MT7027: Project 1

Anton Stråhle, Jan Alexandersson & Max Sjödin

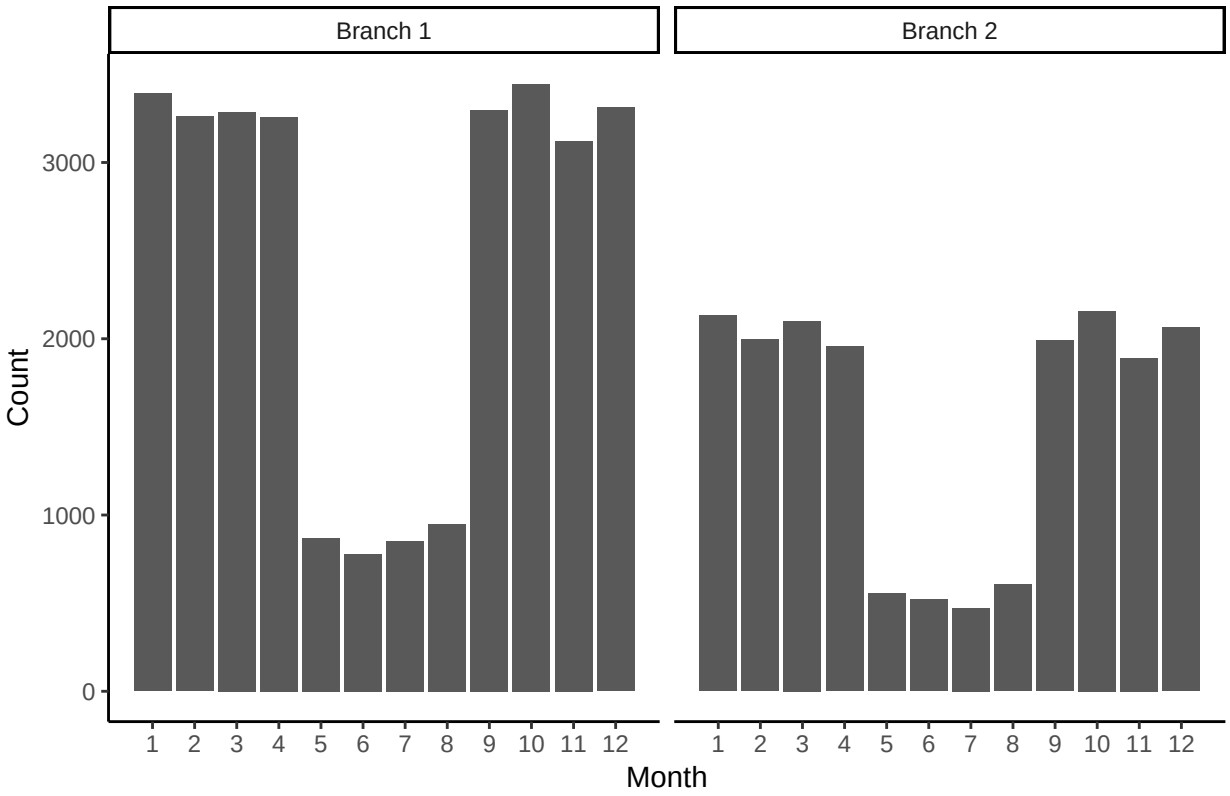
Introduction

In this project we are dealing with data concerning two different insurance branches collected over 10 years. We do not know anything about the sizes of the two insurance portfolios except the fact that their size has not changed over the decade which the data spans. Furthermore, the insurance products are of the non-life type and are paid out in lump payments. It is important to mention that we use Jan Alexanderssons data in this project. In this project we will examine both the claims and cost distributions of these non-life products and how different covers, XL and SL, can be applied to change the total annual cost distributions for the insurance products, and for the insurance provider as a whole.

Exercise 1

In this exercise we want to find trends in the data for the two insurance branches in order to model future claims in a block-wise manner. Since the data is structured in a way such that we only have the claim day (i.e. $1, 2, ..., 3650$) we assume that 365 day/year and that a month is $365/12$ days (to get 12 months).

Figure 1: Number of claims for branch 1 and 2



From Figure 1 it seems reasonable to divide the months into two homogeneous groups with their own claims distributions. One group for months $(1-4, 9-12)$ and one for months $(5-8)$. We also wish to examine if we have homogeneity between the different years during which the claim data was collected.

Figure 2: Number of claims by year for branch 1

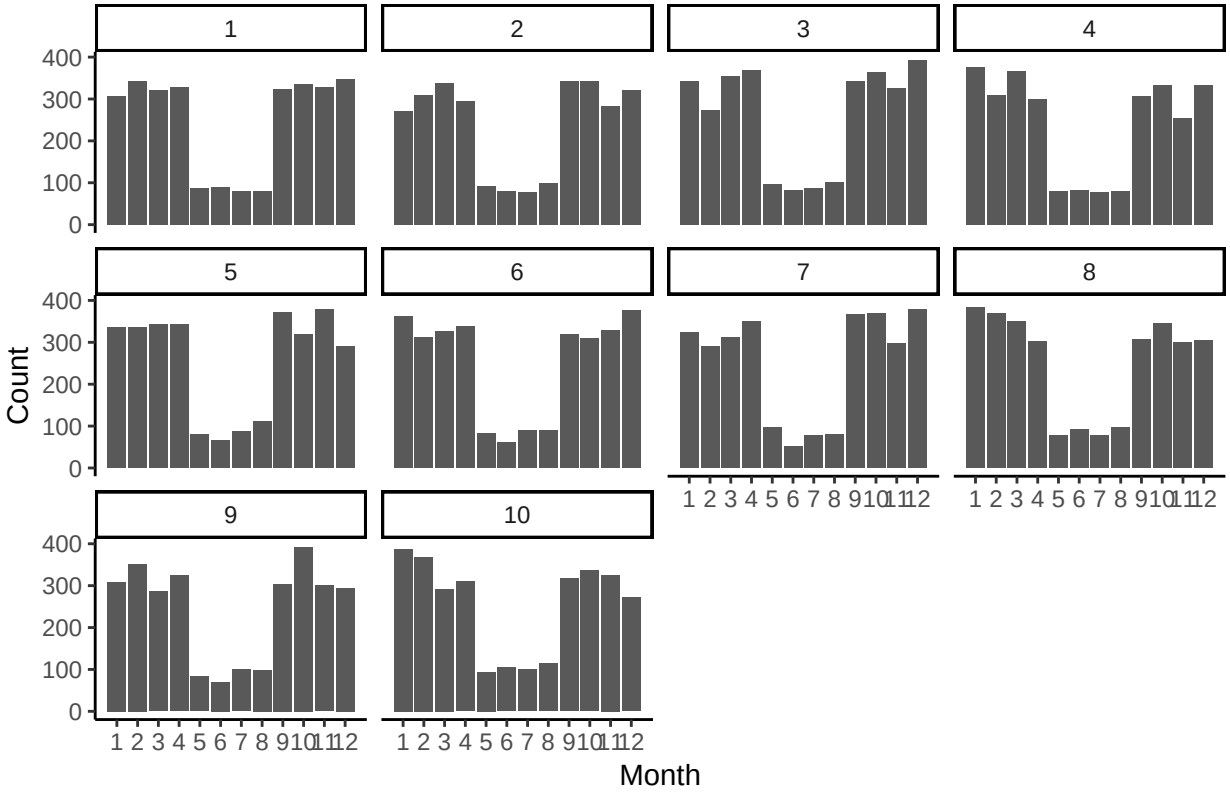
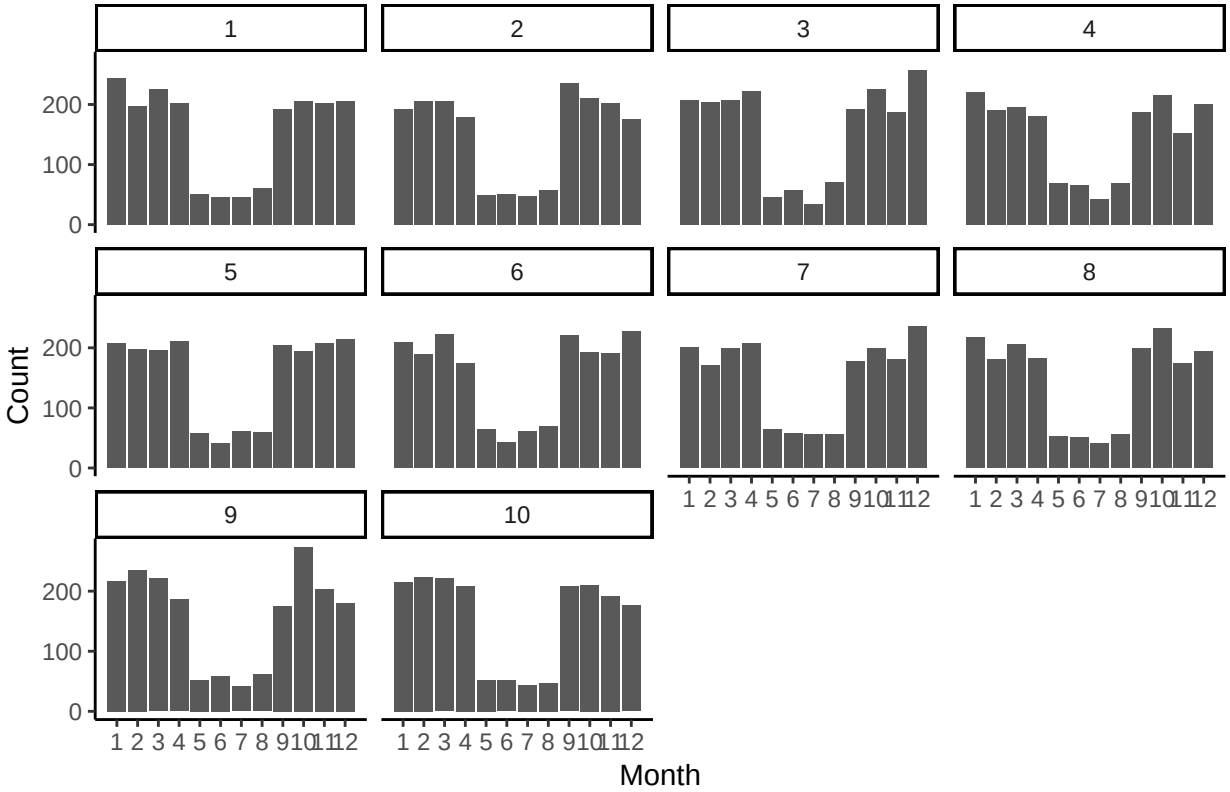


