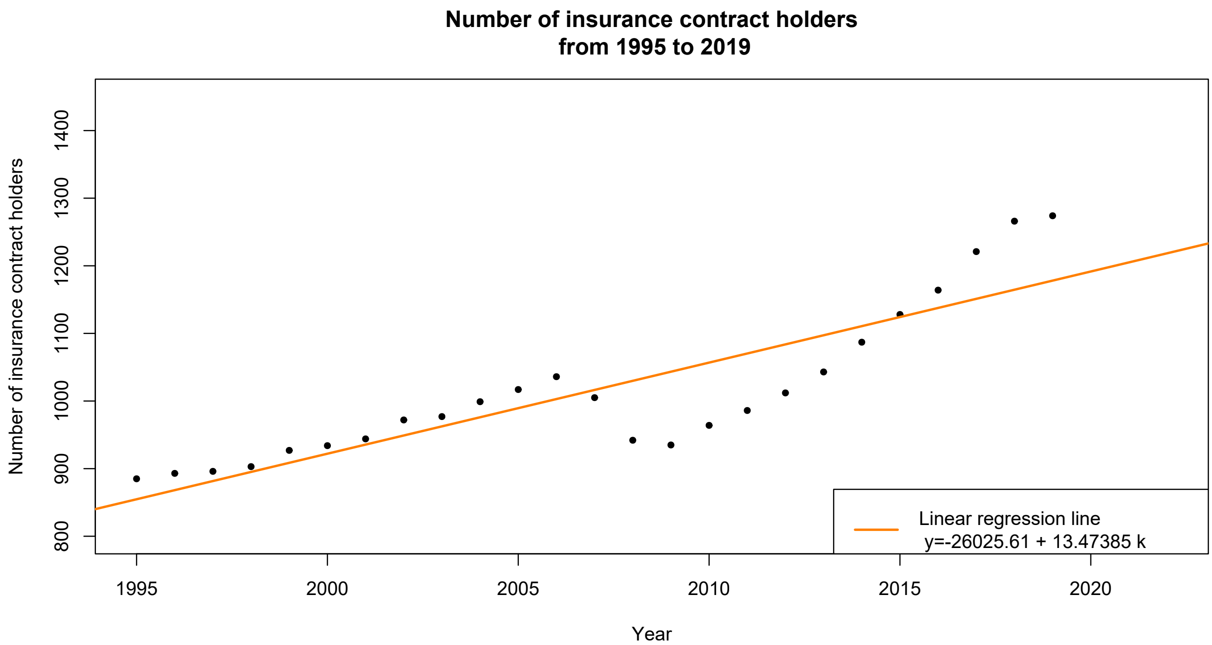

Figure 5. Number of insurance contract holders from 1995 to 2019 and a linear regression line approximating the data from Table 2.

Using the linear regression line estimation for 2020, we get that the predicted number of insurance contract holders is equal to

$$-26025.6100 + 13.4739 \cdot 2020 = 1191.6680.$$

Since the number of insurance contract holders has to be a natural number, we will use 1192 in future calculations.

Using the linear regression, we can extrapolate the number of insurance contact holders for any year in the future. The trendline above shows that a downtrend is predicted for the future. If we apply some reasoning and critical evaluation of this data, we can see that it corresponds with predictions of an impending crisis.

Regarding the fact that neither the number of claims nor the percentage of claims show any trend, we will evaluate the probability that the i-th contract holder has an accident and makes a claim request as

$$\frac{\text{sum of all claims from 1995 to 2019}}{\text{sum of all contracts from 1995 to 2019}} = \frac{3,797}{25,410} \cong 0.1494.$$

The fact that the dataset is extensive so that our sample is large enough, makes our estimation very accurate.

As stated, to get the number of claims in the future, we will use the binomial distribution, which tells us the number of outcomes of a certain event, based on the fact that all contract holders are equally likely to have an accident and make a claim request. From above we can observe that this probability is approximately 14.94% and the total number of contract holders is expected to be 1192.

Now, we have to evaluate the expected value of the binomial distribution.

$$\begin{aligned} E[N] &= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k} = \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} = \\ &= \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k q^{n-k} = np \sum_{k=1}^{n} \binom{n-1}{k-1} p^{k-1} q^{(n-1)-(k-1)} = \\ &= np \sum_{j=0}^{m} \binom{m}{j} p^j q^{m-j} = \\ &= np \end{aligned}$$

In the upper process, we begin with inserting the probability mass function of the binomial distribution in the formula for the expected value of a discrete random variable. Since $k\binom{n}{k}p^k q^{n-k} = 0$ for $k = 0$, we can change the lower boundary of the sum. In the next step, we include the factors of the binomial coefficient and use the equality $k\binom{n}{k} = n\binom{n-1}{k-1}$. Then we factor out the common factor np out of the sum in the second line. In the last step, we substitute $m = n-1$ and $j = k-1$. Simplifying by using $(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k}$ and $p + q = 1$, we get to the result that $E[N] = np$.

