Simulation in Finance and Insurance

Case Study

UNIL | Université de Lausanne
HEC Lausanne

Angelopoulou Antonia
Jaillot Pierre
Hassan Mohamed

Spring 2023

Contents

1. Executive Summary

The insurance company under review exhibits concerning patterns in its operations, particularly related to its loss ratio, premium pricing, and consistency in premium charges. The company's loss ratio, which measures the proportion of claim payouts to premiums earned, exceeds 100%. This indicates that the company's claims expenses exceed the premiums collected, resulting in unsustainable financial performance.

Furthermore, the insurance company charges remarkably low premiums relative to the level of risk it assumes. This pricing strategy raises concerns about the adequacy of reserves and the company's ability to cover future claims. Insufficient premiums can lead to financial instability and potential difficulties in meeting policyholders' claims obligations.

Another area of concern is the lack of consistency in premium charges. Inconsistencies in premium pricing can create confusion among policyholders and undermine the fairness and transparency of the insurance products. This inconsistent approach to premium setting may indicate inadequate risk assessment or a lack of robust underwriting practices.

Addressing these issues is critical for the insurance company's long-term sustainability and profitability. The company should undertake a comprehensive evaluation of its pricing strategy, ensuring that premiums are appropriately aligned with the risks associated with the policies offered. Additionally, the company should prioritize risk management practices, including thorough underwriting processes and robust claims management, to effectively mitigate losses and improve the overall loss ratio.

By implementing necessary corrective measures, such as adjusting premium rates to reflect the true risk exposure and ensuring consistency in premium charges, the insurance company can enhance its financial stability, improve its loss ratio, and ultimately provide better protection for its policyholders. This will position the company for sustainable growth, profitability, and enhanced customer trust and satisfaction in the highly competitive insurance market.

2. Preliminary Work

The first phase of proper data analysis involves preprocessing and cleaning the data: this could be paraphrased as addressing missing values, handling outliers, resolving inconsistencies, standardizing formats, and removing irrelevant or duplicate data entries. Essentially, cleaning data reduces potential bias and ensures that analysis is based on representative data.

2.1 Data Preparation and Data Cleaning

The portfolio data is provided in CSV file format. Initially, all the data was read as a character type. Therefore, our first step was to carefully convert the character data type into the appropriate data types. To achieve this, we utilized Regular Expression patterns, commonly known as Regex to compare the data entries with the initial state of the data and ensure accurate data type conversions.

Following the data type conversions, we conducted sanity checks to identify erroneous data entries and outliers. During the sanity checks, it was discovered that there were two data points with negative premium amounts. The data points had the IDs 100 and 900 respectively. The data contained 11 numerical columns and 4 categorical ones.

The data cleaning criteria used for this process are briefly summarised in the following tables, respectively for the numerical and categorical variables.

Variable	Name	Variable Type	Accepted Range
Claim Frequency	CLM_FREQ	Discrete	$\mathbb{N}$
Claim Amounts	CLM_AMT_i for $i \in [\![1, 7]\!]$	Continuous	$(0, \infty)$
Age	AGE	Discrete	$[\![\text{Legal Driving Age},\ \text{Maximum Attainable Age}]\!]$
Premium	PREMIUM	Continuous	$(0, \infty)$

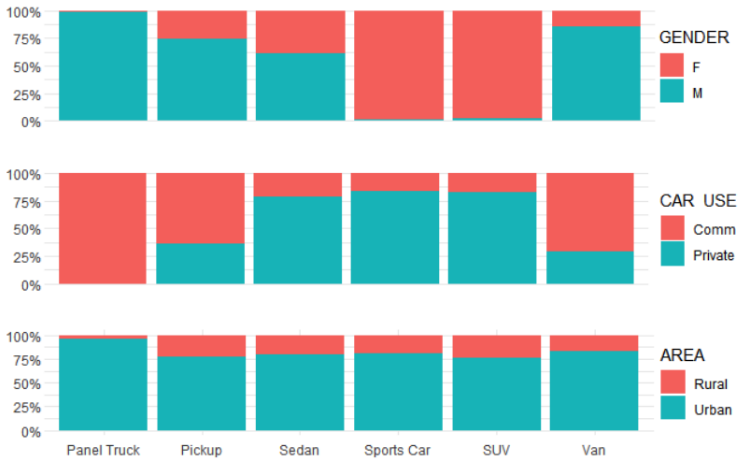
Variable	Name	Accepted Values
Car Use	CAR_USE	Commercial, Private
Car Type	CAR_TYPE	Panel Truck, Pickup, Sedan, Sports Car, SUV, Van
Gender	GENDER	F, M
Area	AREA	Rural, Urban

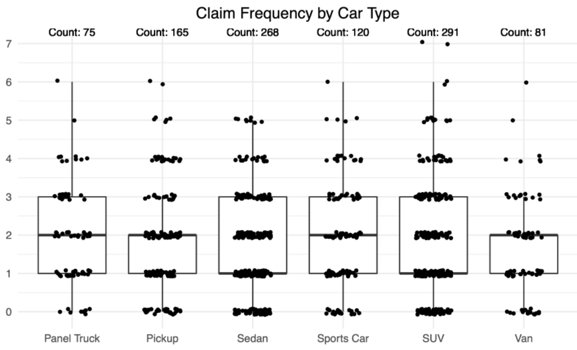
Finally, we performed various data transformations to analyze different aspects of the data. This included excluding null values from the claim amount columns and pivoting the claim amount columns to analyze claim amounts in relation to other variables, among other techniques.