Figure 3: Number of claims by year for branch 2



We note from Figure 2 and 3 that there does not seem to be any major difference in the number of claims between the years. We can therefore model the data separately for low intensity periods and high intensity periods for each branch.

We now wish to fit homogeneous Poisson distributions N_{ij} (where i represents the insurance type and j represents the season) to the periods and each insurance product.

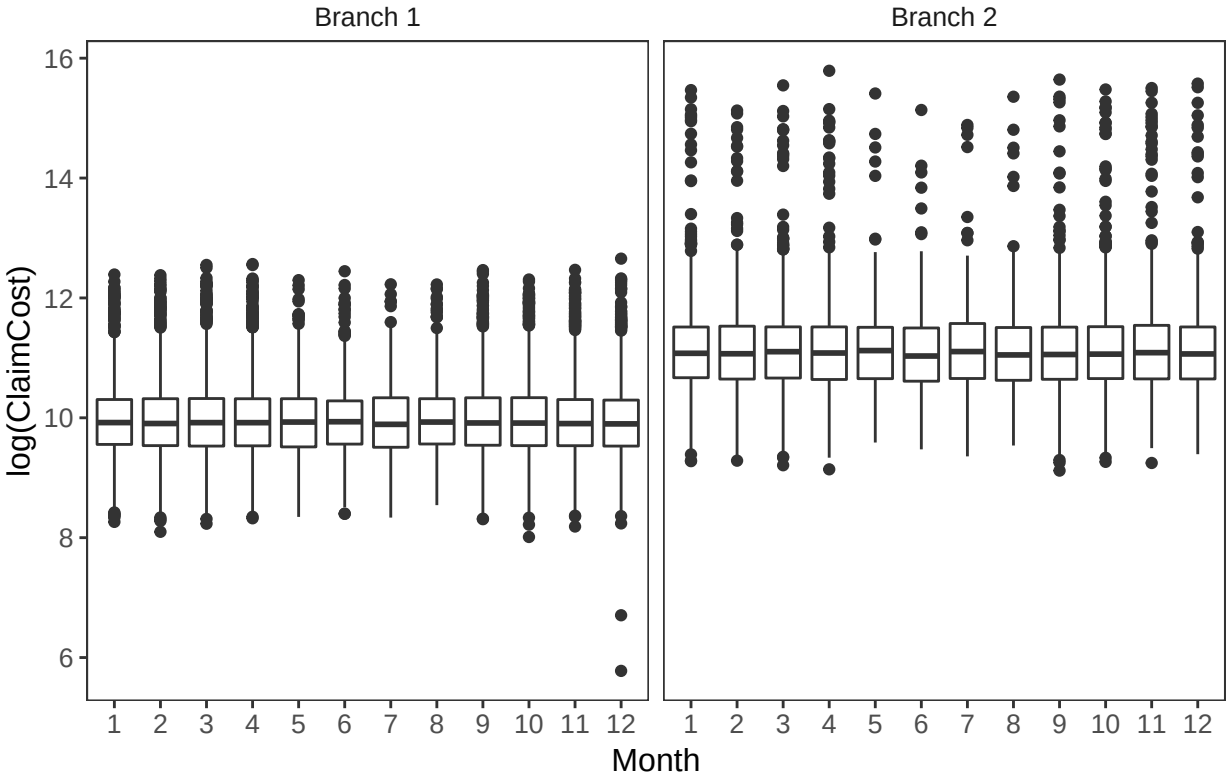
$$N_{ij} \sim \text{Po}(\lambda_{ij})$$

The Poisson variables which have been fit for the different seasons and insurance branches and have had their parameter λ_{ij} estimated through maximum-likelihood methods using the data from the 10 previous years. However, we note from the plots below that the Poisson distribution does not fit the data very well as can be seen in the figures in the Appendix. We seem to have overdispersion for some periods and branches, meaning that the variance is not truly equal to the expectation which is the case for a Poisson variable. For others combinations of seasons and branches the data does however seem to indicate that we have underdispersion, meaning that the variance is lower than the expectation. Therefore we also fit Negative Binomial distributions to N_{ij} as this distribution does not have the property of equal expectation and variance as the Poisson distribution does. We can clearly see from the figures in the Appendix that the Negative Binomial distribution models our data better than the Poisson distribution. We can also see this by examining QQ-plots between the empirical distribution of our data compared to the Poisson distribution and Negative Binomial distribution.

Exercise 2

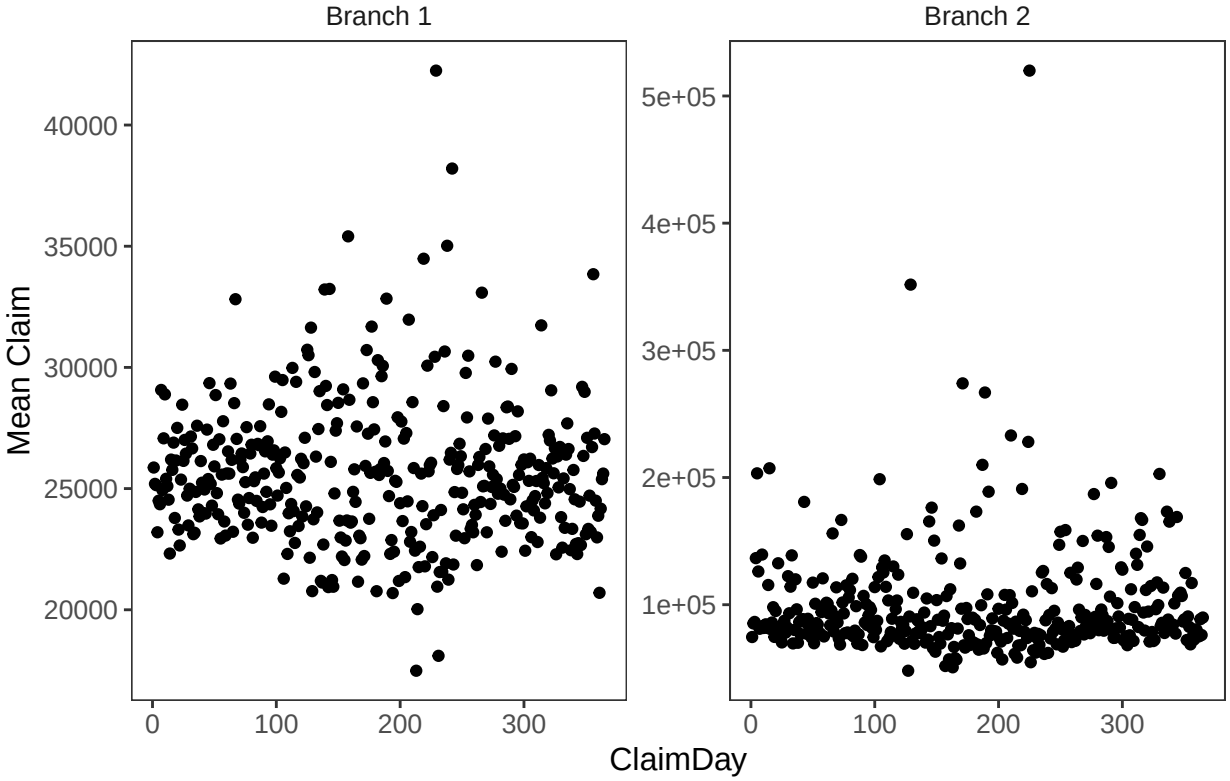
In this exercise we analyze the claim costs over time for each of the two insurance branches. We are looking for possible patterns in the claim costs over time by analyzing the distribution of the claim costs. In the following figures we can observe the average logarithm of the claim cost for each month and insurance branch. The log-transform is used to make the figure more interpretable.

