Simulation in Finance and Insurance 2022

Case study – Report

21.05.2022

Colin Carruzzo Arnaud Castellana Sébastien Traimond

1. Executive Summary:

The calculations performed in our analysis have shown that we cannot reduce the premium amounts, since the expected value of our claims is greater than the total premium income.

The premiums, as they are currently priced by the insurance company amount to a total cash inflow of 198’211 monetary units, whereas the expected cash outflow for our company, which has been computed using our model, amount to 1’544’364 monetary units. This then leads to a deficit for our insurance company of -1’346’153.32 monetary units, as you can see in Figure 1 right below:

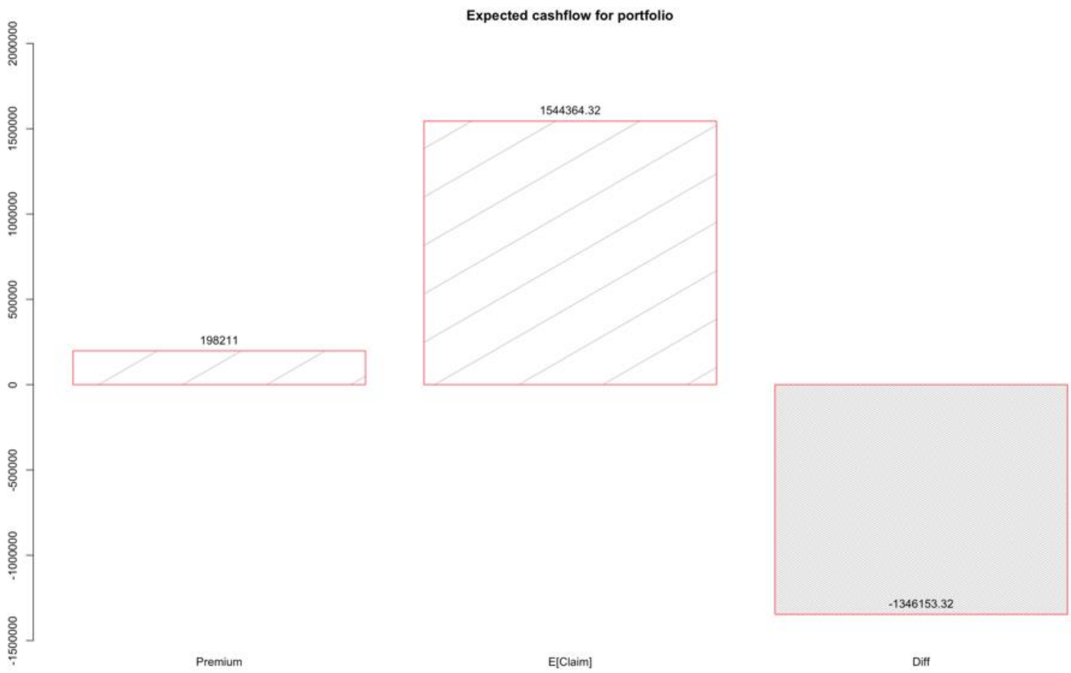


Figure 1

Furthermore, the minimum loss we expect to face annually amounts to 1’439’307 monetary units.

As it currently is the case, the average premium of 198.211 monetary units is already way below the expected loss of a policyholder, which is of 1’555.444 monetary units.

We understand that by lowering the premium tariffs and thus attracting more clients it would then lead to a bigger cash inflow. However, it would not comply with the risk that could be borne by our insurance company, since every new customer we would contract with this premium tariff would lead to an even bigger deficit for our company.

To summarize our point of view, we unfortunately cannot endorse to lower the premiums. The actuarial point of view would even emphasize the need to increase the premium tariffs.

2. Data Exploration, Preparation and Cleaning:

- Data Cleaning:

We excluded the rows concerning the policyholders with id numbers 5 and 6, since they included values for the claim amounts which did not comply with the rest of the claim amounts.
For the policyholder with id number 5, we decided to exclude his data since he had a claim amount of -3000 monetary units.
For the policyholder with id number 6, we decided to exclude his data since he had a claim amount of 200000 monetary units, which is significantly above the other claim amounts of our data set and thus inconsistent with our data set.