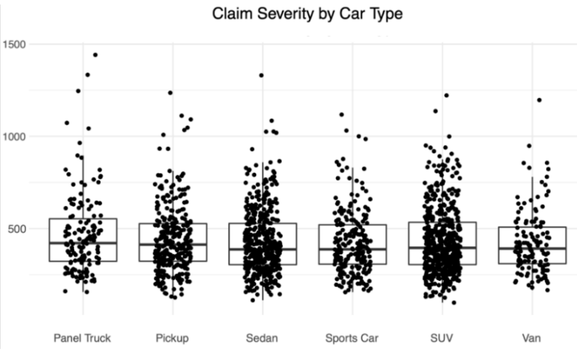
2.2 Data Exploration

The portfolio comprises 1000 policyholders, including 56.4% females with a median age of 44 and 45 for males. The majority, representing 80.4% percentage of the policyholders, reside in urban areas and possess private vehicles 63.8%. Below we present summary statistics of our data:

	Claim Frequency	Claim Amounts							Age	Premium
	CLM_FREQ	CLM_AMT_i							AGE	PREMIUM
		$i = 1$	2	3	4	5	6	7		
Min.	0	125.0	125.0	126.0	126.0	126.0	252.0	615.0	19	500.0
Q1	1	304.0	302.5	293.0	280.8	293.0	344.0	631.8	39	631.8
Median	2	391.0	391.5	403.0	406.5	412.0	367.0	648.5	44	758.0
Mean	1.763	427.2	427.5	438.1	440.0	437.9	438.9	648.5	44.56	752.4
Q3	3	517.2	524.2	552.2	577.0	606.8	615.0	665.2	51	875.0
Max.	7	1442.0	1442.0	1442.0	1331.0	818.0	682.0	682.0	80	1000.0







2.2.1 Claim Frequency Analysis

Here we analyse the frequency of claims per car type while considering other data variables. We are interested in finding what causes a car to engage in a claim or possibly multiple. Analysis of the portfolio using the above plots show that panel trucks, pickup followed by sports car tend to have a relatively higher median frequency of claims. This does not include SUVs which contribute twice to 7 claims per policy. Although the average number of claims per SUV is 1.76 the median is much lower with a value of 1 due to the large number of non-claim occurrences.

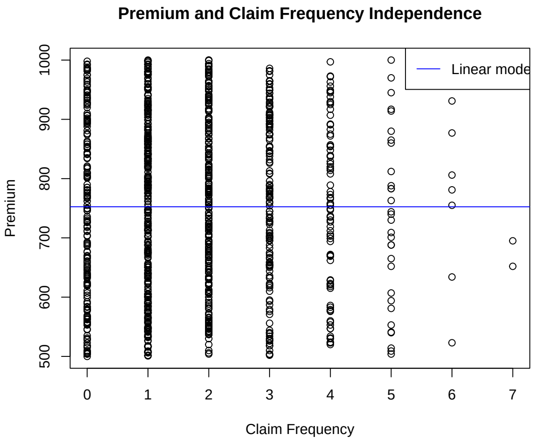
Panel trucks and pickups are predominantly commercial and driven by males, however, vans are also commercial and driven predominantly by males but witness fewer claims frequencies. A Possible explanation could be the type of job associated with the car or possibly its dimensions leading to higher claim frequencies. Sports cars appear to be driven almost exclusively by females 99.16%. A reasonable explanation for the relatively high claim frequency is that sports cars participate in racing events where the likelihood of an accident is relatively high.

2.2.2 Claim Severity Analysis

The portfolio shows that across all six car types, the severity of claims appears to be uniformly distributed around the value of 434. Further analysis concluded no patterns with regard to claim severity in relation to other variables present in the data set.

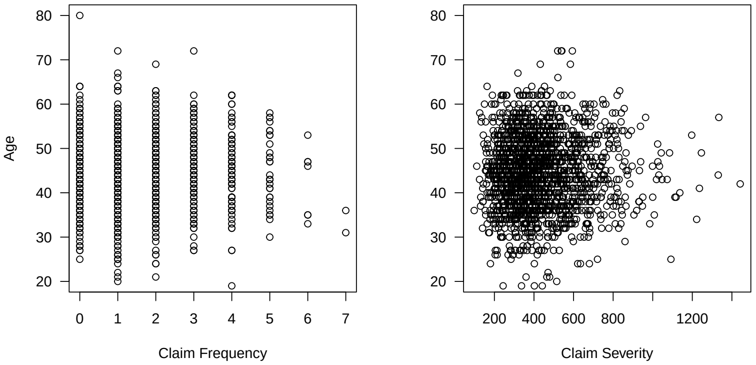
2.2.3 Premium Analysis

Premium is uniformly distributed with a minimum value of 500 and a maximum value of 1000, this leads to an average premium of 752. It also seems to be independent of the number of claim occurrences.



The premium amount also appears to show no patterns across car types, areas, gender, and car use. This means that when a policyholder joins they are charged a randomly assigned premium that follows a uniform distribution with the aforementioned parameters. This leads to some inconsistencies, for instance, the oldest female policyholders are 72 years old and drive sports cars and there was one policyholder with 3 claims paying 667 whereas another one with only one claim paying 868. Another remarkable finding is that the oldest policyholder is a male of age 80 with no prior claim experience paying 802.

2.2.4 Age Analysis



The ages in our data set vary between 19 and 80 years of age. There is no particular pattern for low frequency and severity claims, however, it appears to be that most high frequency and severity claims are attributed to policyholders of age between 35 and 57. This means that not only that they engage in more accidents, but also in more dangerous ones.

3. Model Selection and Validation

The selection of appropriate and representative models is crucial for insurance companies and a key factor in determining their product development policy. With respect to this, we examined several possible models for both the frequency and severity of claims and ran several tests to decide which would better interpret our data set.

In the following sections, we will survey the selection process of respectively the claim frequency and the claim severity models. Indeed, based on our data exploration, it appears to be reasonable to adopt a collective risk model. Such a model assumes independence among claim amounts and also between the number of claims and claim amounts. Spearman's correlation coefficient between claim frequency and claim amounts is of 0.0139, indicating a negligible correlation and validating the suitability of this approach.

3.1 Claim Frequency Modelling

3.1.1 Candidate Models

In order to select of model for the claim frequency, we take into consideration the following criterion. The claim frequency of a given policy can take any value that is a non-negative integer. Thus, the models considered have to accommodate a discrete random variable that has no known upper bound. Indeed, although the claim frequencies observed in the data set attain a maximum value of 7, there is no theoretical argument for the existence of such a bound for the number of claims filed by a policyholder. Consequently, although the binomial distribution has the advantage of simplicity and is sometimes used to model the frequency of events, we have excluded it from our analysis on the grounds of its admission of an upper bound. Using this criterion, we pre-select the following common distributions as candidates for our model.