Figure 4: Boxplot of log of claim cost vs month



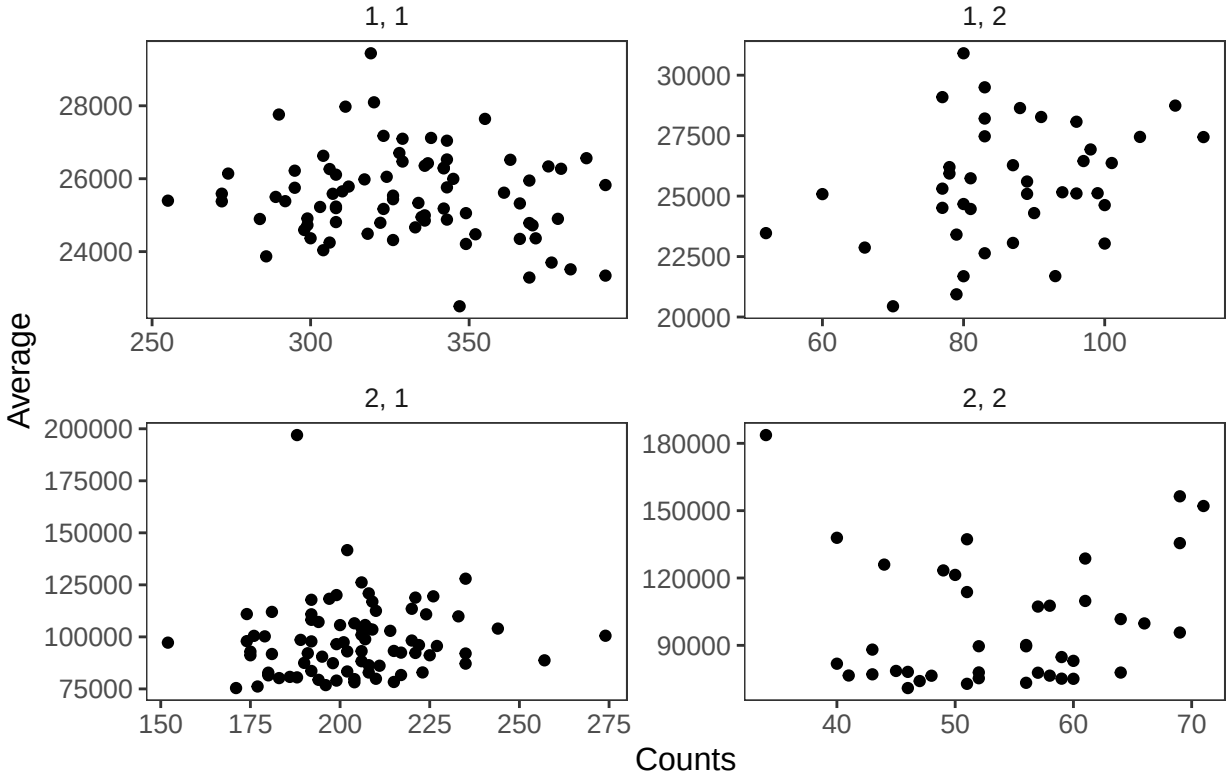
From Figure 4 it seems as if the average claim costs are time independent, at least when aggregated on a monthly basis, from the two previous plots. The same can also be said if aggregated on a yearly basis. We can also observe the mean of the claim sizes for every day in order to further strengthen the assumption of time independence.

Figure 5: Average claim cost for each day



We note from Figure 5 that the average claim sizes each day seem to be time independent. An interesting factor is however that all extremely deviating claim costs seem to occur during the summer (i.e period 2), which may be a result of a lower number of claims during this period. In conclusion it seems as if the claim costs are independent of time.

Figure 6: Insurance Product, Season



From Figure 6 we cannot see any systematic correlation between claims and claim costs for any of the combinations, meaning that we can model the number of claims and claim costs independently. We do not observe any time dependencies in terms of average claim costs and there seem to be no systematic correlation between claims and claim costs. In conclusion, the lack of time dependencies for the claim costs means that we can model the claim costs with the assumption that the claim costs are time independent.

We also want to investigate the distribution of the claim costs.

Figure 7: Zoomed histogram of claim cost for branch 1

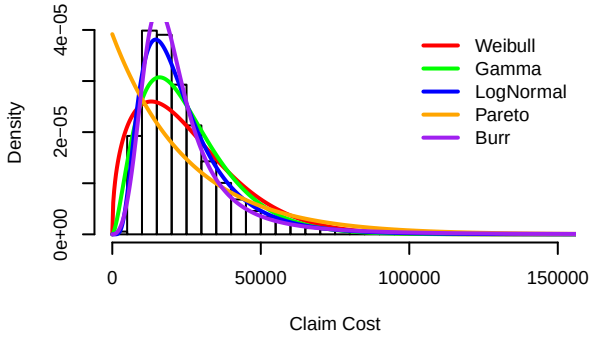


Figure 8: Empirical and theoretical CDFs for branch 1

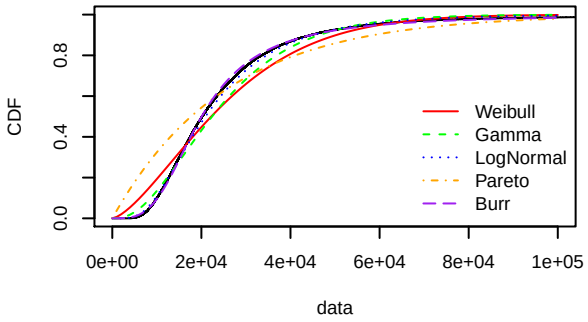


Figure 9: Q−Q plot for branch 1

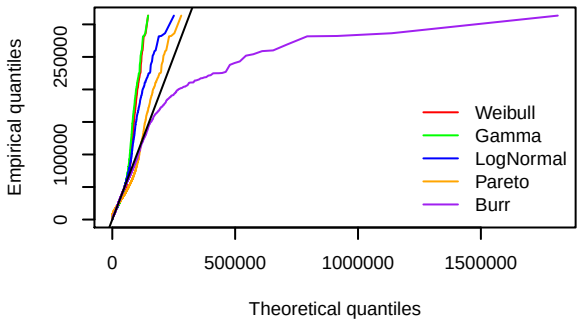


Figure 10: Zoomed histogram of claim cost for branch 2

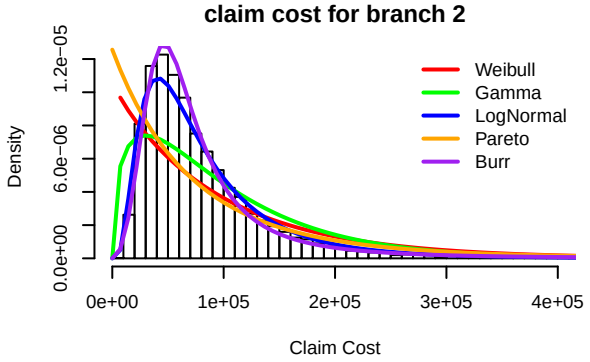


Figure 11: Empirical and theoretical CDFs for branch 2

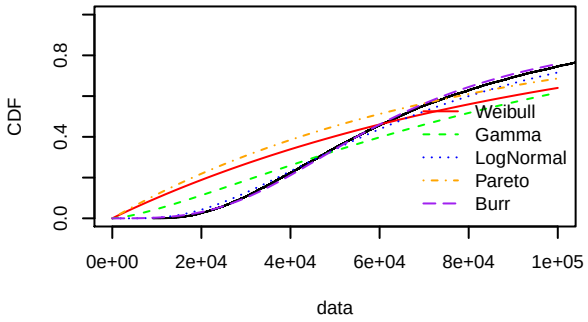
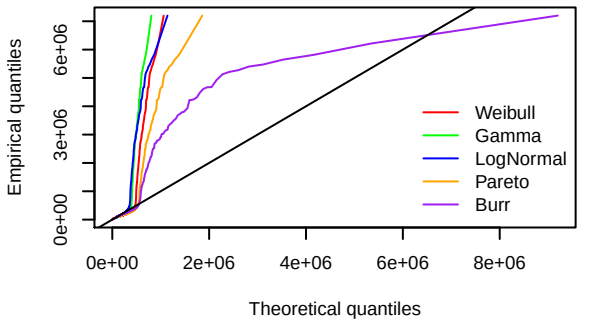


Figure 12: Q−Q plot for branch 2



We can see from the histograms in Figure 7 and 10 that a fitted Burr distribution seems to model the data well, however we can see from the QQ-plots in Figure 9 and 12 that we run into problems in the tails.

Therefore we tried to fit different distributions for the tail, compared to the rest of the data. That is, we fit a distribution to claim cost below a chosen number and another distribution for cost above that number. However, all our attempts led to a bad fit. Since we could not achieve a good fit using common distributions we fit the data using splines, using the R package `logspline`, which gives a very good fit, however one has to be careful of overfitting when using splines.