From Figure 6, we can infer that the most probable possible number of claims will lie between $k = 140$ and $k = 220$. Substituting p and n in the $E[N] = np$, we get the expected value for N.

$$E[N] = 178.0848$$

Now that we have the estimated values for both X and N, we can finally get the estimated value for S. Using Wald's identity, we get

$$E[S] = E[X_i] \cdot E[N] \cong 178.0848 \cdot 3.7343 \cong 665.0221. \tag{1}$$

After converting the result from the thousands number form, we get that our insurance company can expect to have the total amount of claims of €665,022 in the following year.

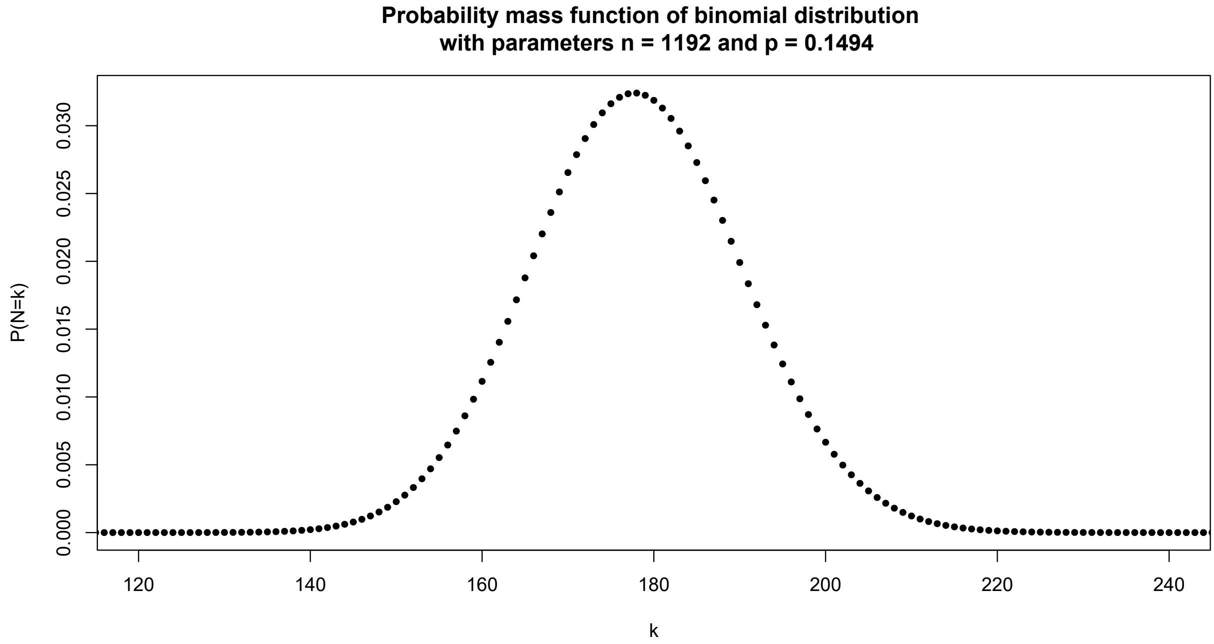

Figure 6. Probability mass function of the binomial distribution.

To critically evaluate our research, we can conclude that it is best for this result to be rounded to the whole value because it would be inaccurate to specify the results to one decimal place because of the accuracy of some predictions used in our model and inclusion of the error in our result. Furthermore, in the context of this article, it is unnecessary for an insurance company to predict the total amount so accurately, due to the order of magnitude of the money trades they are dealing with. This result differs from my assumptions that were around half a million euros for approximately 33%, meaning that it is a higher value than initially expected. To comment on this result the actual value for the year 2020 might be much lower because of the pandemic. Since some of the ski resorts were closed, there were no injuries, resulting in fewer helicopter transportations what means that an insurance company will have fewer claim demands and consequently need less money for the year 2020.

6. Modelling the total loss

To generalise the result we got using the Wald's Identity, we generate the values of S. First, we generate 100,000 random values of $Bin(1192, 0.1494)$ distributed variables. Each of these values is then used to generate its specific number of $Pareto(4.9712, 2.9831)$ distributed variables, which are in each specific instance summed, to get the value of S. Hence, we get 100,000 values of S, which can be further used to generate the empirical cumulative distrubution function of the random variable S. The $F_S(s)$ can be observed in Figure 7. If we calculate the average value of S, we get 664.9587, what is only 0.01% less than the expected value $E[S]$ which we calculated using the Wald's Identity in the previous chapter. It is important for insurance companies to estimate the probabilities that the total claims will exceed a given level. This can be read from the function $F_S(s)$. The 90^{th}, 95^{th} and the 99.5^{th} percentiles of $F_S(s)$ can be observed in Table 3.

