Collegio Carlo Alberto

Assignment on Insurance Products

Ramo Danni

Alessandro Canola Gavioli[1], Naimeh Ghare Daghi[1], Pasquale Paolicelli[1], and Daniel Sulluchuco[1]

[1]Master in Insurance and Innovation

January 7, 2025

Abstract

Understanding how to calculate premium prices is an important part of financial decision-making. In this report, we'll take a closer look at a Loss dataset provided to us and explore the process of determining premium prices.
The goal is to analyze the data, apply the right methods to describe them and finally calculate the premiums in several different conditions.
In the last part, we proceed to analyze how reinsurance can be used to modify the distribution of losses.

1 EDA

The dataset analyzed comprises 20,000 records, each representing the financial loss incurred by a company for individual claims.

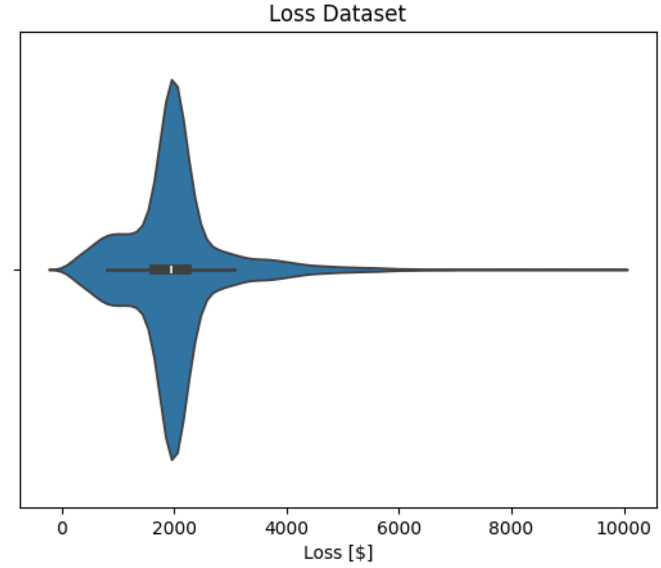

Figure 1: Loss dataset - Box Plot. The continuous line represents the Kernel Density Estimate KDE.

Statistic	Loss [k · $]
Maximum Loss	9.80
Minimum Loss	0.03
Mean Loss	2,00
Median Loss	1.96
StD	0.91

Table 1: Summary Statistics for Loss Claims

Most claim amounts are concentrated around the median value of 1,955.47, but the presence of a few high claims skews the data distribution. Additionally, the minimum compared to the maximum indicates a stark difference in claim magnitudes, which highlights the diverse nature of the dataset. This features suggest a large asymmetry in the dataset.

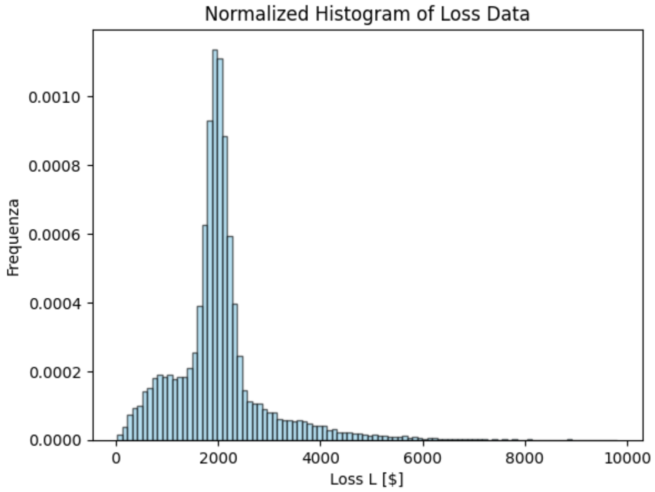

Figure 2: Normalized Losses L distribution dataset

Figure 2 shows a noticeable asymmetry with a long tail extending towards large claim amounts. However, most of the data is clustered around the mean and appears roughly Gaussian in shape. There's also an underlying "background" component that seems to drive the long tail towards higher claim values, with the peak of the distribution around ∼ 1000$.

2 Key data features analysis

The graphs of the Empirical Distribution, Empirical Loss Index, and Mean Excess Function are essential visual tools for analyzing the loss data of an insurance company. These representations provide critical insights into the behavior of losses, helping to understand their distribution, frequency, and severity. By examining these graphs, we can:

1. Empirical Distribution: Visualize the cumulative probability of losses to identify patterns and determine probabilities for specific thresholds.
2. Empirical Loss Index: Evaluate the relative magnitude and frequency of losses, offering a comprehensive view of the risk exposure.
3. Mean Excess Function: Analyze the expected loss beyond a given threshold, which is crucial for setting reinsurance limits or estimating tail risks.

These graphs collectively enable better decision-making for risk management, pricing, and strategic planning in the insurance context. To plot these graph one important thing was to order all the losses in size.