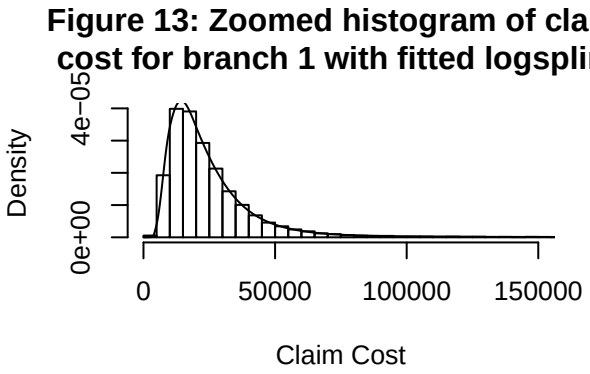


Figure 13: Zoomed histogram of claim cost for branch 1 with fitted logspline

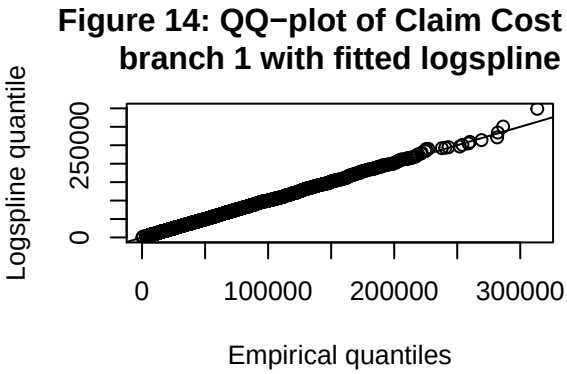


Figure 14: QQ−plot of Claim Cost for branch 1 with fitted logspline

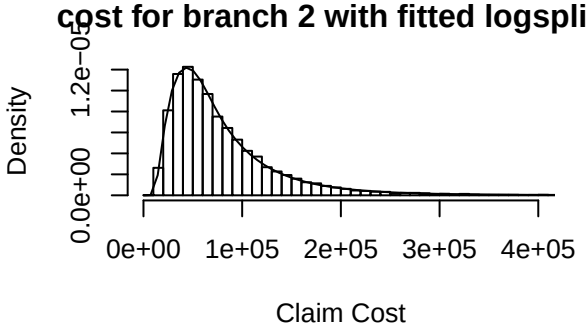


Figure 15: Zoomed histogram of claim cost for branch 2 with fitted logspline

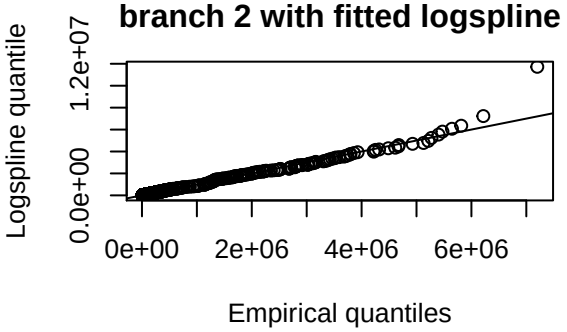
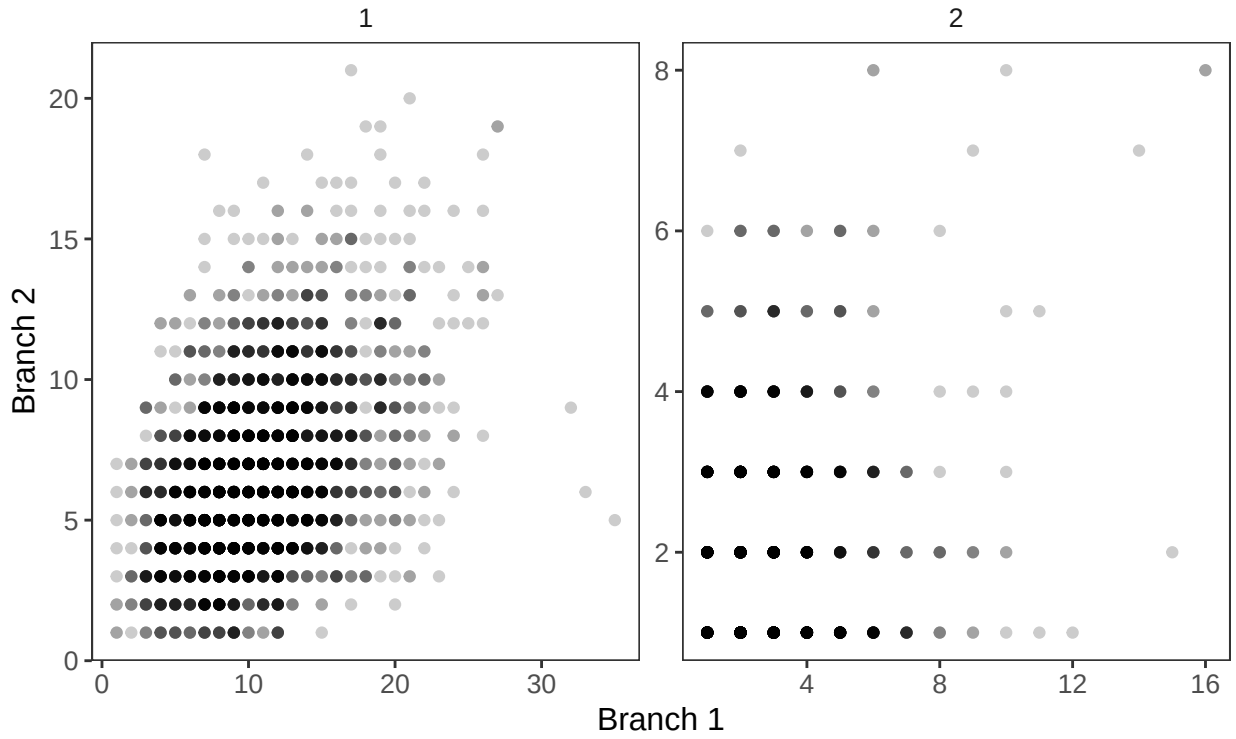


Figure 16: QQ−plot of Claim Cost for branch 2 with fitted logspline

Exercise 3

In order to accurately assess the total claim cost for the entire company it is of importance to examine whether or not we have dependence between the two insurance branches. We look at the total number of claims for each branch on a daily basis and plot the number of claims against each other, for each season.

Figure 17: Daily number of claims for each branch plotted against each other for seasons 1 and 2



We note from Figure 17 that there seems to be some kind of positive correlation between the two insurance products, meaning that we cannot model the claims for each product separately, but that we instead have to model them jointly. We can however still model the claim costs of the two branches independently as we saw in the previous exercise.

Table 1: Correlation between the branches for total cost and total number of claims per month under the 10 year period.

Month	Cost Correlation	Claims Correlation
1	0.1316952	0.1900403
2	0.4018696	0.3696001
3	-0.2385072	-0.6387608
4	0.2670056	0.6790218
5	-0.4364984	-0.3144967
6	0.3886900	0.0925294
7	0.5359388	-0.0961634
8	-0.3130146	-0.4049584
9	0.1585718	0.0867293
10	0.3137511	0.7734773
11	0.2673423	0.6607376
12	0.5033946	0.8279092

From Table 1 we can see that there seems to exist some dependency between the branches considering the total number of claims for a certain month under this 10 year period. We can see the same for the total claims cost both not as strong as for the number of claims. As a consequence we should not use the marginal

distributions to simulate future claim costs and instead use a joint distribution.

Exercise 4

As we mentioned previously the number of claims for the two branches are not independent, meaning that the have to be sampled from a bivariate distribution rather than two univariate distributions. This can be done by either deriving the bivariate distribution analytically and then sampling from it or by generating a bivariate sample by bootstrapping or using Gauss-copula method.

For the bootstrap method we begin by sampling the number of claims for each month of the following year from the data of the correct season. We previously mentioned that the claim costs are time independent for both branches as well as that they are independent between the two insurance branches. Due to this we can, as previously mentioned, model claims and the individual claim costs independently. An advantage of using bootstrap is that we do not need to know the distributions for the number of claims or the claim costs, which we do not know for the claim cost since we could not find a good fit for any of the distributions we tried.

We can also use the Gauss-copula-method to simulate a 2-dimensional number of claims for each day, one for each branch. The Gauss-copula-method keeps the dependency structure, which we observed, between the two branches with regards to the number of claims. We simulate separately for each season since we observed that the day distributions are different in these periods. More specifically we follow these steps to simulate number of claims for one month:

1. Simulate i.i.d pairs $(Z_{i,1}, Z_{i,2})$, $i = 1, ..., 30$ and $Z_{i,k} \sim N(0, 1)$, $k = 1, 2$.
2. Let $(X_{i,1}, X_{i,2}) = (Z_{i,1}, \rho Z_{i,1} + \sqrt{1 - \rho^2} Z_{i,2}$
3. Let $(Y_{i,1}, Y_{i,2}) = (G^{-1}(\Phi(X_{i,1})), H^{-1}(\Phi(X_{i,2})))$, where G is the marginal cumulative distribution function for branch 1 and H for branch 2.