The information concerning these two policyholders were removed because they were judged as being mistakes.
Furthermore, we also verified that no NA-values were included in our dataset, but this was not the case in our data set.

- Data Preparation:

In order to facilitate our coding experience, we also decided to save the data contained in the columns of the frequency and the claim amounts in two new shorter variables, named CLM_FREQ and CLM_AMT.

- <u>Data Exploration:</u>

```
       id            CLM_FREQ        CLM_AMT        CAR_USE            CAR_TYPE              AGE
 Min.   :   1.0   Min.   :0.000   Min.   :   0   Length:1000        Length:1000        Min.   :19.00
 1st Qu.: 252.8   1st Qu.:1.000   1st Qu.:1837   Class :character   Class :character   1st Qu.:39.00
 Median : 502.5   Median :1.000   Median :1962   Mode  :character   Mode  :character   Median :44.00
 Mean   : 502.5   Mean   :1.572   Mean   :1549                                         Mean   :44.56
 3rd Qu.: 752.2   3rd Qu.:2.000   3rd Qu.:2036                                         3rd Qu.:51.00
 Max.   :1002.0   Max.   :8.000   Max.   :2283                                         Max.   :80.00
    GENDER              AREA              PREMIUM
 Length:1000        Length:1000        Min.   :100.0
 Class :character   Class :character   1st Qu.:147.0
 Mode  :character   Mode  :character   Median :195.0
                                       Mean   :198.2
                                       3rd Qu.:249.0
                                       Max.   :300.0
```

Firstly, let us have a quick look at the summary of our cleaned data set, which gives us a first idea of the concerned data set.

We will now go into more details concerning the claim frequency and claim amounts.
Concerning the claim frequencies, we notice that the mean is equal to 1.572 with a minimum value of 0 and a maximum value of 8.
The variance is of 1.756573 and the standard deviation is of 1.325358.
Figure 2 gives a visual representation of the different claim frequencies.

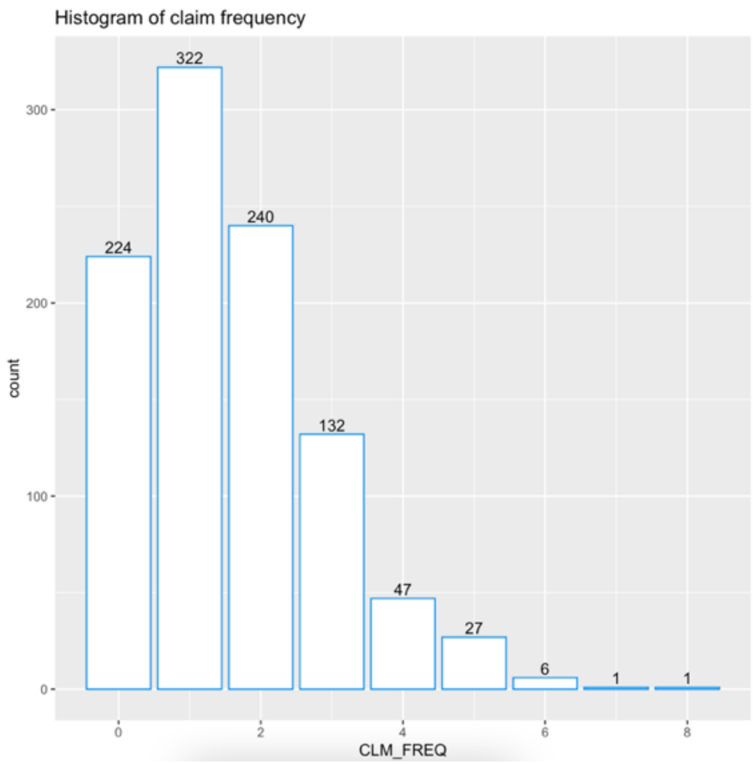


Figure 2

Let us now have a look at the data concerning the claim amounts. For the claim amount we need to exclude the zero entries in order to have a relevant data set. That being done, the mean claim amount is of 1995.737 monetary units. The minimum value amounts to 1726.876 monetary units and the maximum to 2283.425 monetary units.