Empirical Distribution

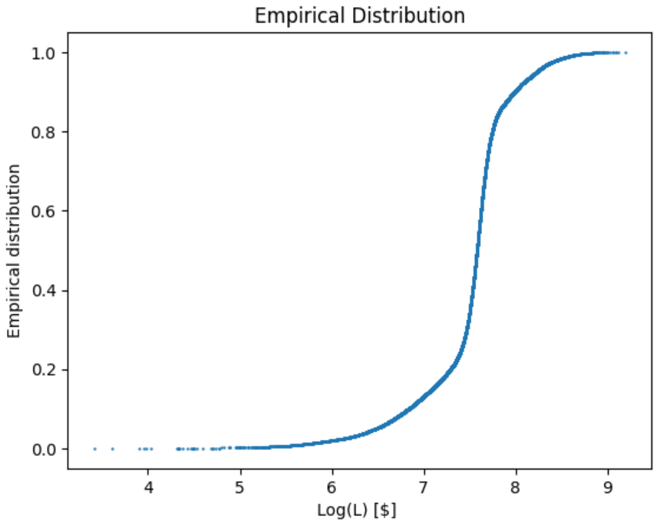

Figure 3: Empirical Distribution

- **Observation**: The graph depicts the cumulative distribution of logged claim sizes, showing how the probability of claims accumulates as the size increases. The steep increase indicates a concentrated range, between 7 and 8, where most claim sizes occur. The flattening near the upper end suggests that extremely large claims are rare.

Empirical Loss Size Index

The Empirical Loss Size Index is defined as follows:

$$I_x[i] = \frac{\mathbf{E}[X\mathbf{I}_{x<x_j}]}{\mathbf{E}[X]} = \frac{\frac{1}{N}\sum_{j=1}^{N} x_j \cdot \mathbf{I}(x_j \leq x[i])}{\frac{1}{N}\sum_{j=1}^{N} x_j} \tag{1}$$

This is the empirical plot obtained with our dataset.

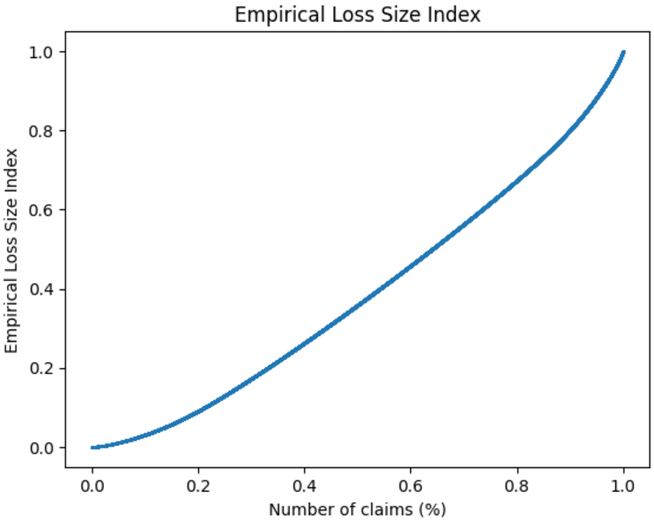

Figure 4: Empirical Loss Index

- **Observation**: This graph illustrates the relationship between the cumulative proportion of claims and the corresponding proportion of the total loss attributed to those claims. The linearity observed in the graph indicates an absence of a small subset of claims disproportionately contributing to the overall loss. In other words, the loss distribution appears evenly spread across claims, with no indication of extreme losses dominating the total.

Mean Excess Function

The Mean Excess Function, or Expected Shortfall, is defined as:

$$E[i] = \frac{1}{N}\sum_{j=1}^{N}(x_j - x[i]) \cdot \mathbf{I}(x_j > x[i])$$

This is the empirical plot obtained with our dataset.

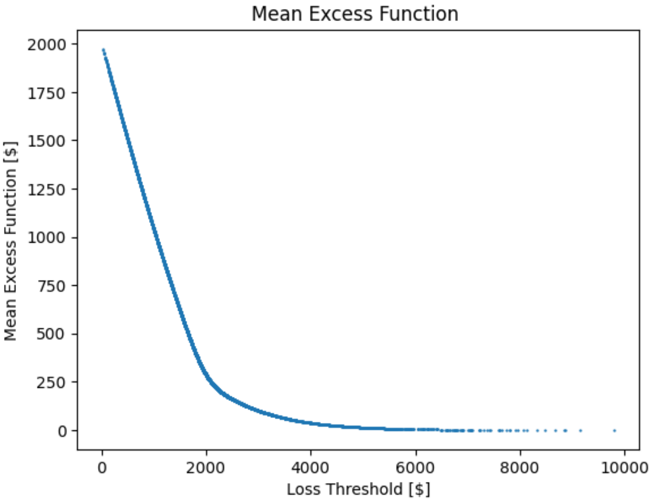

Figure 5: Mean Excess Function

- **Observation**: This plot illustrates the average loss amount exceeding a given threshold. The sharp decline suggests that the average loss decreases as the threshold increases, which is typical when extreme events have already been captured at lower thresholds. The fact that it approaches zero for thresholds above 4000 indicates that there are very few claims exceeding this value, highlighting the limited presence of extreme loss events in the dataset.