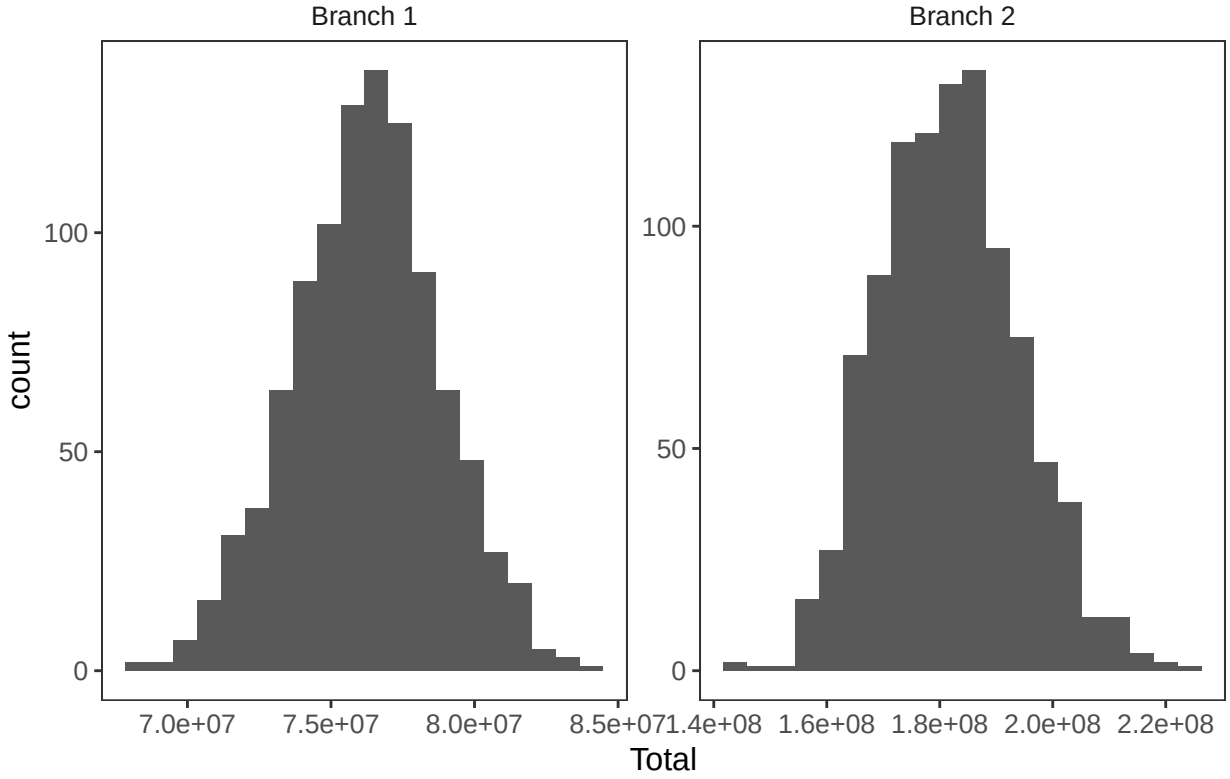
Note that these steps simulate the number of claims for one month and thus the ρ differs from month to month and G and H differs between seasons, as discussed in exercise 1. For each month we use the ρ we have in Table 1.

We could also have used the Gauss-copula-method to simulate claim costs, however since we do not have as much correlation as for the number of claims and that it is assumed that we have an equal number of claims for each branch which is not necessarily true in our case, we chose not to use copulas for the claim costs. We assume that the dependency structure between the branches is captured in the number of claims. Therefore we continue by, for each claim, simulating a cost from the corresponding logspline approximation to assign to that claim and assume independence between costs.

However, since our results in the following exercises did not differ much, using bootstrap compared to the Gauss-copula method, we chose to only include results from the bootstrap method.

We bootstrap 1000 years of data, using the data for our 10 year period, and we can see the distributions of the total annual costs in Figure 18 below. We can see that branch 2 has both a higher expectation and variance compared to branch 1.

Figure 18: Distribution of sampled annual costs of the two branches



Exercise 5

We now want to implement two separate XL-covers. The covers caps losses for the 10% worst claims for each respective insurance branch. For our two branches, based on a large number of simulations, these cut-offs M_1 and M_2 turn out to be the following

Table 2: Cut-offs

Type	M
1	44551
2	153861

We note that these cutoffs are vastly different for the two branches which of course is to be expected since they insure against different things. Graphically the distribution of the individual claim costs looks as follows when, and when not, taking the XL-covers into account.

Figure 19: Simulated claim costs without XL−covers

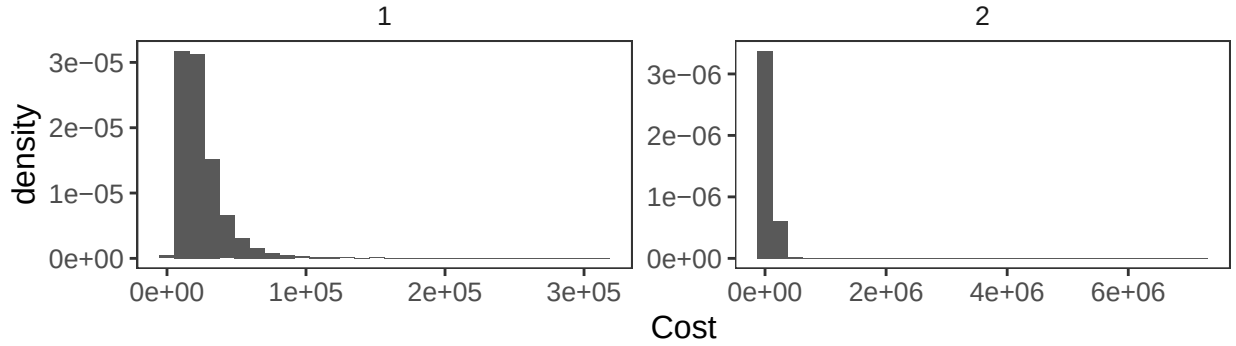
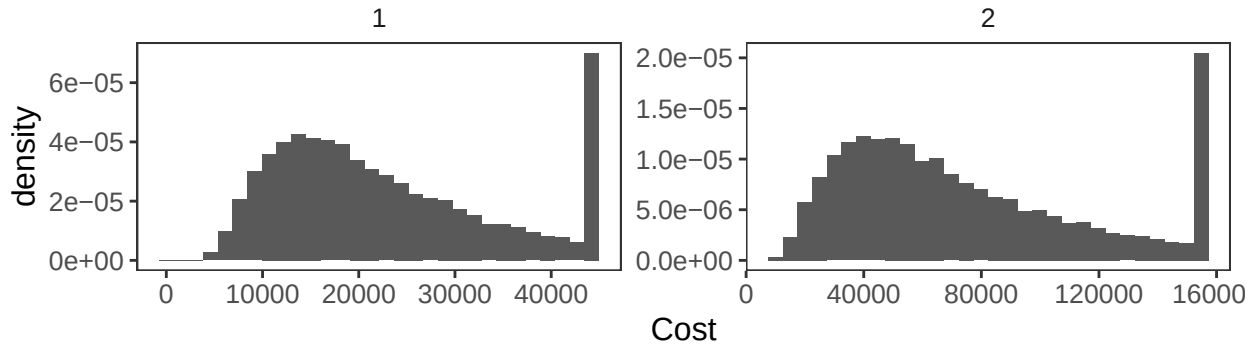


Figure 20: Simulated claim costs with XL−covers

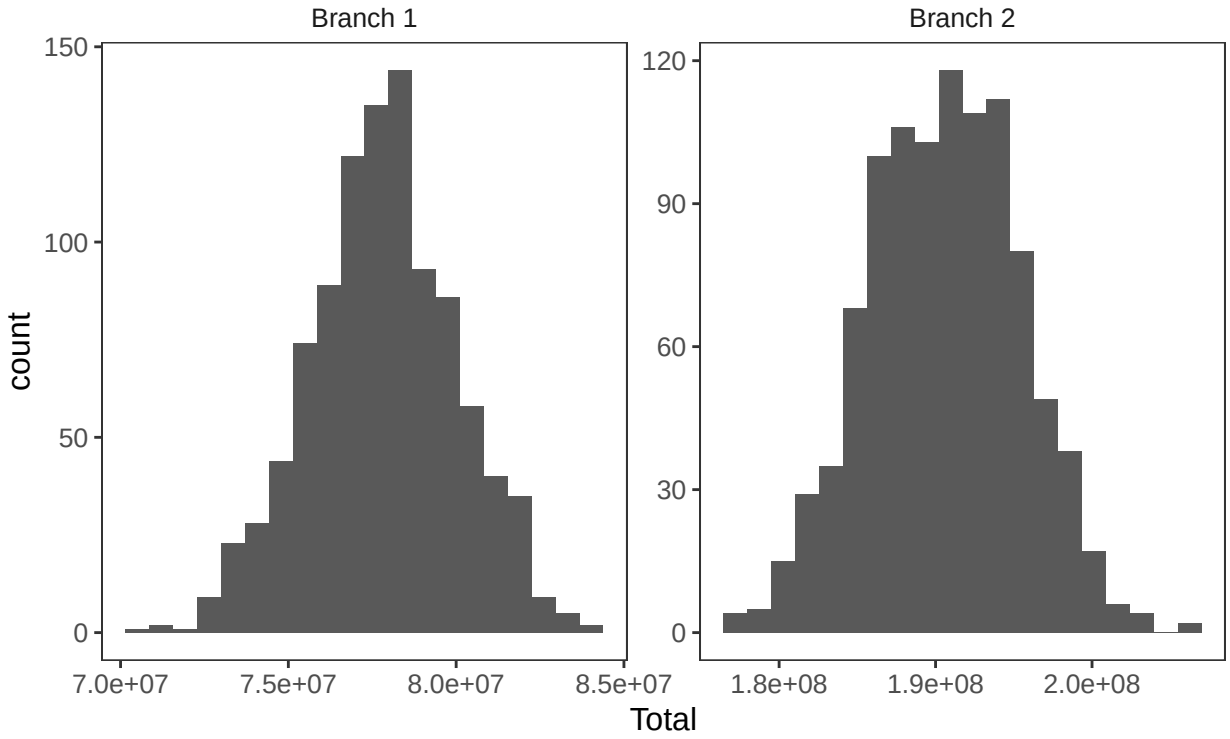


From Figure 19 and 20 we note that both XL-covers seem to cover some very deviating claim costs, specifically in the case of insurance branch 2 which in some sense could be expected due to the higher variance compared to branch 1. Due to the possibility of these large outliers the cost of the XL-covers will be very high.