90%	95%	99.5%
726.6958	745.1000	790.7612

Table 3. Percentiles of distribution of S.

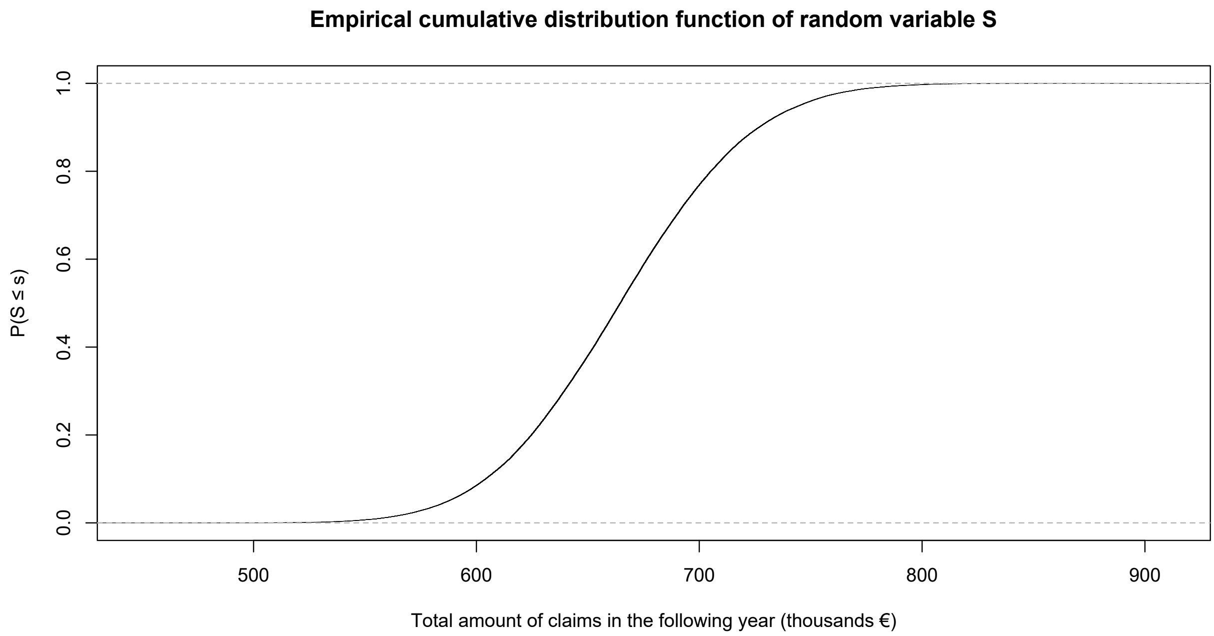

Figure 7. Empirical cumulative distribution function of random variable S (100,000 values).

7. Conclusion

Our initial aim was to calculate the expected amount that the insurance company will need to maintain solvent. As we can see, we approached the problem by collecting the data from the insurance company, which we further used to predict the amount for the year 2020. Here we must also be aware that these predictions were based on the current trends of insurance claims and neglected any fluctuations in the cost of helicopter transportation. However, these factors are very insignificant as they could affect the final result by a maximum of 2%, which is slightly above the inflation rates for Slovenia, predicted by the European Commission for the year 2021 [19]. Insurance companies could therefore use the predicted inflation rates and make and even more accurate predictions.

The method that we used for reaching our aim, was based on the collective risk model. In order to calculate the expected value of the total amount of insurance claims, we used Wald's identity, for which we defined two random variables, one discrete and one continuous. We needed to evaluate their expected values using parameters of two different types of distribution, which were carefully but correctly chosen by examination of how their characteristics suit my data and real-life context. By knowing the expected values of the two random variables, we were finally able to come to the expected value of the total amount of insurance claims. So, the main goal of this research has been reached.

We must be aware, however, of some assumptions that were made in this process for the purpose of the theoretical model that we used. Firstly, as stated above, we assumed that the price of helicopter transportation does not change. Should we have the data of those prices for the same time period as we had for the number of insurance claims, we could also estimate the trend value for the year 2020. This could be a possible improvement to this article since it would give an even more representative value.

Furthermore, our definition of the collective risk model is independent of time. As the value of money changes through time (inflation), such models could be upgraded to incorporate the time component. To do so, we would have to use stochastic processes. Another possible further extension of my model would be to investigate how insurance companies calculate the amount of insurance premium, which will allow them to make a profit as well as cover the expected amount of claims.