These graphs collectively provide a comprehensive view of the loss behavior and help in making informed decisions on risk management and premium pricing.

3 Data model representation

This section aims to find the best model that fits the data, to achieve this two methods have been tested, being the first one the Maximum Likelihood (ML) and the second one Methods of moments.

Maximum Log Likelihood (ML)

Taking into account the considerations made about Figure 1, a two functions convolution have been chosen to fit the Loss data. In particular two combinations have been tested:

1. LogNormal $\oplus$ Normal
2. Log Normal $\oplus$ Gamma

This combination choice is due to considerations about the data asymmetry they may be able to represent.

LogNormal $\oplus$ Normal - Fit_1

$$f_{LogN}(x \mid \theta, s) = \frac{1}{x\theta\sqrt{2\pi}} \exp\left(-\frac{(\ln(x) - s)^2}{2\theta^2}\right), \quad x > 0$$

$$f_N(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x > \infty$$

A weigh w is adopted to measure their correspondent contribution. In total, 5 parameters define the adopted model, being $w, \theta, s, \mu, \sigma$. [1]

$$f_{Fit1} \;= w \cdot f_{LogN}(x \mid \theta, s) + (1-w) \cdot f_N(x \mid \mu, \sigma)$$

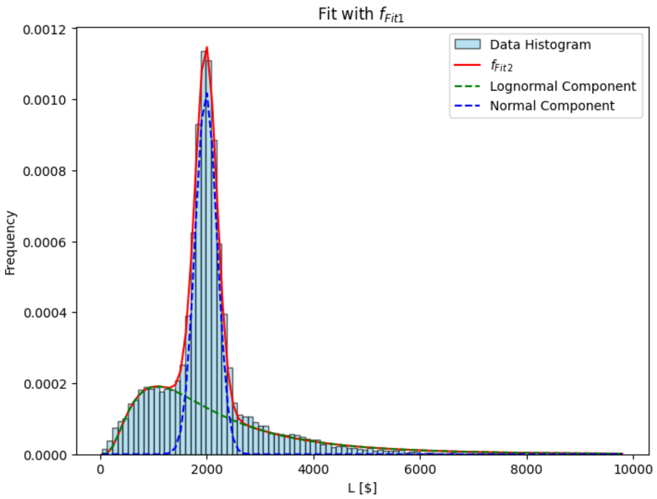

Figure 6: Fit over the Loss distribution with f_{Fit1} using the ML method

The fit over the Loss distribution Figure 6 gives us an almost satisfying result, however it underestimates the distribution around 3000$ due to a residual asymmetry which this function was not able to represent.
Indeed, the $\chi^2/Df = 20.26$, which highlight the fact that this convolution function does not represents the Loss data.

LogNormal $\oplus$ Gamma - Fit_2

In order to include a larger degree of asymmetry, a Gamma distribution function has been tested, convoluted with a LogNormal distribution.

$$f_{Gamma}(x \mid \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)}, \quad x > 0$$

Also in this case the parameters are five, being $w, \theta, s, \alpha, \beta$

$$f_{Fit2} \;= w \cdot f_{LogN}(x \mid \theta, s) + (1-w) \cdot f_{Gamma}(x \mid \alpha, \beta)$$

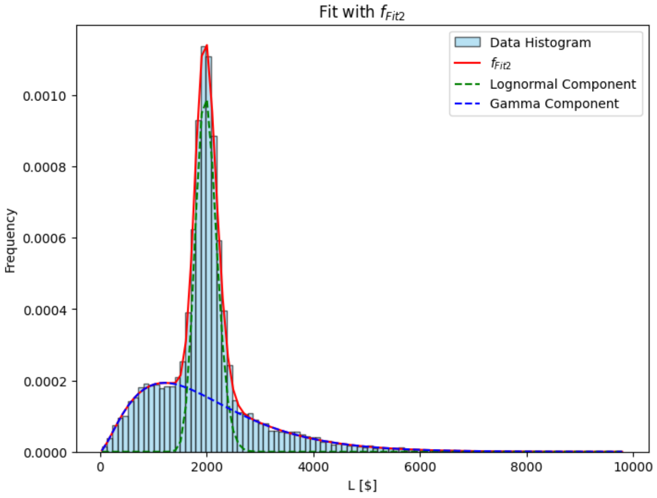

Figure 7: Fit over the Loss distribution with f_{Fit2} using the ML method

In this case the fit provides a better representation of the dataset, doing right were the f_{Fit1} failed. As matter

[1] The fit does not take directly the s parameter, but its exp(s) value.

of fact, it has been obtained a $\chi^2/Df = 0.93$, which is almost a perfect result.

To validate the results the p_{value} has been calculated for both fits, f_{Fit1} and f_{Fit2}.