Table 3: Prices for the two XL-covers

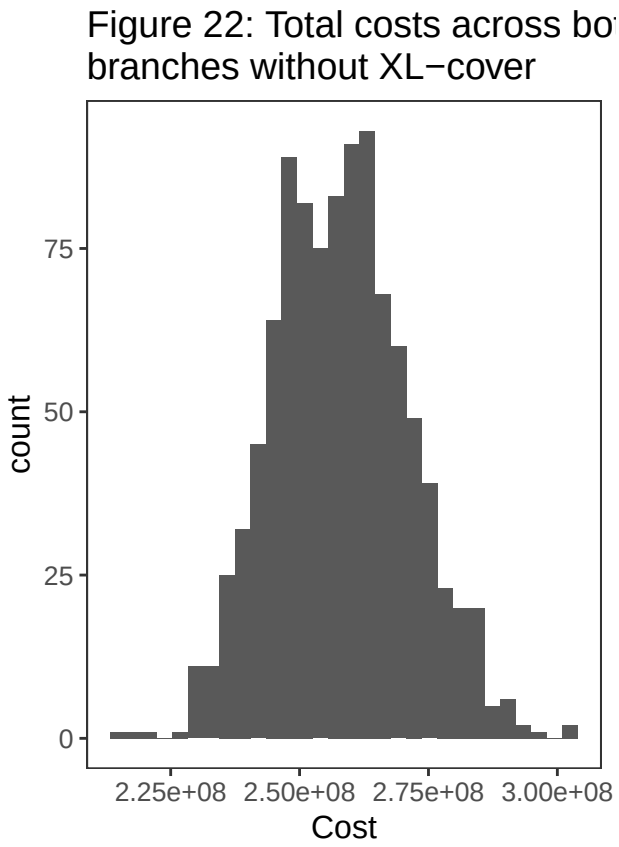
Branch	Price
1	9425536
2	52569247

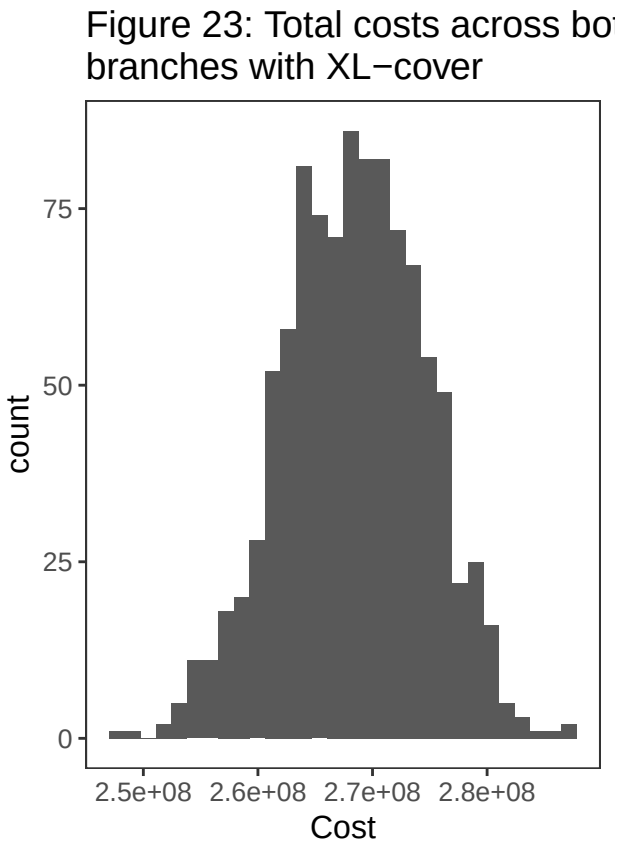
As previously mentioned both covers turn out to be fairly expensive, specifically the cover for branch 2. We now want to examine how the purchase of these XL-covers would impact our total costs for the forthcoming year. We saw a simulated distribution of the losses for the two branches without the XL-covers in Figure 18. If we implement the covers we get the following simulated distributions.



When comparing Figure 18 and Figure 21 we note that that the approximate expected cost for the forthcoming year is slightly higher when purchasing the XL-cover for the respective branches. This increase in the expected cost does however come with the benefit of reducing the weight of the tails of the distributions. As such, the purchase of the XL-covers seem to increase the average cost whilst reducing the overall risk. For a risk averse insurance provide these XL-covers could therefore serve as an excellent tool to reduce the overall risk.

We can also observe the how the purchase of both covers affect the distribution of the overall annual cost.





Once again we see from Figure 22 and 23 that the distribution is lighter-tailed whilst having a higher expectation.

Exercise 6

We now want to implement an SL-cover instead of an XL-cover for both insurance branches. Like the XL-covers the SL-covers insure against the 10% worst annual costs at a price of 120% of the expected cost. We can once again plot the simulated annual costs with and without the SL-covers.

The cut-offs for the SL-covers, based on a large number of simulations, turn out to be the following.

Table 4: Cut-offs

Type	M
1	79545248
2	197907455

Figure 24: Simulated costs without SL covers

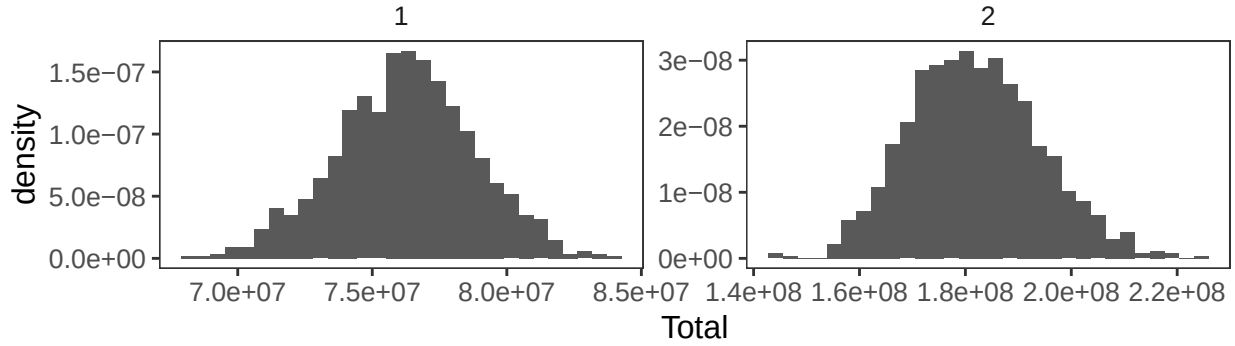
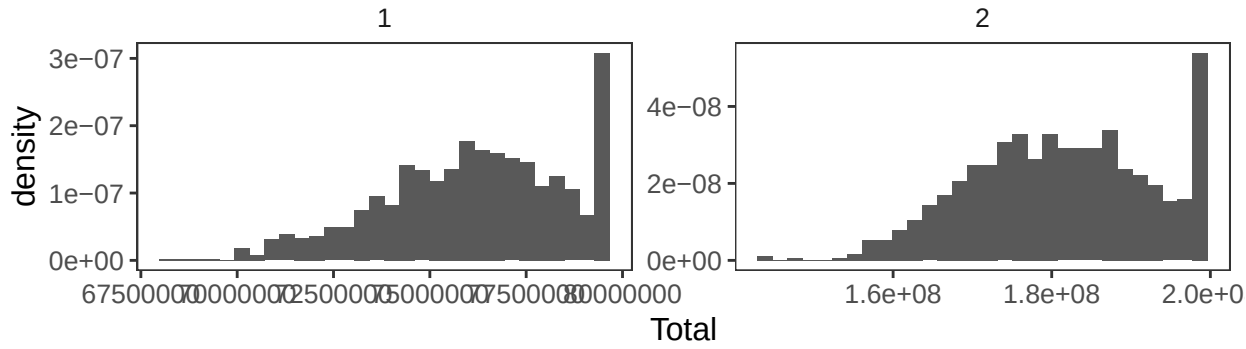


Figure 25: Simulated costs with SL covers



In the case of the SL-covers we note from Figure 24 and 25 that, unlike for the XL-covers, that we have no extreme deviations in the annual cost (compared to the deviations in the distribution of the claim costs). This is due to the fact that the annual cost is the aggregation of a very large number of individual claim costs, thus an outlier of the same proportions as those in the case of the individual claim costs will be theoretically extremely unlikely and practically never occur. In other words, by aggregating the claims we reduce the overall risk of large deviations. As such the prices for the SL-covers should not be as extreme compared to those of the XL-covers.