Poisson

Our first candidate for the claim frequency model is the Poisson distribution. This distribution is commonly used for such models in insurance, and admits a single real positive rate parameter λ. In order to fit such a distribution to our data set, we use the method of moments to approximate the parameter λ with the estimator $\hat{\lambda}$ using the sample mean, as follows:

$$X \sim \text{ Pois}(\lambda) \quad \Rightarrow \quad \mathbb{E}(X) = \lambda \qquad\qquad \text{thus: } \hat{\lambda} = \overline{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

where $\{x_i\}_{i=1}^{n}$ denotes the claim frequencies observed in the data set. In this case, the estimator $\hat{\lambda}$ takes the value 1.763.

Negative Binomial

Our second candidate for the claim frequency model is the Negative Binomial distribution. It is a more popular alternative to the Poisson distribution for the modelling of claim frequency. Formally, the negative binomial distribution models the number of failures in a sequence of independent and identically distributed Bernoulli trials before a specified number of successes occurs. Accordingly, this distribution admits two parameters, a positive number of successes denoted by r and a fixed probability of success for the Bernoulli trials denoted by p. As for the Poisson distribution, we proceed by respectively approximating the aforementioned parameters r and p by the estimators $\hat{r}$ and $\hat{p}$. Again, using the method of moments, we can observe that:

$$X \sim \text{NB}(r,p) \quad \Rightarrow \quad \mathbb{E}(X) = \bar{x}_n = \frac{(1-p)\cdot r}{p} \quad \text{and} \quad \text{Var}(X) = s_n^2 = \frac{(1-p)\cdot r}{p^2}$$

allowing us to express the parameters and estimators as:

$$r = \frac{\mathbb{E}(X)^2}{\text{Var}(X) - \mathbb{E}(X)} \quad \text{and} \quad p = \frac{\mathbb{E}(X)}{\text{Var}(X)} \quad \Rightarrow \quad \hat{r} = \frac{\bar{x}_n^2}{s_n^2 - \bar{x}_n} \quad \text{and} \quad \hat{p} = \frac{\bar{x}_n}{s_n^2}$$

where s_n^2 denotes the sample variance given by $\frac{1}{n-1}\Sigma_{i=1}^{n}(x_i - \overline{x}_n)^2$. Here, these take the values:

$$\hat{r} = 20.48274 \quad \text{and} \quad \hat{p} = 0.9207489.$$

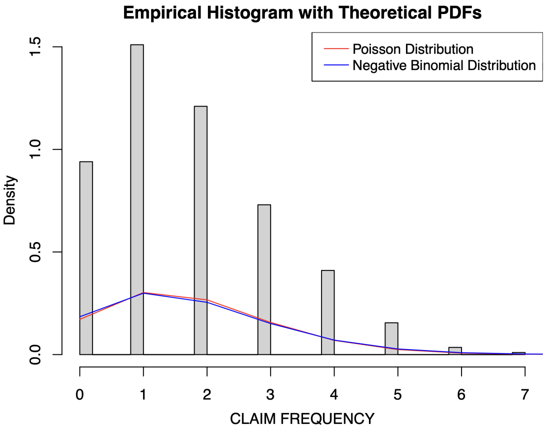
3.1.2 Model Selection

Having fitted both of our candidate model to the data set, it is now possible to compare their adequacy. In the following section we will use a series of model diagnostics as well as appropriate test statistics as criteria to select the distribution we will chose to retain.

Probability Density Function

First among those tests is the plotting of the theoretical probability density functions of the fitted Poisson and Negative Binomial functions along with the density of the claim frequencies observed in the data set.

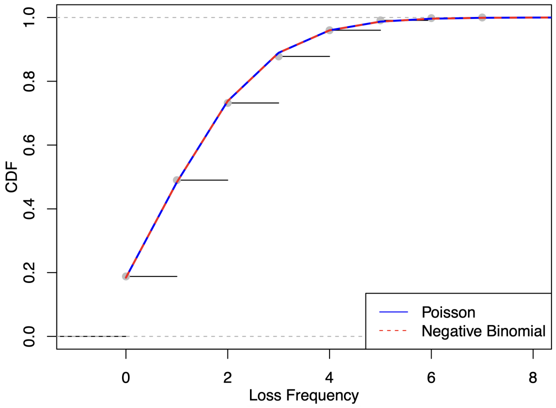
On this plot, we can observe that the Negative Binomial has more mass than the Poisson in zero and in the tail and less in the 2 to 3 range.



FF Plot

Secondly, we constructed the so-called F-F plots of our candidate distributions and the claim frequencies observed in the data set. This type of plot consists of the display of the cumulative distribution function of a theoretical distribution against the empirical cumulative distribution function of the data set, that is the ordered points from said set against their percentiles.

On this plot, one can see that both distributions are good fits for the modeling of the claim frequency.



Chi-squared test

Finally, we evaluate the pertinence of our models through a mathematical test of their fit. In order to assess the credibility of our hypothesis that the observed sample could be originating from our candidate fitted distributions, we consider the following hypotheses:

$$H_0\text{: Our sample follows the model distribution}$$

$$H_1\text{: Our sample does not follow the model distribution}$$

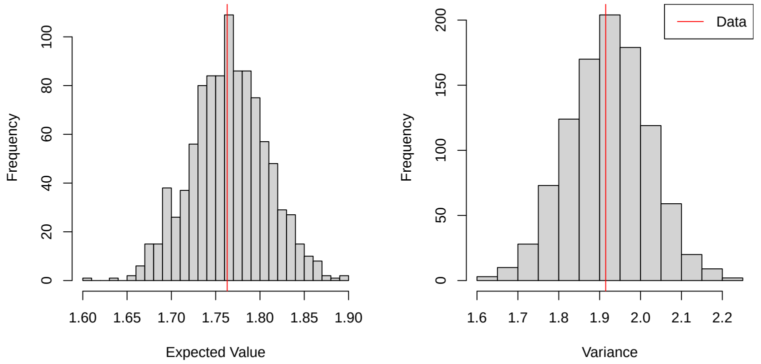
To test the null hypothesis we arranged our observations appropriately and to proceed with the chi-square Pearson test, ultimately resulting in a p-value equal to 0.2303 for the fitted Poisson distribution. Similarly, for the fitted Negative Binomial distribution, we obtained a p-value also equal to 0.2303. Consequently, we could not reject the null hypothesis that our observed claim frequency data could have been derived by such distributions with the estimated parameters.