$f(x)_{Fit}$	p_{value}
f_{Fit1}	~ 0
f_{Fit2}	81.54%

Table 2: p_{values} table for both tested models. The confidence level is set to be 5%

A final confront between both, f_{Fit1} and f_{Fit2} model is presented and two statistical measures are implemented, being the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion). These measures are define through two fit's quantities, the number of parameters of the model $k_{\hat{\theta}_{ML}} = 5$ in this case for both PDFs. The Maximum Log Likelihood value $\log(\mathcal{L}^{f_{Fit}}(x \mid \hat{\theta}_{ML}))$ and lastly the number of data points n, the dataset size.

$$AIC = -2 \cdot \log(\mathcal{L}^{f_{Fit}}(x \mid \hat{\theta}_{ML})) + 2 \cdot k_{\hat{\theta}_{ML}}$$

$$BIC = -2 \cdot \log(\mathcal{L}^{f_{Fit}}(x \mid \hat{\theta}_{ML})) + 2 \cdot \log(n)$$

model	**AIC**	**BIC**
$f_{Fit\,1}$	$318'046.11$	$318'085.63$
$f_{Fit\,2}$	$317'080.61$	$317'120.12$

Table 3: AIC and BIC values obtained with their respectively fit parameters.

These statistical models have have pros and cons. However, in both cases a lower BIC and AIC value indicates a better Fit model adaptation to the dataset.

All the performed test unequivocally indicates that the $f_{Fit\,2}$ model fits at best the dataset. Hence, *the PDF convolution LogN + Gamma has been chosen to represent the whole dataset*, and later in this report its parameters will be used for the Monte Carlo (MC) simulation.

Parameter	**Value**
w	0.497 ± 0.005
θ	0.100 ± 0.001
s	1988.626 ± 0.118
α	2.503 ± 0.012
β	801.972 ± 0.113

Table 4: Parameters obtained with the f_{Fit2} model

Methods of Moments (MM)

To find empirically the 5 parameters for the f_{Fit2} model, the MM has been tested. However it hasn't produce result due to the complexity of the problem.
While the MM method is handy and easy to implement for models with few parameters, for instance the a Gaussian function $f(x|\mu, \theta)$, in which the arithmetic mean and empirical Standard Deviation of the data can be used to solve the two equation system

$$\begin{cases} \mu = \bar{x}(x) \\ \sigma = StD(x) \end{cases}$$

In the case under study, 5 known quantities and 5 equations are needed. The resulting system has to be solved. However the equations are not linear, which result into a tricky method to implement.
For this reason the optimal ML method results will be adopted for the following analysis in this report.

To visually summarize the results, a joint plot of the different is here provided. Where, the single distribution functions, *Gaussian*, *Gamma* and *LogN* have been tested to fit the data. However is visually evident that these function alone can not represent the loss dataset.

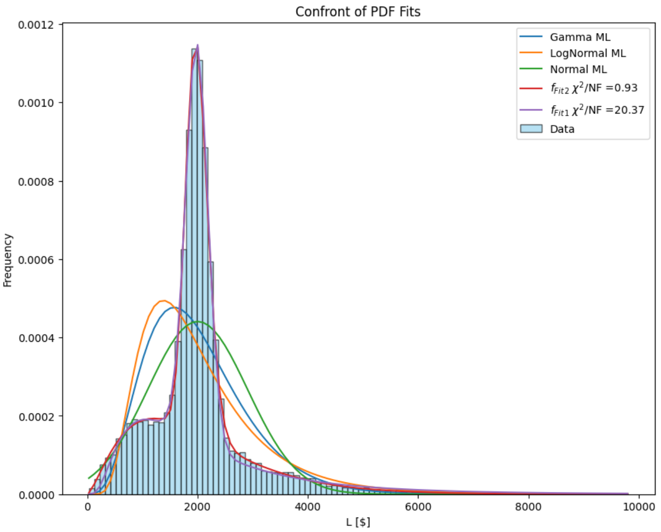

Figure 8: Visual summary plot for the distribution functions which have been tested over the Loss dataset

4 Monte Carlo simulation

The MC process relies on the f_{Fit2} PDF parameters estimated in the previous section, which will allow us to generate the Loss (L) amounts in line with the initial Loss distribution dataset.
This case study will consider a portfolio with $n_{policies} = 10'000$ policies. For each contract, a number N of claims will be generated accordingly to a Negative Binomial (NBinom) with parameters $r = 8$ and $p = 0.8$