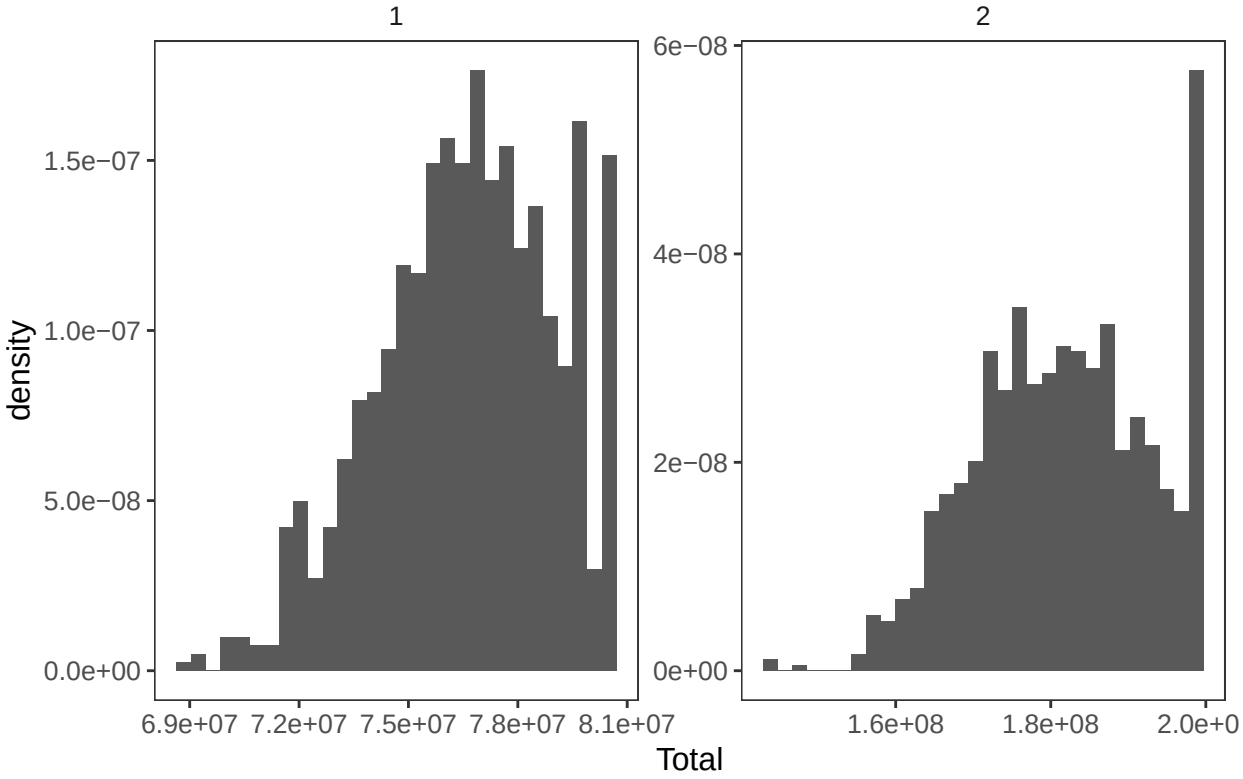
The prices of the two SL-covers are as follows

Table 5: Prices for the two SL-covers

Type	Price
1	137963.8
2	792533.6

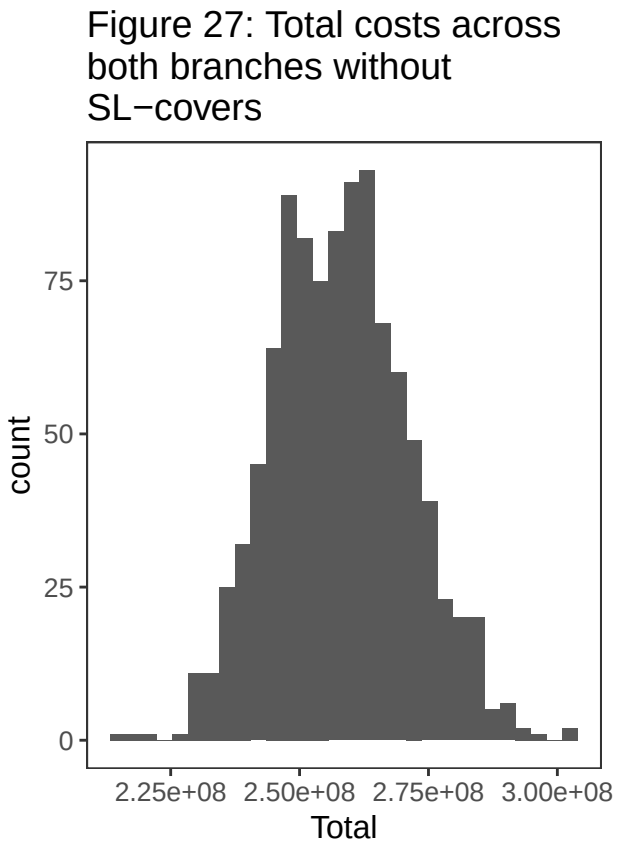
Here we note from Table 5 that the prices are actually extremely low, for both insurance branches the annual cost of the SL-cover is but a drop in the ocean compared to the overall annual claim costs. The price for the SL-cover for branch 2 is a bit more expensive which is due to the fact of the higher variance of the annual claim costs as was previously noted.

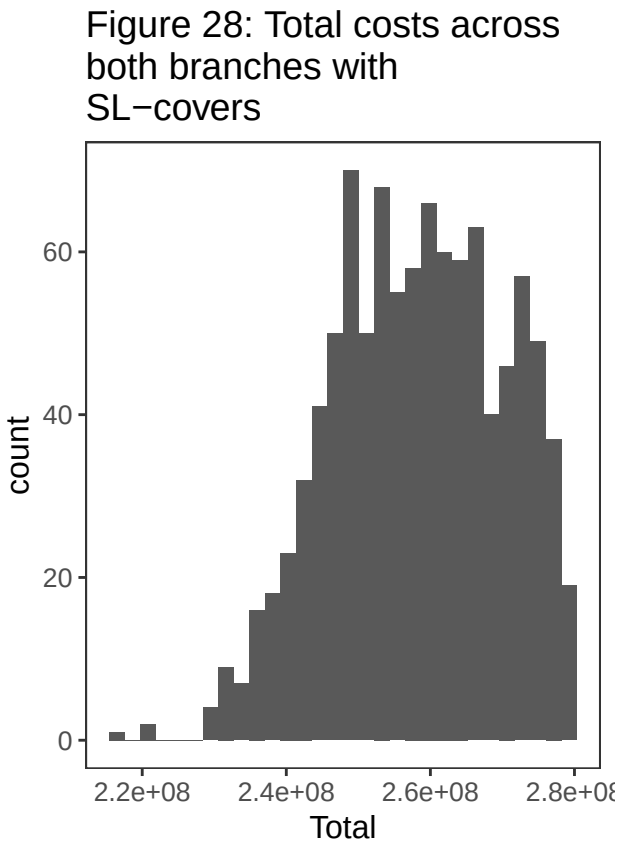
Figure 26: Sampled annual costs with SL−covers



When comparing branch 1 and 2 in Figure 26 we note that the expected cost is slightly higher due to the price of the SL-cover but more importantly that we have cut-offs which essentially prohibits any larger deviations in the overall cost past the cut-off point. By doing this the distributions for the overall costs are skewed slightly positively.

We can also examine how the purchase of both covers impact the overall annual claim costs.





From Figure 27 and 28 we can see that by purchasing both covers we increase the expectation slightly whilst also skewing the distribution positively as in the case of the individual cost distributions of the branches.

Exercise 7

Lastly we want to implement an SL-cover which insured against 10% of the worst total annual costs, i.e. summed over both branches. This cover will be priced and insure like the individual SL-covers in Exercise 6.