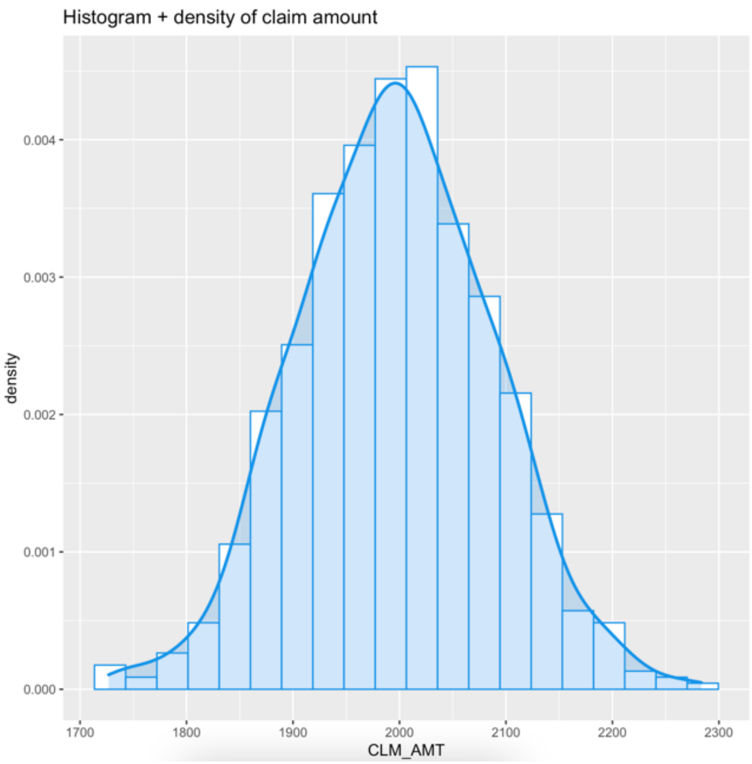


Figure 3

Figure 3 gives us a visual representation of the claim amounts, with the histogram and the density of the sample.

Figure 4 is a boxplot which quantifies the median, the first (25%) and the third (75%) quantile. The 25%-quantile is equal to 1933, the median equal to 1995 and finally the 75%-quantile is equal to 2056. This means that 50% of our values are contained in the interval [1933,2056]. This information already gives us a clear idea that the variance of the claim amounts will not be very big. Indeed, it amounts to 8125.862 and the standard deviation to 90.14356.

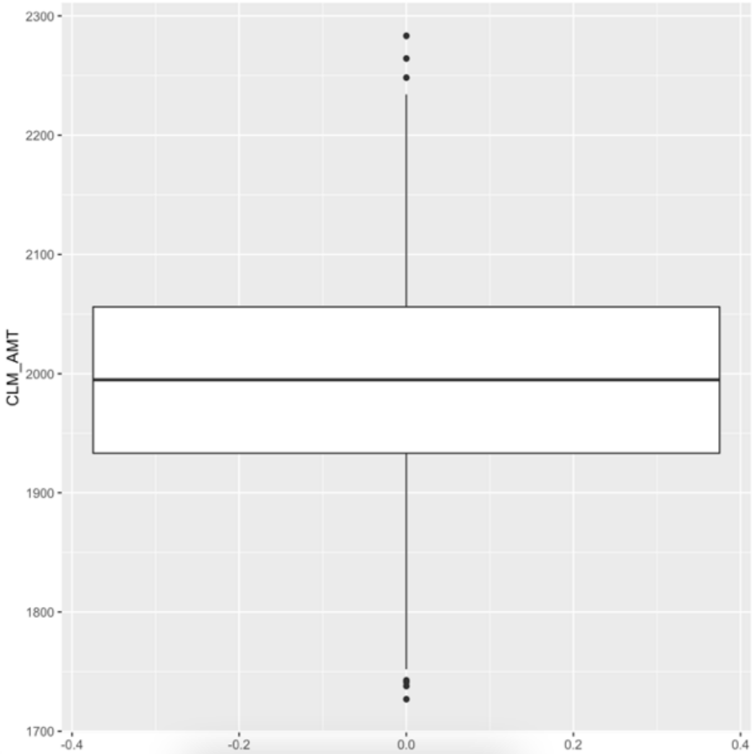


Figure 4

3. Model Selection and validation

- Frequency:

First thing to note is that we have to apply the Chi-square Goodness-of-Fit test for the frequency of the claims since the frequency of the claims is a discrete distribution. This statistical test is used to compare the sample with a specified distribution in order to determine whether our data set is likely to come from this distribution or not. If the resulting p-value of the test is greater than the significance level alpha of 0.05, we can conclude that the observed data is not significantly different from our specified distribution and we can therefore use this distribution for our model.
We know that the most adequate distributions that can be used to model the claim frequency are the poisson distribution, the binomial distribution, and the negative binomial distribution.
We first performed the Chi-square Goodness-of-Fit test on the poisson distribution which resulted in a p-value of 0.09901, which means that our data can be well described via a poisson distribution since the p-value of 0.09901 is greater than the significance level of 0.05. We continued to perform the Chi-square Goodness-of-Fit test on the other mentioned discrete distributions.

For the negative binomial distribution, we achieve a p-value of 0.7055, which means that the data is also well described by a negative binomial distribution.
A p-value of 1 means that the data is perfectly described by the chosen model. Since the p-value of the negative binomial distribution is closer to 1, we confirm the negative binomial distribution as our chosen model.
In Figure 5 we have a visual representation, which shows the good match of the negative binomial distribution with the claim frequency of the data.

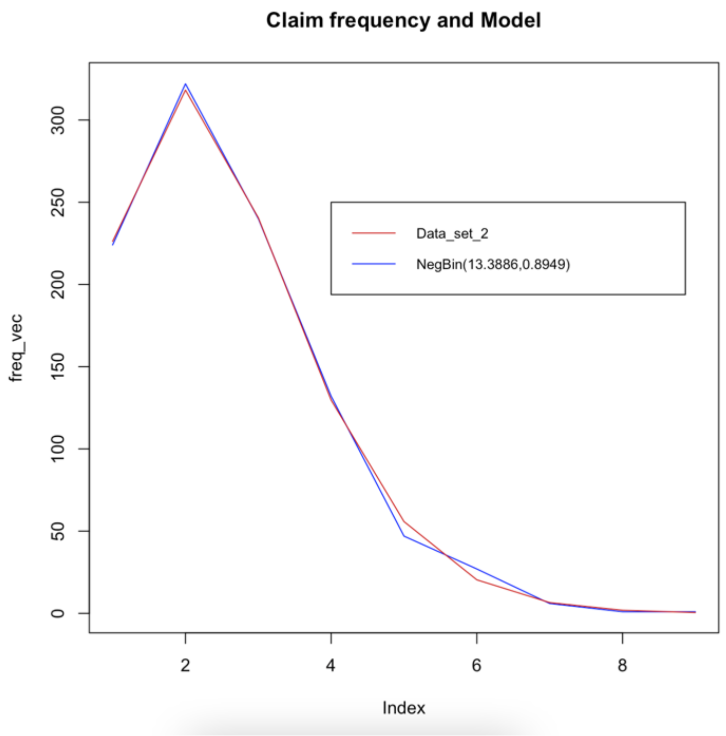


Figure 5

- <u>Severity:</u>

Since the severity of the claim sizes are continuously distributed, we can perform the Kolmogorov-Smirnov Goodness-of-Fit test, which also is a statistical test with the same purpose than the Chi-square Goodness-of-Fit test with the only difference that the Kolmogorov-Smirnov Goodness-of-Fit test is used for continuous distributions.
This statistical test determines the distance between the cumulative distribution function of the empirical distribution, which is being described by our data, and the specified distribution we want to compare our sample to.
Performing the Kolmogorov-Smirnov Goodness-of-Fit test on the log-normal distribution, the test returns a p-value of 0.9888. The Kolmogorov-Smirnov Goodness-of-Fit test for the normal distribution is 0.9987. Performing the Kolmogorov-Smirnov Goodness-of-Fit test with the gamma distribution, we achieve a p-value of 0.9992, which is a good result since it is above the significance level of 5% and very close to 1.