In the process, the biggest challenge I faced was to choose the two distributions that give the best approximation. I overcame this by examining the characteristics of each and compare those to my specific situation. Besides, it was intriguing, yet challenging to acquire new programming skills. Overall, I believe this research was an excellent opportunity for me to explore my personal interest. Although it required a lot of additional exploration greatly outside the mathematical curriculum, I was able to achieve the aim I set at the beginning.

REFERENCES

[1] A. Hayes and J. Young, *Investopedia: Solvency*, (2020), URL: https://www.investopedia.com/terms/s/solvency.asp (uporabljeno 30. 1. 2021).

[2] European commission, *Solvency II Overview – Frequently asked questions*, (2015), URL: https://ec.europa.eu/commission/presscorner/detail/en/MEMO_15_3120 (uporabljeno 16. 1. 2021).

[3] Kognity AB, *Probability and statistics: Discrete random variables*, URL: https://app.kognity.com/study/app/ibdp-mathematics-analysis-and-approaches-hl/probability-and-statistics/discrete-random-variables/the-big-picture/ (uporabljeno 12. 1. 2021).

[4] *Random variables*, New Haven: Yale University, Department of Statistics and Data Science, URL: http://www.stat.yale.edu/Courses/1997-98/101/ranvar.htm (uporabljeno 8. 1. 2021).

[5] Fakulteta za gradbeništvo in geodezijo, *Slučajne spremenljivke*, Katedra za mehaniko, (2010) URL: http://km.fgg.uni-lj.si/PREDMETI/sei/Slucajne%20spremenljivke.PDF (uporabljeno 16. 1. 2021).

[6] I. Gordon, *Continuous probability distributions – A guide for teachers (Years 11–12)*, (2013), URL: https://amsi.org.au/ESA_Senior_Years/PDF/ContProbDist4e.pdf (uporabljeno 14. 1. 2021).

[7] L. Prislan, *Kolektivni modeli tveganja v zavarovalništvu*, (2010), URL: https://repozitorij.uni-lj.si/IzpisGradiva.php?lang=slv&id=96760 (uporabljeno 15. 1. 2021).

[8] C. McKay, *Probability and Statistics*, (2019), URL: https://books.google.si/books/about/Probability_and_Statistics.html?id=w-nEDwAAQBAJ&redir_esc=y (uporabljeno 13. 1. 2021).

[9] S. Chewi, *Wald's Identity*, (2017), URL: https://inst.eecs.berkeley.edu/ ee126/fa17/wald.pdf (uporabljeno 17. 1. 2021).

[10] R Core Team, *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing, (2020), URL: https://www.R-project.org/ (uporabljeno: 8. 1. 2021).

[11] *Histograms*, (2012), URL: https://www.radford.edu/ scorwin/courses/200/book/50Histograms.html (uporabljeno: 14. 1. 2021).

[12] T. Peternelj, *Analiza in modeliranje rasti premoženjske zavarovalnice*, (2016), URL: http://www.cek.ef.uni-lj.si/magister/peternelj5274.pdf (uporabljeno 17. 1. 2021).

[13] C. Kleiber and S. Kotz, *Statistical Size Distributions in Economics and Actuarial Sciences*, (2003), New Jersey: Wiley.

[14] *Pareto distribution*, Wikipedia: The Free Encyclopedia, URL: https://en.wikipedia.org/wiki/Pareto_distribution (uporabljeno 27. 1. 2021).

[15] C. Dutang, V. Goulet and M. Pigeopn, *Actuar: An R Package for Actuarial Science*, Journal of Statistical Software, **15(4)** (2008), 1–37.

[16] Kognity AB, *Probability and statistics: Mean and variance of a continuous random variable*, URL: https://app.kognity.com/study/app/ibdp-mathematics-analysis-and-approaches-hl/probability-and-statistics/continuous-random-variables/mean-and-variance-of-a-continuous-random-variable/ (uporabljeno 30. 1. 2021).

[17] P. Menz and N. Mulberry, *Calculus Early Transcendentals: Integral & Multi-Variable Calculus for Social Sciences: Improper Integrals*, (2020), URL: https://www.sfu.ca/math-coursenotes/Math%20158%20Course%20Notes/sec_ImproperIntegrals.html (uporabljeno 14. 1. 2021).

[18] Kognity AB, *Probability and statistics: The binomial distribution*, URL: Available on: https://app.kognity.com/study/app/ibdp-mathematics-analysis-and-approaches-hl/probability-and-statistics/binomial-distribution/the-binomial-distribution/ (uporabljeno 15. 1. 2021).

[19] European commission, *Economic forecast for Slovenia*, (2020), URL: https://ec.europa.eu/info/business-economy-euro/economic-performance-and-forecasts/economic-performance-country/slovenia/economic-forecast-slovenia_en (uporabljeno 2. 2. 2021).