Result

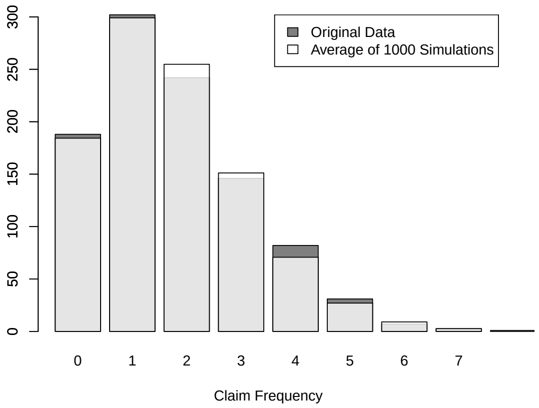
Having proceeded with these tests, we have determined that both the Poisson and Negative Binomial distributions equipped with the parameters estimated above represent acceptable distributions for the modelling of the claim frequencies observed in our data set. However, in spite of its greater complexity due to its two parameters, we choose to retain the Negative Binomial distribution for our model. Indeed, in addition to the marginally greater fit observed during the series of tests above, these multiple parameters allow for a greater flexibility valuable in our case. In particular, this additional degree of freedom allows us to overcome the limitation of the Poisson distribution of having a mean equal to its variance, which does not correspond to what is observed in our sample.

3.1.3 Monte Carlo

In order to test our claim frequency model, we used the Monte Carlo method and simulated 1000 portfolios. Then, we compared the obtained probability mass function to the empirical one derived from the claim frequencies observed in the data set. These simulations yielded a mean claim frequency of 1.76 when the sample mean of those observed in the data set was of 1.763. We conducted a t-test to verify whether the mean and the variance of the portfolios claim frequency could respectively serve as a mean and a variance to our sample simulations. At 5% confidence level we were unable to reject the hypothesis that the claim frequency mean was different from our sample simulations.



We also adopted an alternative approach suitable for discrete distributions. We averaged the results from 1000 simulations for each frequency bin and compared them to our portfolio data. The results below demonstrate a strong fit between the negative binomial simulations and our actual claim frequency data in our portfolio.



3.2 Claim Severity Modelling

3.2.1 Candidate Models

As for the model of claim frequency, we pre-select a model for the claim severity by taking into consideration a criterion on the values taken by this metric. The claim severity of a given claim can take any value that is a positive real number. Hence, the models that we consider have to accommodate a continuous positive random variable that has no known upper bound. Indeed, as for the claim frequency, although the observations of the data set have a maximum value of a claim amount of 1442.0, there is no theoretical argument for the existence of such a bound for the severity of a given claim. Thus, we pre-select for our model the following common positive continuous distributions as a result of the use of this criterion.

Exponential

First among our candidate for the model of the claim severity is the exponential distribution. It is a common option for a light tailed model, in particular due to its simplicity afforded by its unique real positive parameter λ. The simple expression of the mean of such a distribution allows for a simple approximation of this parameter by the method of moments estimator $\hat{\lambda}$ derived as follows:

$$X \sim \text{ Exp}(\lambda) \quad \Rightarrow \quad \mathbb{E}(X) = \frac{1}{\lambda} \quad \Rightarrow \quad \lambda = \frac{1}{\mathbb{E}(X)}$$

yielding the estimator $\hat{\lambda} = \frac{1}{\bar{x}_n}$, taking in our case a value of 0.002303217

Gamma

Then, we consider the Gamma distribution as a candidate to model the claim severity. It is a popular choice for such models, due to its versatility, in particular with respect to skewness and different tail behaviors. This distribution admits two real positive parameters, namely a shape parameter k and a scale parameter θ. As such distributions admit well-defined moments, expressed with closed-form formulas, it is possible to approximate these parameters with estimators $\hat{\theta}$ and $\hat{k}$ derived through the method of moments as follows:

$$X \sim \text{Gamma}(k, \theta) \Rightarrow \mathbb{E}(X) = k \cdot \theta \quad \text{and} \quad \text{Var}(X) = k \cdot \theta^2$$

allowing us to express the parameters and estimators as:

$$k = \frac{\mathbb{E}(X)^2}{\text{Var}(X)} \quad \text{and} \quad \theta = \frac{\text{Var}(X)}{\mathbb{E}(X)} \quad \Rightarrow \quad \hat{k} = \frac{\overline{x}_n^2}{s_n^2} \quad \text{and} \quad \hat{\theta} = \frac{s_n^2}{\overline{x}_n}$$

In our case, the estimators $\hat{k}$ and $\hat{\theta}$ take the values 5.678094 and 76.46497 respectively.

Inverse Gaussian

Next among our candidates for a model of the claim severity, we consider the Inverse Gaussian distribution. It is a popular choice among heavy-tailed distributions, making it fit for the accurate modeling of extreme events, which in the case of claim severity can be particularly important. The Inverse Gaussian distribution admits a mean parameter μ and a shape parameter λ, both positive real numbers. Once again, these can be approached by the method of moments, and yield estimators $\hat{\mu}$ and $\hat{\lambda}$ as follows:

$$X \sim \text{IG}(\mu, \lambda) \quad \Rightarrow \quad \mathbb{E}(X) = \mu \quad \text{and} \quad \text{Var}(X) = \frac{\mu^3}{\lambda} \quad \Rightarrow \quad \lambda = \frac{\mathbb{E}(X)^3}{\text{Var}(X)}$$

Eventually, those estimators equal :

$$\hat{\mu} = \overline{x}_n \quad \text{and} \quad \hat{\lambda} = \frac{\overline{x}_n^3}{s_n^2}$$

and the values 434.1753 and 2465.288 respectively in our case.