All the above-mentioned distributions would be a really good choice to model the claim amounts, but once again we chose to select the one yielding the closest value to 1, which in our case corresponds to the gamma distribution.

Figure 6 confirms that a gamma distribution does well explain the data, with a QQ-plot showing a straight 45 degrees line.

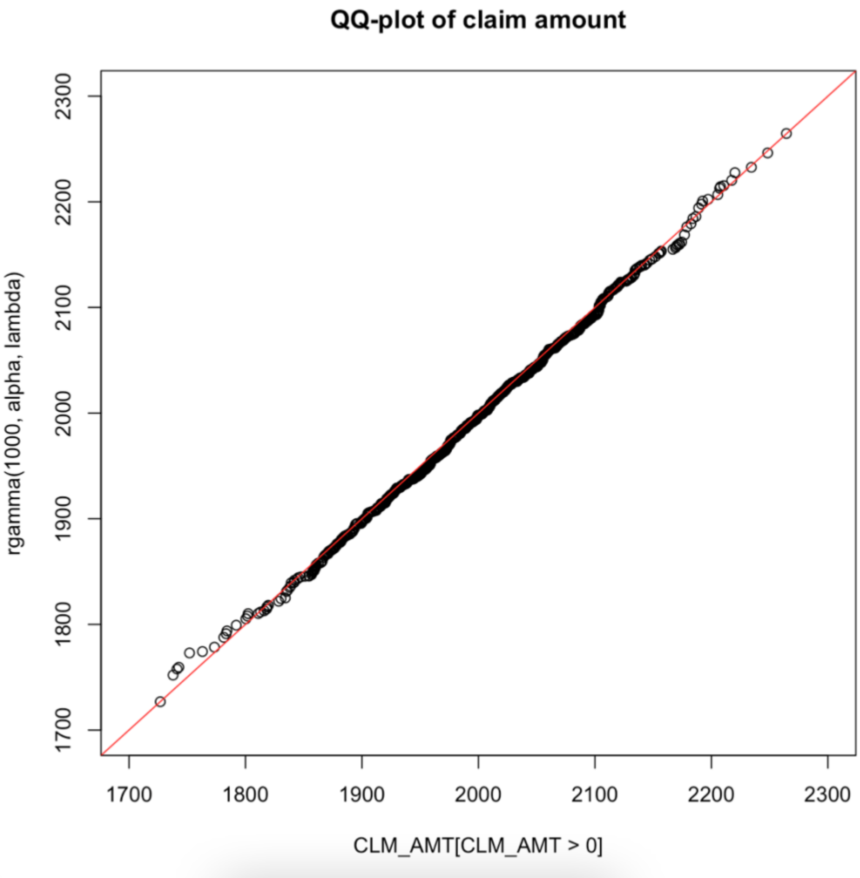


Figure 6

Figure 7 gives us a good visual representation of the good match of the gamma distribution with the claim amounts of the data.