$$P(X = n) = \binom{n + r - 1}{n} \cdot p^r \cdot (1 - p)^n, \quad n \geq 0$$

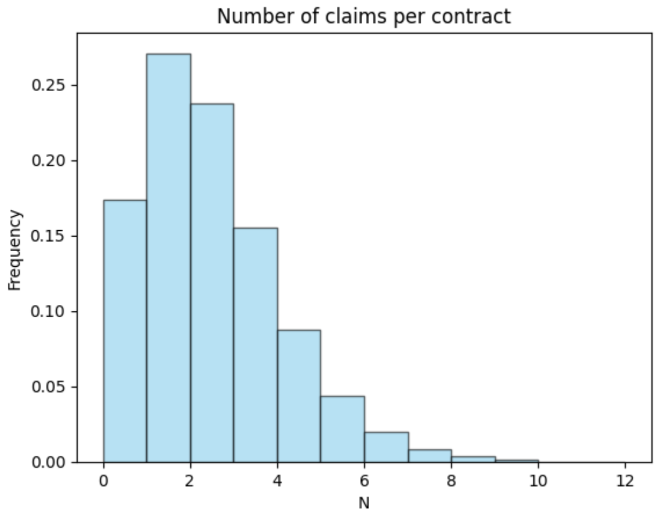

Figure 9: Negative Binomial distributed sample of $n_{policies}$ contracts, normalized

The expectation value of the given NBinom is $E[n] = r \cdot \frac{1-p}{p} = 2$. This expected value will be set as default for the generation of the number of claims for the portfolio.
The total amount of generated claims for the whole portfolio is roughly 20'000, the same order of magnitude of the initial dataset.

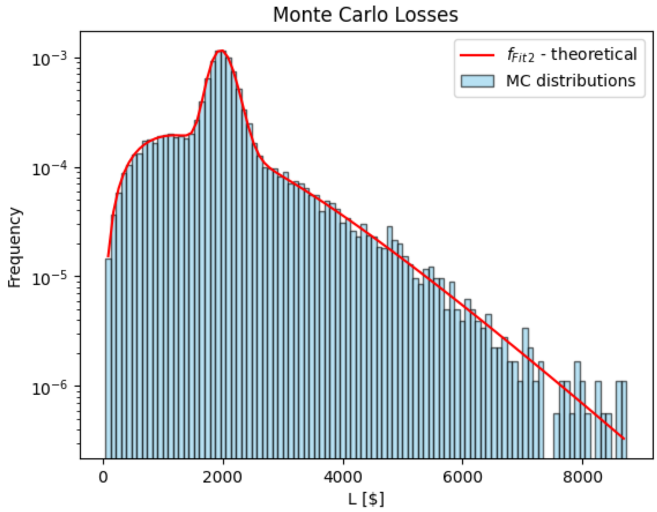

Figure 10: Loss values distribution generated with the MC method in blue, in log scale frequencies and normalized, in red the f_{Fit2} PDF expectation

Figure 10 shows a perfect Loss values generation over almost the entire range. However, the frequencies for the highest Loss values, on the extreme on the right tail, the obtained frequencies have a surplus with respect to the expected values. However, they are ultra rare events, and we expect to not have an appreciable effect over the final results.

The total amount Loss per contract has been computed as [2]

$$S_i = \sum_{j=1}^{N} L_{ij}$$

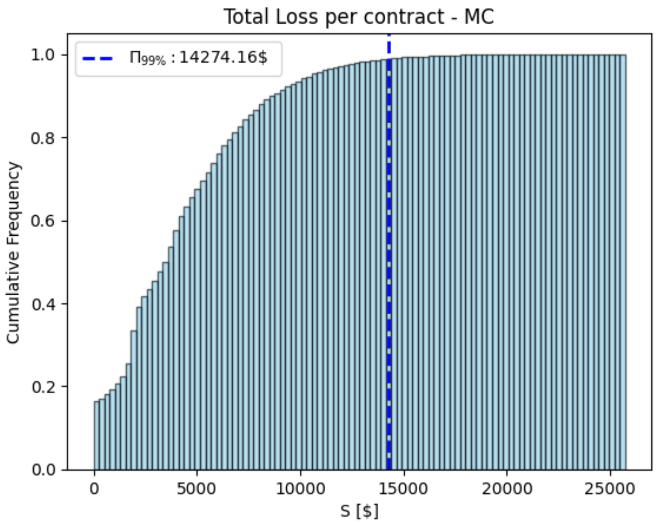

Figure 11: Cumulative S distribution. The Vertical line represents the Pure Premium price determined through the Percentile Premium method at 99% level of confidence.

Fair Premium price determination

The Fair Premium price Π_{fair}^{MC} has been evaluated as follow

$$\Pi_{fair}^{MC} \equiv E[L] \times E[N]$$

The $E[L]$ is the expected value for the f_{Fit2} PDF, which is equal to the weighted sum of the convoluted PDFs

$$E_{f_{Fit2}}[L] = w \cdot E_{LogN}[L] + (1 - w) \cdot E_{Gamma}[L]$$

The expected Loss value is $E_{f_{Fit2}}[L] = 2002.85\,\$$. Hence, the fair premium price has been found to be $\Pi_{fair}^{MC} = 4005.70\,\$$.