Log-Normal

Finally, we consider an alternative common heavy-tailed distribution for our model of the claim severity, that is the Log-Normal distribution. As for the Inverse Gaussian distribution, this property makes it popular for applications in insurance. It also admits two parameters, both positive and real valued, the location, or mean of its logarithm, μ and its variance σ^2. Again, it is possible to approximate these parameters by the method of moments estimators $\hat{\mu}$ and $\hat{\sigma}^2$, derived as follows:

$$X \sim \text{LN}(\mu, \sigma^2) \quad \Rightarrow \quad \mathbb{E}(X) = \exp(\mu + \frac{\sigma^2}{2}) \quad \text{and} \quad \text{Var}(X) = (\exp(\sigma^2) - 1) \cdot \exp(\mu + \frac{\sigma^2}{2})^2$$

allowing us to express the parameters as:

$$\mu = \ln(\mathbb{E}(X)^2) - \frac{1}{2} \cdot \ln(1 + \frac{\text{Var}(X)}{\mathbb{E}(X)^2}) \quad \text{and} \quad \sigma^2 = \ln(1 + \frac{\text{Var}(X)}{\mathbb{E}(X)^2})$$

yielding the estimators:

$$\hat{\mu} = \ln(\overline{x}_n) - \frac{1}{2} \cdot \ln(1 + \frac{s_n^2}{\overline{x}_n^2}) \quad \text{and} \quad \hat{\sigma}^2 = \ln(1 + \frac{s_n^2}{\overline{x}_n^2})$$

that in our case respectively take the values 5.99234 and 0.162217.

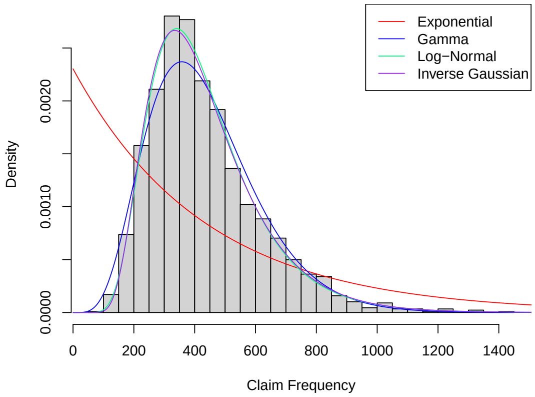
3.2.2 Model Selection

As for claim frequency, having obtained these fitted candidate models, we proceed by evaluating and comparing their adequacy for our goal to model the claim severity. Accordingly, the following section consists of a series of tests aiming to ease the selection of the distribution most suitable for our model.

Probability Density Function

As for the selection of the claim frequency model, we first consider a plot of the theoretical probability density functions of our fitted candidate models along with their empirical counterpart for the claim severity.

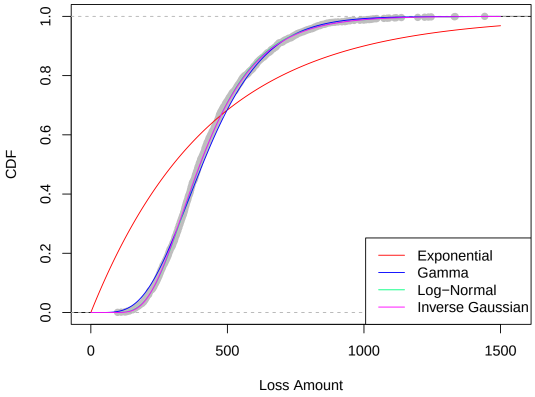
On this graph, we can see that the Exponential distribution has far too much mass near zero. The other candidate models seem better fitted for our purpose. We can also observe that the Gamma distribution stands out form the Inverse Gaussian and the Log-Normal by its lack of a heavy tail.



FF Plot

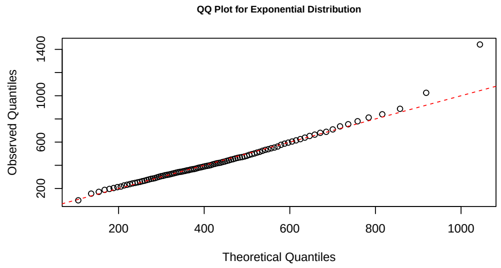
Then, we consider the F-F plots for our fitted candidate distributions and the values for the claim severity observed in our data set.

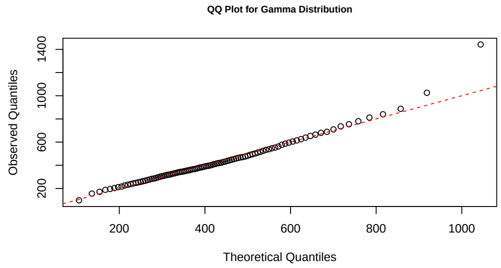
As above, this graph disqualifies the Exponential distribution as a candidate for out model. In addition to this, it reinforces the apparent superiority of the fits of the Inverse Gaussian and the Log-Normal over that of the Gamma distribution.

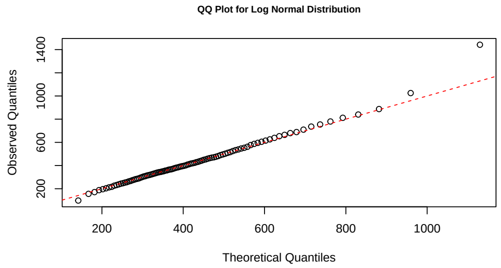


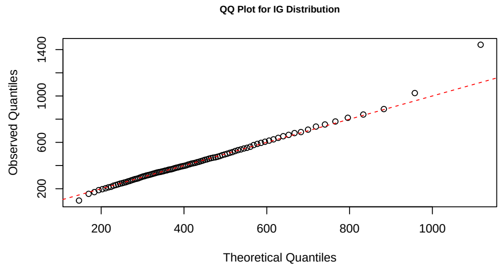
Q-Q Plot

Our final graphical criteria for the selection of the claim severity model is their Q-Q plots.









This final series of plots corroborates our assessment that the Inverse Gaussian and the Log-Normal distributions represent a superior model, especially for the modeling of the tails.