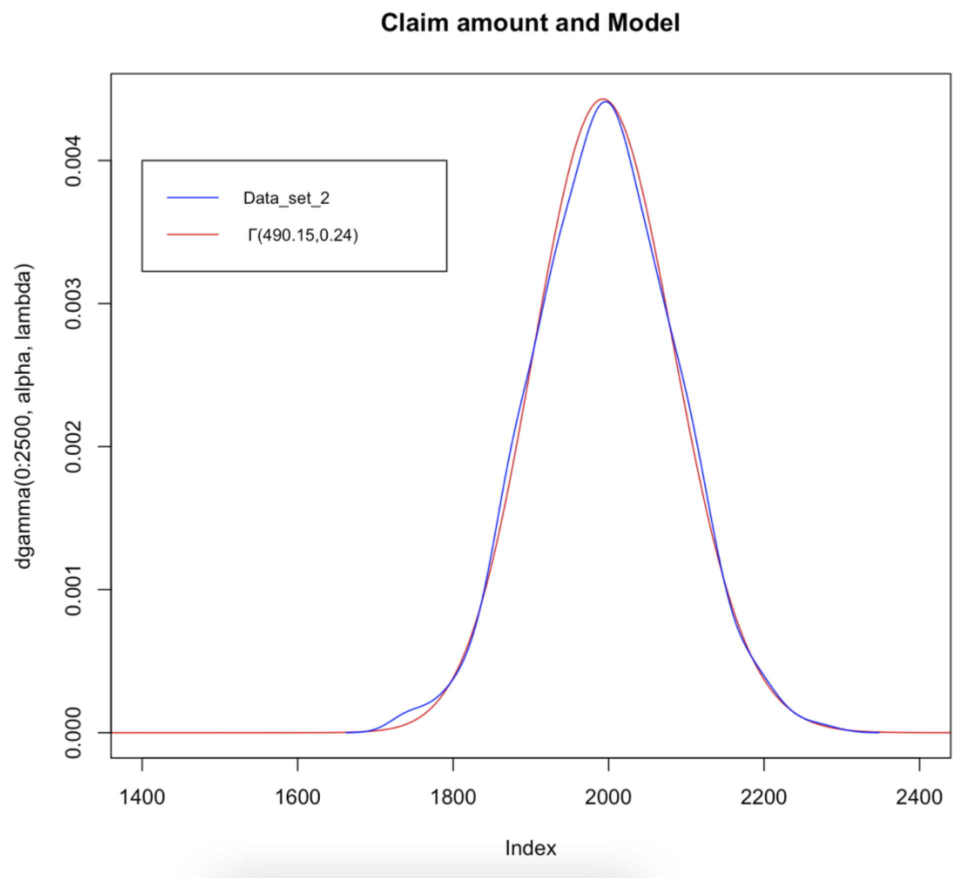


Figure 7

4. Application

The average of the premium amounts our customers pay is of 198.211 monetary units.

The risk premium based on the data provided amounts to 1'548.692 monetary units.

We will now explain how we computed the risk premium based on our model:
With this specific dataset, we notice that the aggregate claim size is modelled by the gamma distribution, and this is independent of the number of claims, i.e. the claim frequency. This means that if we consider only one policyholder, what is of interest for us in this specific dataset is only the claim size and the probability of having zero claims (and obviously the probability of having at least one claim.)
We know that the aggregate claim size per policyholder is well-described by a gamma distribution.
We can use a Bernoulli random variable to describe the claim frequency, with the parameter p equal 1-P(0 Claims) = 1- NegBin(0,r,p).

Let us now simulate our expected loss per policyholder with the Monte Carlo Method. We find a result of 1'555.44, which is quite close to the risk premium based on the data.

The difference between the premium based on our model and the effective average premium is huge: we end up with a cash outflow of 1'357.229 monetary units per policyholder.

While looking at the overall portfolio instead of only one policyholder, we can simulate from the binomial distribution, with n = 1000 and p = 1 – P(having 0 claims).
For the claim amounts we can then keep our model of the gamma distribution. Using this new model, we end up with an aggregate claim size for the whole portfolio of 1'544'364 monetary units, which is way bigger than the overall premium income of 198.211 monetary units. You can refer to Figure 1 for the visual representation.

Concerning Solvency 2, we find that the company should possess a minimum capital of 1'611'317 monetary units, in order to be bankrupt only one year over two hundreds. This represents the value at risk, with alpha equal to 99.5 %.