Pure Premium price determination

The pure premium price has been determined through the cumulative distribution plot of S. Particularly, the premium has been calculated thorugh the cost of capital using risk measures based on Value at Risk (VaR_α).
The VaR_α has been calculated at three levels of confidence $\alpha = (95\%\,, 99\%\,, 99.5\%)$

[2] The number of bins for the S distribution will be set to 100 as default

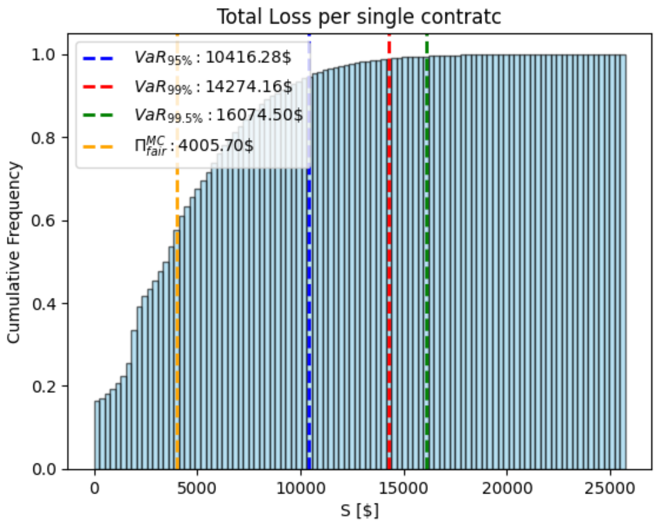

Figure 12: Cumulative distribution function for S. The vertical colored line bars indicate the VaR_α at different confidence levels and the Fair Premium price.

Lastly, the premium is calculated as follow, where $r_{CoC} = 9\%$

$$\Pi^{\alpha}_{CoC} = \Pi^{MC}_{fair} + r_{CoC} \cdot VaR_\alpha$$

VaR_α	Π^{α}_{CoC}
$\alpha = 95\%$	4943.17$
$\alpha = 99\%$	5290.38$
$\alpha = 99.5\%$	5452.41$

Table 5: VaR_α values for the MC simulation

5 BootStrap Simulation

This method does not make assumptions about the Loss distribution. Instead, it uses the empirical Loss distribution to extract randomly Loss values from it.
This method considers all the Loss values in the dataset evenly probable to be extracted, which means that Loss values with higher multiplicity and thigh clusters of data are more probable to be randomly extracted.

In this case of study, a sample of $nPolicies = 10'000$ are extracted and they constitute the homogeneous portfolio of the insurer company. The number N of claims for each contract is randomly extracted using the same parametrized $nBinomial$ used in the Monte Carlo section.

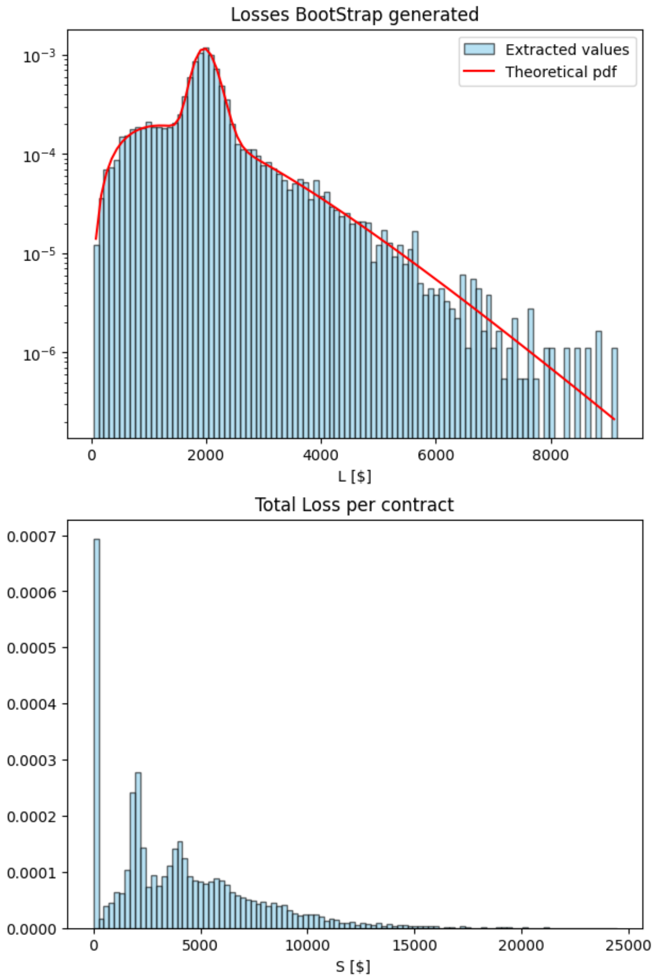

Figure 13: Bootstrap Loss amounts. The Top figure, in log scale, depicts in blue the randomly extracted distribution, in red the curve f_{Fit2} PDF expectation. The Bottom figure is the S amount loss per contract distribution.

The fair premium is computed and is equal to $\Pi_{fair} = E[L] \times E[N] = 4008.39\,\$$. The Expected Loss value has been computed as an empirical mean over the Loss generated values, $E[L] \equiv \bar{L} = 2004.20\,\$$.