Kolmogorov-Smirnov Test

Our final criterion is to run the Kolmogorov-Smirnov test and set a null hypothesis similar to the chi-squared test of the claim frequency. In principle, the Kolmogorov-Smirnov test is a measure of the distance between the cumulative distribution functions of our model and the empirical distribution function of the observed claim severity and an indicator of whether our observations are indeed derived from the distribution in question. In summary, the resulting p-values allowed us to reject the hypothesis that our sample severity data were derived from the exponential distribution, while not rejecting the null hypothesis for the other distributions.

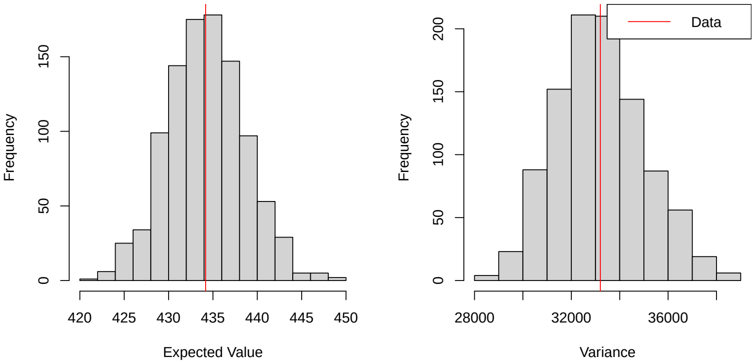
This test yields encouraging results, as it supports our rejection of the Exponential distribution, as it obtains a p-value close to zero. The testing of the Gamma distribution results in an encouraging p-value of 0.9986 which supports its use. However, both the Inverse Gaussian and the Log-Normal distributions exceed this result, as their tests both yield a p-value of 1. The D-value finally breaks the tie between these two distributions, as they are respectively equal to 0.019648 and 0.016971, thus confirming that the Log-Normal is a slightly better fit.

Results

As a result of these tests, we have chosen to retain the Log-Normal fitted distribution as a model for the claim severity. Indeed, the graphical tests strongly indicated that both of the heavy-tailed distributions, namely the Inverse Gaussian and the Log-Normal, were a superior match for our data points. Then, the result of the Kolmogorov-Smirnov Test suggested that the fit of the Log-Normal distribution was marginally superior to that of the Inverse Gaussian, motivating our final choice, along with the Log-Normal's greater mathematical ease of use.

3.2.3 Monte Carlo

As for the claim frequencies, we proceed by testing our model for the claim severity using the Monte Carlo method. To this end, we simulate that of 1000 claims and compare the obtained probability mass function to the empirical probability mass function given by the claim amounts observed in the data set. As a result of these simulations, we have obtained a sample mean claim severity of 434, when that of those observed in the data set was 434.1753. We also conducted a t-test to verify whether the mean and the variance of the portfolios claim severity could serve as a mean and a variance to our sample simulations. At 5% confidence level we were unable to reject the hypothesis that the claim severity mean was different from our sample simulations.



3.3 Variance Reduction Techniques

Now that we conducted our Monte Carlo Estimator we shift our attention to applying variance reduction techniques to speed up the simulation time. We simulated 1000 samples using different variance reduction techniques, namely; Antithetic Method, Control Variate Method and Importance Sampling Method. We tried to also employ the Conditional Monte Carlo Method and the Stratified Sampling Methods however, our portfolio did not exhibit any patterns that we could leverage to employ such methods. In the following plots we will show the histograms of 1000 simulated Estimators' variances and compare them to the Crude Monte Carlo method. at the top right of each plot we will present a sample representation of the new expected value from one of the 1000 simulations.

3.3.1 Antithetic Method

This method relies on creating a negative dependence structure between the uniformly distributed random numbers such that the variance is lower.
The new theoretical variance using this technique is provided below:

$$\mathrm{Var}(\hat{f}_{\text{anti}}(U)) = \frac{1}{2} \cdot \frac{\mathrm{Var}(f(u))}{n} + \frac{1}{2n} \cdot \mathrm{Cov}(f(u), f(1-u))$$

3.3.2 Control Variates Method

This method introduces a new random variable X_Y known as the control covariate and simulates from it instead of the original random variable distribution. The idea is that the newly introduced random variable is simple to simulate from and correlates with the original random variable. As such we achieve an estimator with lower variance. The new theoretical variance using this technique is provided below:

$$\text{Var}(\bar{X}_Y) = \frac{1}{n \text{ Var}(X - Y)} = \frac{1}{n}(\text{Var}(X) + \text{Var}(Y) - 2 \text{ Cov}(X, Y))$$