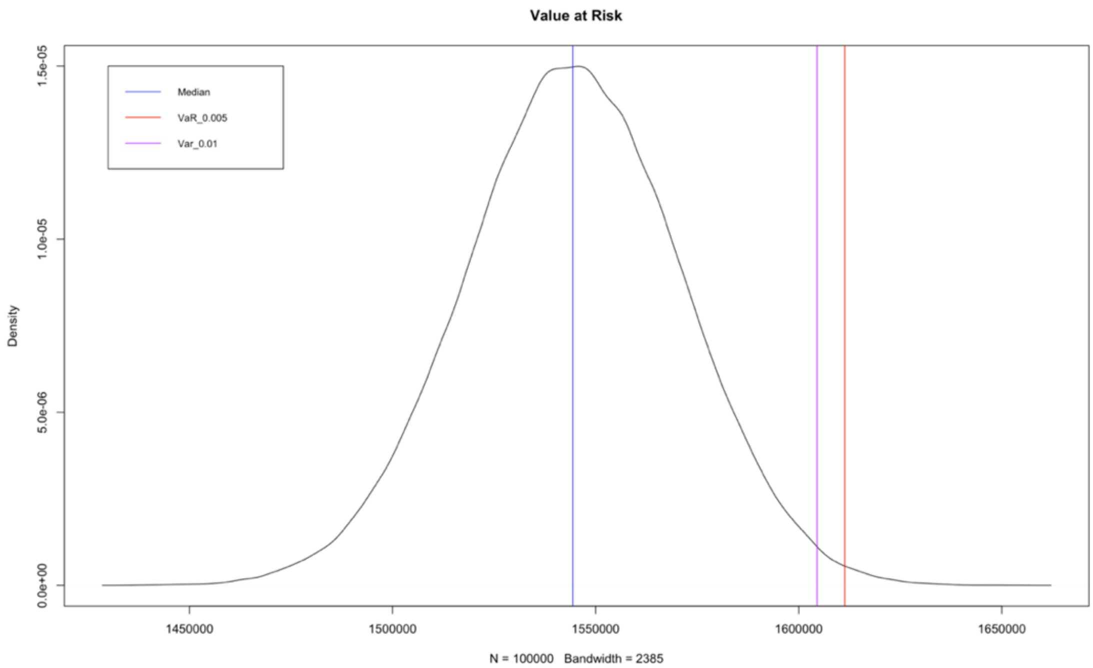
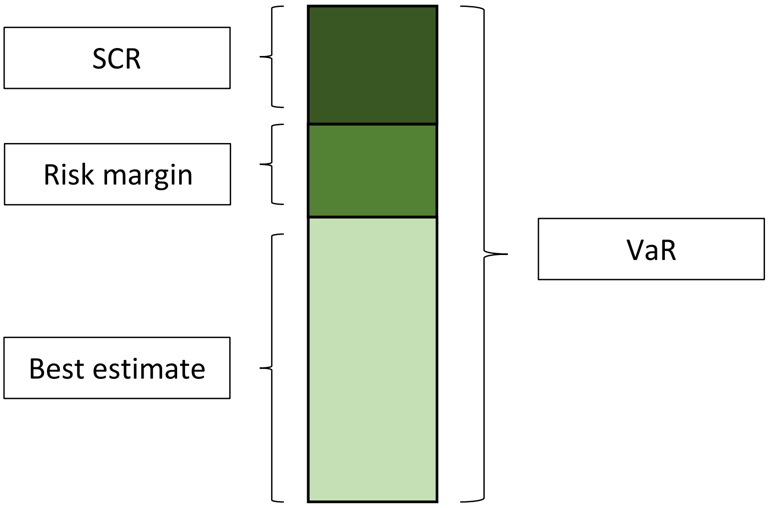


Figure 8

You can refer to figure 8, with the median in blue, the quantile of 99% in purple and the quantile of 99.5% in red.

In order to raise the necessary capital, the insurance company should increase the tariff on the premiums of the portfolio. This part will consist in the best estimate and the risk margin. The Solvency Capital Requirement could be raised via equity.



Another option would be to keep the tariff as they currently are, but instead increase the Solvency Capital Requirement by raising more equity.

In any case, the company is obliged by the regulator to raise the necessary capital.

5. **Model explanation:**

In order to compute our results described in part 1, we needed to be able to predict how much a policyholder would cost our insurance company per year.

To do so, we looked at our data and concluded that, in this particular motor insurance portfolio, the number of claims a policyholder declares, has no impact on his overall annual cost, except if he has zero claims. In other words, to be able to predict the costs of a policyholder, it only matters to know if a policyholder declared any number of claims or not

We can therefore separate our policyholder in two distinct categories, on one hand the category of policyholders with any number of claims and on the other hand the category of policyholders without any claim. Based on the data, we found out that approximately 22% of the policyholders in this portfolio will not declare any claim.

Obviously, if a policyholder declares no claims, he will not cost us anything. Once again, based on the data, we were able to model the cost for the policyholders that would declare any claim.

After having done the calculations, we ended up with our conclusions mentioned in part 1.