6 Probability of Ruin

In this section the probability of ruin of the insurance companies will be computed, in two scenarios, being the first study-case the Boostrap scenario in section 5, and the MC scenario from section 4.

An insurance company faces the risk of ruin when it cannot cover all the claims made by its policyholders. This situation arises when the premiums collected are insufficient to offset the total losses incurred.
The probability of ruin, P_{ruin}, has been estimated by simulating the ratio of cases where the insurer experiences ruin to the total number of simulations performed.

BootStrap - case study

As afore anticipated, the same setup used in section 5 will be used for the simulation. However, only $nPolicies = 5'000$ will be generated.
The new Fair Premium price is $\Pi^{BS}_{fair} = 4014.28\,\$$.

The simulation comprises $N_{sim} = 500$. At each simulation, it has been checked if the company experiences ruin

$$\sum_{i=1}^{n_{Policies}} S_i > n_{Policies} \times \Pi_{fair}$$

This process produces a $P^{BS}_{ruin}(\Pi^{BS}_{fair}) = (51.40 \pm 1.58)\%$ [3]

Monte Carlo - case study

This case study performs an analogous procedure to estimate the probability of ruin of the company. In this case the exact setup used in the MC section is used, the number of simulation is $N_{sim} = 1000$.
This process produces a $P^{MC}_{ruin}(\Pi^{MC}_{fair}) = (51.10 \pm 1.58)\%$

The probability of ruin have been also estimated at $\Pi_{VaR_{95}}$, $\Pi_{VaR_{99}}$ and $\Pi_{VaR_{99.5}}$ levels, for both case studies. However, these estimations delivered a null probability in both cases, indicating that, the total amount of cached premium prices at this levels are enough to cover all the claims.

7 Reinsurance

To determine the premium for a reinsurance policy, we begin by fixing the cumulative distribution function (CDF) at the 97th percentile, known as the Value-at-Risk (VaR$_{97\%}$). This threshold represents the point below which 97% of the loss outcomes lie, isolating the upper 3rd percentile of the loss distribution.

The next step involves calculating the mean of the losses within this upper 3rd percentile. This is done by averaging all the loss values that exceed VaR$_{97\%}$. The value at the 97th percentile itself is then subtracted from this average to determine the expected excess loss for the reinsurer. This process ensures that the premium calculation is aligned with the actual risk exposure borne by the reinsurer.

$$\Pi^{reinsurer}_{fair} = \frac{\sum_{i=\hat{N}}^{N} S_i - VaR_{97\%}}{N} \tag{2}$$

where the $VaR_{97\%}$ is taken as the mean value of the 97th bin of cumulative S distribution. This process produces a premium price $\Pi^{reinsured}_{fair} = 75.13\,\$$.

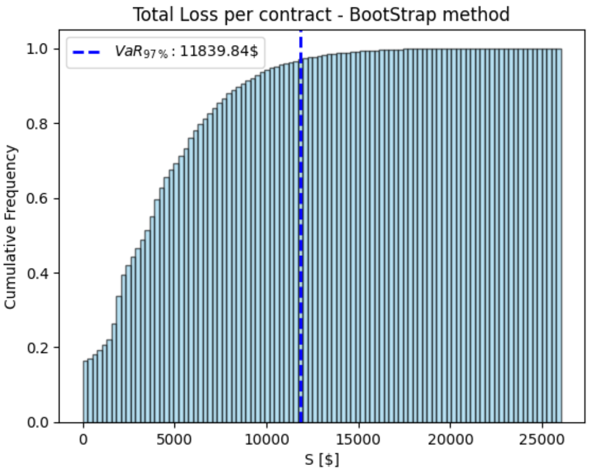

Figure 14: The distribution of losses highlighting the Value-at-Risk (VaR$_{97\%}$), which separates the top 3rd percentile of losses from the remaining outcomes. This percentile serves as the threshold for determining reinsurer payouts.

The reinsurer's payout structure is shown in Figure 15. It covers only those losses that exceed the $VaR_{97\%}$, effectively transferring the extreme tail risk from the primary insurer to the reinsurer. This payout structure forms the basis of the reinsurance agreement.

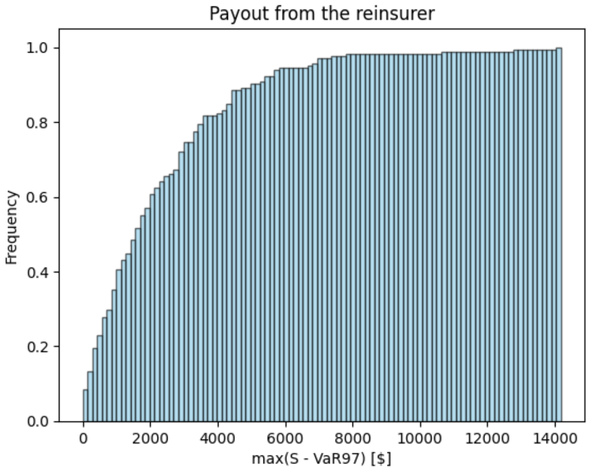

Figure 15: The payout structure for the reinsurer. The reinsurer compensates for losses above the 97th percentile threshold, ensuring coverage for extreme outcomes.