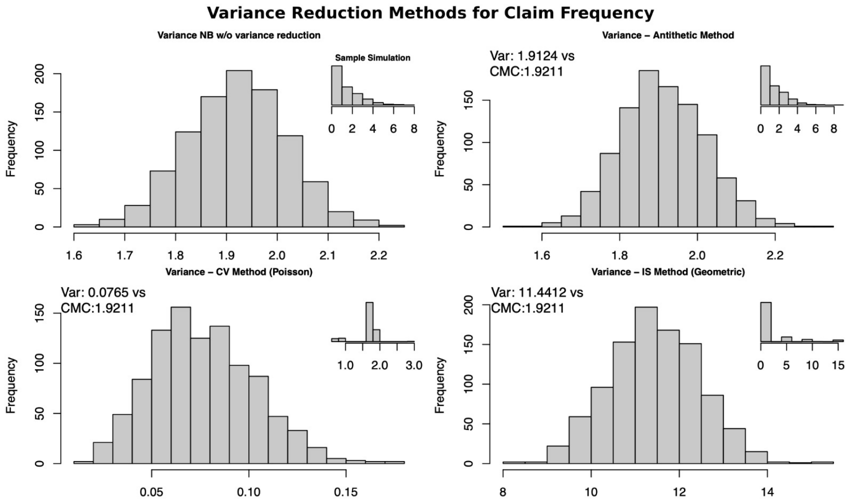
3.3.3 Importance Sampling Method

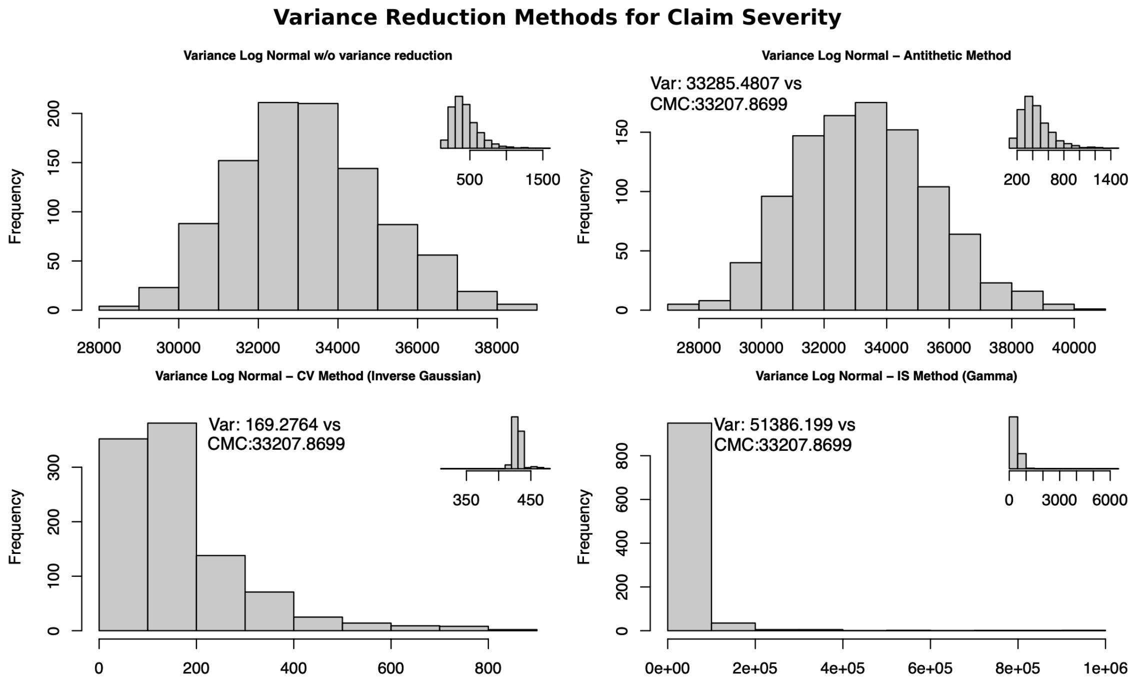
The Importance Sampling method also implements a technique that avoids sampling from the original distribution. We instead sample from a proposal distribution $\tilde{f}$. The Importance Sampling Estimator is then defined by:

$$\begin{aligned} \text{Var}(X^{IS}) &= E_{\bar{f}}\left[\left(\frac{f(X) \cdot g(X)}{\bar{f}(X)}\right)^2\right] - E_{\bar{f}}\left[\frac{f(X) \cdot g(X)}{\bar{f}(X)}\right]^2 \\ &= \int \frac{f(x) \cdot g(x)^2}{\bar{f}(x)} f(x) d\mu(x) - (E[g(X)])^2 = E\left[\frac{f(X) \cdot g(X)^2}{\bar{f}(X)}\right] - E[g(X)]^2 \end{aligned}$$

Where $W_i = \frac{f(x_i)}{\bar{f}(x_i)}$ is known as the weight. In our case the function $g(x)$ was always the identity function.

3.3.4 Results





The Crude Monte Carlo Method exhibits a variance ranging from 1.6 to 2.2, with an average value of 1.9161. The Antithetic Variance Method shows only a negligible improvement, with a mean variance of 1.9144. Similarly, for the severity model, the Crude Monte Carlo Estimators have a variance ranging from 28,000 to 40,000, with a mean value of 33,312. The Antithetic Variance Method also shows a minimal improvement, with an average variance of 33,229.

To reduce variance, we applied the Control Variates Method by selecting a Poisson random variable with a parameter of $\lambda = 1.736$. This resulted in a new variance ranging from 0.01 to 0.15, approximately 27 times lower than the Crude Monte Carlo Method. For the severity model, we selected an Inverse Gaussian random variable with parameters $\mu = 434.1753$ and $\lambda = 2465.288$. This led to a variance reduction of approximately 165.0591, equivalent to around a 200-fold improvement.

We chose $\tilde{f}$ to be a geometric distribution with a parameter of $p = 0.0567215$ for the Importance Sampling Estimator. However, the imprecision of this estimator resulted in variances ranging from 0 to 15.8. Over 1000 samples, the variance converged to approximately 11.5, which is 10 times higher than the Crude Monte Carlo Method, rendering it as inefficient. The same pattern was observed for the severity model, where we selected $\tilde{f}$ to be a Gamma random variable with the parameters shape $= 5.678094$ and scale $= 76.46497$. The imprecise estimator led to a higher variance of approximately 62688.15.

4. Applications

4.1 Risk Premium

Risk premium in insurance context refers to the additional return that an insurance company demands in order to offset the exposure to a higher level of risk. In our calculations we consider risk neutral valuation. This means that we are not taking into account safety loading or additional expenses such as administrative fees, taxation, etc.

4.1.1 Risk Premium Calculation

Empirical Risk Premium

Based on the data the risk premium was calculated by the following formula:

$$\text{Empirical Risk Premium} = \frac{\text{Empirical Aggregate Claim Amount}}{\text{Number Of Policyholders}}$$

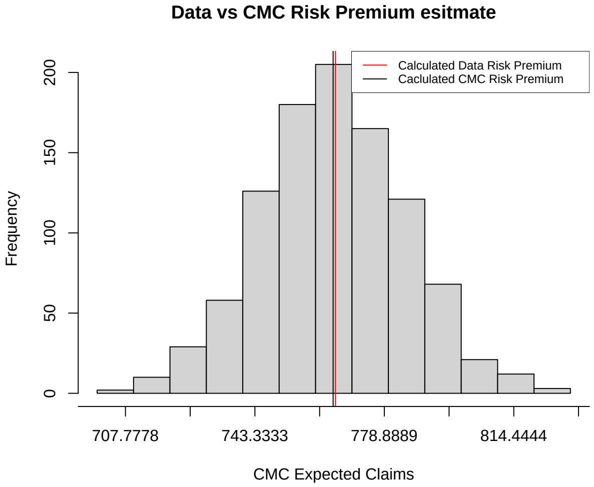
Monte Carlo Risk Premium

Based on our Monte Carlo simulations we computed the risk premium as follows:

$$\text{Monte Carlo Risk Premium} = \frac{\text{Monte Carlo Aggregate claim amount}}{\text{Number Of Policyholders}}$$

By using the Monte Carlo method we simulated 1000 portfolios of 1000 policyholders. For each portfolio, we calculated the Monte Carlo risk premium using the above formula.