The Pure Premium price for the reinsurance policy is calculated using the principle of expected value, which accounts for both the reinsurer's expected payouts and an additional loading factor to ensure profitability. The formula for the premium is given by:

$$\Pi^{reinsured}_{pure} = (1 + k) \cdot \Pi^{reinsurer}_{fair}$$

where k is the loading factor, set to $k = 0.25$ in this case, to cover administrative costs, risk margins, and profits.

By substituting the values, the reinsurance premium is calculated as:

$$\Pi^{reinsured}_{pure} = 93.92\,\$.$$

The final step involves evaluating the total losses under the reinsurance contract. These losses include the cost of

[3] The estimated error has been calculated by using the Standard Error $= \sqrt{\frac{P_{ruin} \cdot (1 - P_{ruin})}{n_{trials}}}$

the reinsurance premium, which is paid for every contract, and the capped losses covered by the reinsurer. The total losses are plotted to provide a comprehensive view of the financial impact of the reinsurance agreement.

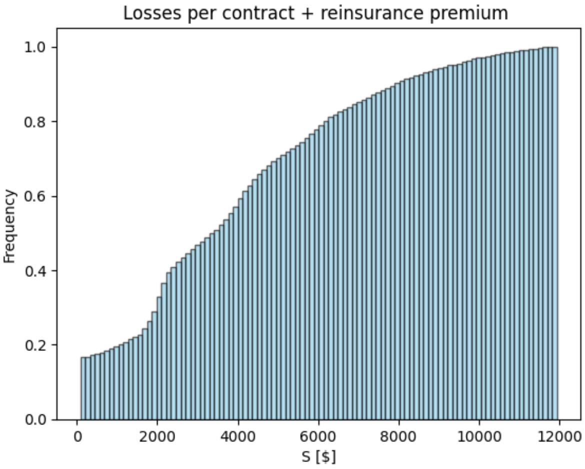

Figure 16: Total losses for the insurer company under the reinsurance agreement, incorporating the reinsurance premium and the capped payout provided by the reinsurer beyond the 97th percentile threshold. This plot illustrates the reduced variability in losses achieved through the reinsurance coverage.

This approach ensures that the reinsurance premium is both fair and reflective of the risk transferred, while also providing financial stability to the primary insurer by limiting exposure to extreme losses.

Results and considerations

A convolution PDF, $LogNormal \oplus Gamma$ functions gave proof to represent the dataset, verified through three different statistical tests. Such PDF model parameters have been used to generate MC loss values, for a portfolio with $10'000$ policies. Where each contract could have a number N of claims (loss values), distributed accordingly to a Negative Binomial. These characteristics defined the insurance company setup for the MC case study.
The Fair premium price in this setup was found to be $\Pi^{MC}_{fair} = 4005.70\,\$$. And the Probability for the company to experience ruin has been determined to be $P^{MC}_{ruin}(\Pi^{MC}_{fair}) = (51.10 \pm 1.58)\%$

The BootStrap (BS) method has also been tested. This method does not make assumption about the distribution of the dataset, but instead use directly the empirical data to generate values, providing results aligned to the MC ones. This case was studied with two size generated samples, $10'000$ and $5'000$ policies.
This method provided a Fair Premium price $\Pi^{BS}_{fair} = 4008.39\,\$$ for the $10'000$ policies case. While, $\Pi^{BS}_{fair} = 4014.28\,\$$ for the $5'000$ policies case.
For the latter case the probability for the company to experience ruin has been determined to be $P^{BS}_{ruin}(\Pi^{BS}_{fair}) = (51.40 \pm 1.58)\%$.

In both cases, MC and $BS^{5'000}$ the probabilities to experience ruin are compatible and both are $\approx 50\%$, which indicates that the Fair Premium in half of the cases is enough to cover all the claims. Indeed, the majority of the cases in which the company experienced ruin, during the simulation, the total premium price cached was close to cover the entire portfolio claims. Which indicates that *the Fair Premium price represents statistically the threshold value that equals the expected total loss ($E[S] \equiv \Pi_{fair}$) for a single contract.*

The case of reinsurance has also been analyzed. In this scenario, we calculated the reinsurance cost based on the expected payouts in the upper tail of the loss distribution $\Pi^{reinsured}_{fair} = 75.13\,\$$ and then it has been loaded with the expected value principle. Then the cost $\Pi^{reinsured}_{pure} = 93.92\,\$$, is applied across all contracts. By incorporating this reinsurance cost , the plot of the losses is shifted to the right, reflecting the additional premium expense. Furthermore, the losses are capped at the maximum threshold stipulated by the reinsurance agreement, effectively limiting the insurer's exposure to extreme outcomes while transferring the excess risk to the reinsurer.