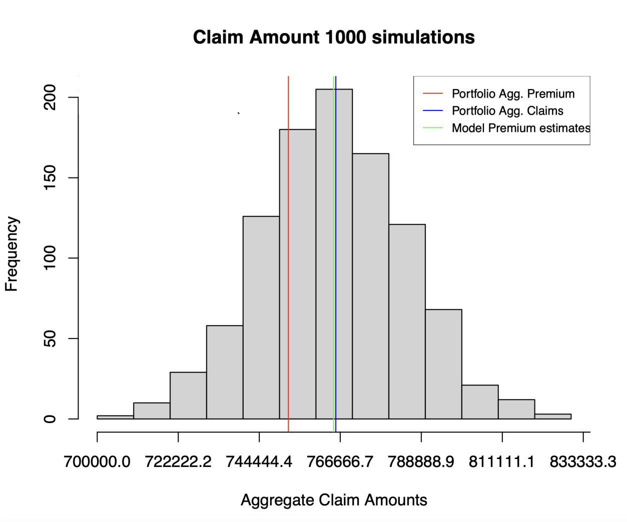
The calculated empirical risk premium was 765.451, while the Monte Carlo risk premium converged to an average value of 766.2344. These numbers represent the average premium charged by our company to a policyholder. The proximity of said numbers indicate a strong alignment with our model.



4.1.2 Risk Premium and Empirical Premium Analysis

Our portfolio has a total premium of 752,431 and an aggregate claims amount of 765,451. This results in a loss ratio of 101.73%.

Based on our model assumptions, we observe that the premium we currently charge our policyholders can only cover approximately one-third of the incurred losses.



4.2 Risk Measures

A risk measure is a mathematical function or metric used to quantify and assess the level of risk associated with an uncertain event or investment. It provides a framework to measure the potential losses or adverse outcomes that may arise from taking on a particular risk.

Under the European directive Solvency II Insurance companies are obligated to maintain an economic capital that ensures the occurrence of ruin with a maximum probability of α within a one-year period. This measure is commonly known as value at risk (VaR_α). Solvency II mandates setting the probability α at 0.5%.

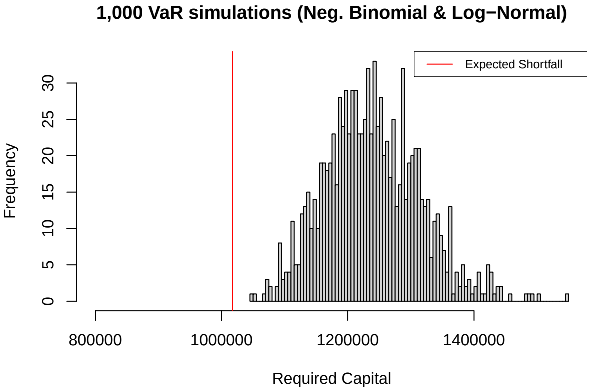
Another commonly known risk measure that is mandated by other regulatory authorities is known as the Expected Shortfall (ES_α). The expected shortfall at percentage α is the expected return on the portfolio in the worst α of cases. In Switzerland for instance the regulatory authority FINMA mandates to set the level α to 5%.

4.2.1 Mathematical Formulae

The *value at risk at level* α is given by: $\mathrm{VaR}_\alpha(X) = q_\alpha = F^{-1}(\alpha) = \inf\{x \mid F(x) \geq \alpha\}$

The *expected shortfall at level* α is given by: $\mathrm{ES}_\alpha(X) = \mathbb{E}(X \mid X > \mathrm{VaR}_\alpha(X))$

4.2.2 Results



After running 1000 simulations we get a VaR ranging between 1,046,588 and 1,549,250 at the 0.5% level. The average VaR value of those simulations at the 0.5% level was 1,239,205. This means that for our portfolio to comply with the European regulation of Solvency II we have to hold capital of around 1,239,205. This implies that there is approximately a 0.5% probability of experiencing ruin (total loss of capital) with the given amount of capital.

The Expected Shortfall at the 5% level appears to be less strict in terms of capital requirement. The simulation provided values ranging between 929,414 and 1,129,192. The average Expected Shortfall value at the 5% level was 1,017,706. This amount of capital would be able to survive experiencing the average of the worst 5% scenarios.

5. Model explanation

After conducting our analysis of the portfolio, we have reason to believe that losses follow a particular pattern. It appears to be that on average a policyholder faces 1.763 claims on a yearly basis and with an amount of 434.1753. If we multiply those numbers we get an aggregate claim of 765,451.05 which is very similar to what we currently experience in our company, our portfolio experiences an aggregate claim amount of 765,451.

To reach numbers with high accuracy as such we construct two mathematical models. You can view these models as formulae aiming to resemble reality. To select such models (formulae) we try to come up with mathematical functions that mimics our data. To put into context, we try to find mathematical function that behave in a similar way as the average number of claims, and another one for the average size of claims.

This approach of modeling both the frequency and severity of claims is referred to as a collective risk model. For the number of claims we used a model called the negative binomial, it represents the probability of observing a certain number of successes (e.g., sales, conversions, or customer acquisitions) before a specified number of failures (e.g., unsuccessful attempts, customer churn) in a series of independent events. This model (formula) is characterized by two parameters the number of successes (r) and the success probability (p). The negative binomial is useful in claim frequency modelling.

For modeling claim severity we used the Log-Normal distribution. It is often used to model quantities that are always positive and have a wide range of possible values. The log-normal distribution has some interesting properties. One key property is that it is skewed to the right. This means that there is a long tail on the right side of the distribution, indicating the possibility of some very high claims. Integrating these models allows the company to align its pricing and marketing strategies,

leading to a comprehensive understanding of the risks in its portfolio. This integration facilitates data-driven decisions that ensure the company's sustainability and foster market growth.