Figure 29: Simulated costs without SL covers

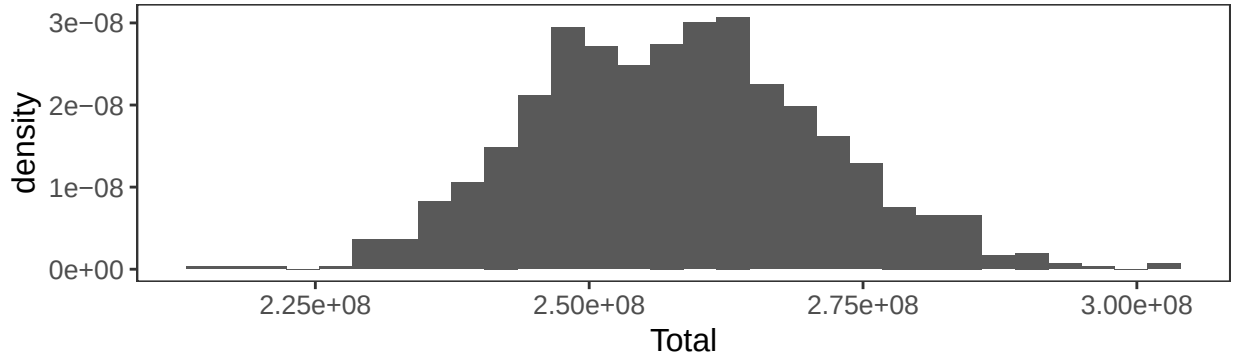
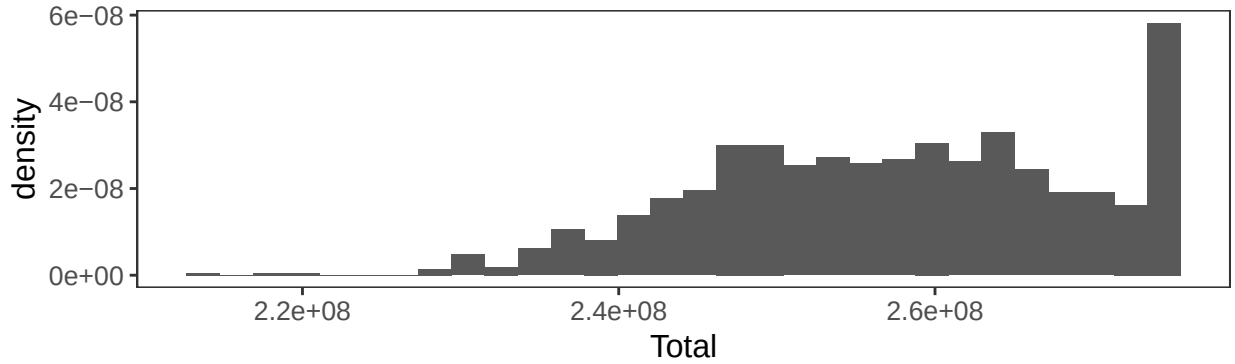
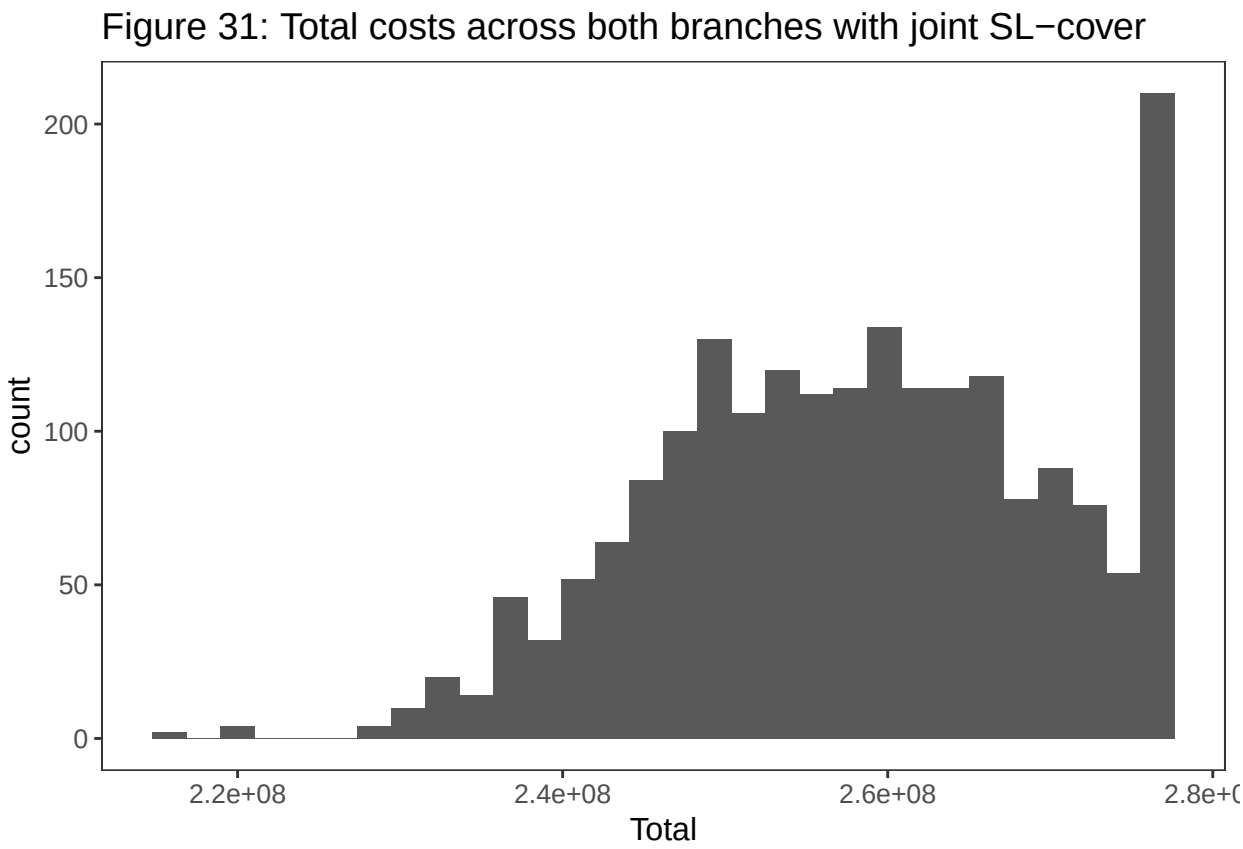


Figure 30: Simulated costs with SL covers



Like for the previous SL-covers we note from Figure 29 that the distribution of the total annual cost without the purchase of the cover lacks any extreme outliers and as a result the overall cost for the SL-insurer and as a consequence the price will be lower.

The cost of this cover, based on simulated data, is 7.672261×10^5 which is once again lower than than the previously examined covers, this is because we have aggregated on yet another level which has further reduced the risk.



We see from Figure 31 that the total SL-cover implies a higher expected cost but also limits the total annual cost to $M = 2.7543585 \times 10^8$. As such, a more risk averse insurance provider might opt for the SL-cover to completely exclude the possibility of large deviations in the annual costs whilst a more risk-taking insurance provide might avoid the SL-cover to reduce the expected annual cost.

Appendix

Figure A1: Histogram of N_{11}

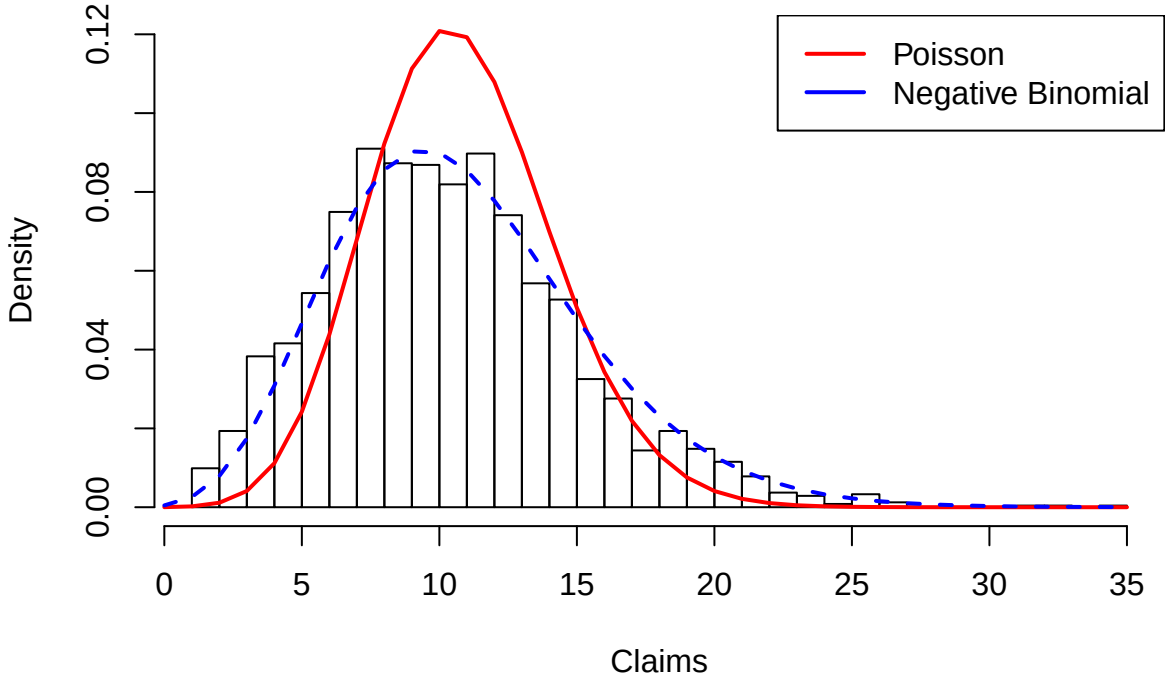


Figure A2: Empirical and theoretical CDFs for N_{11}

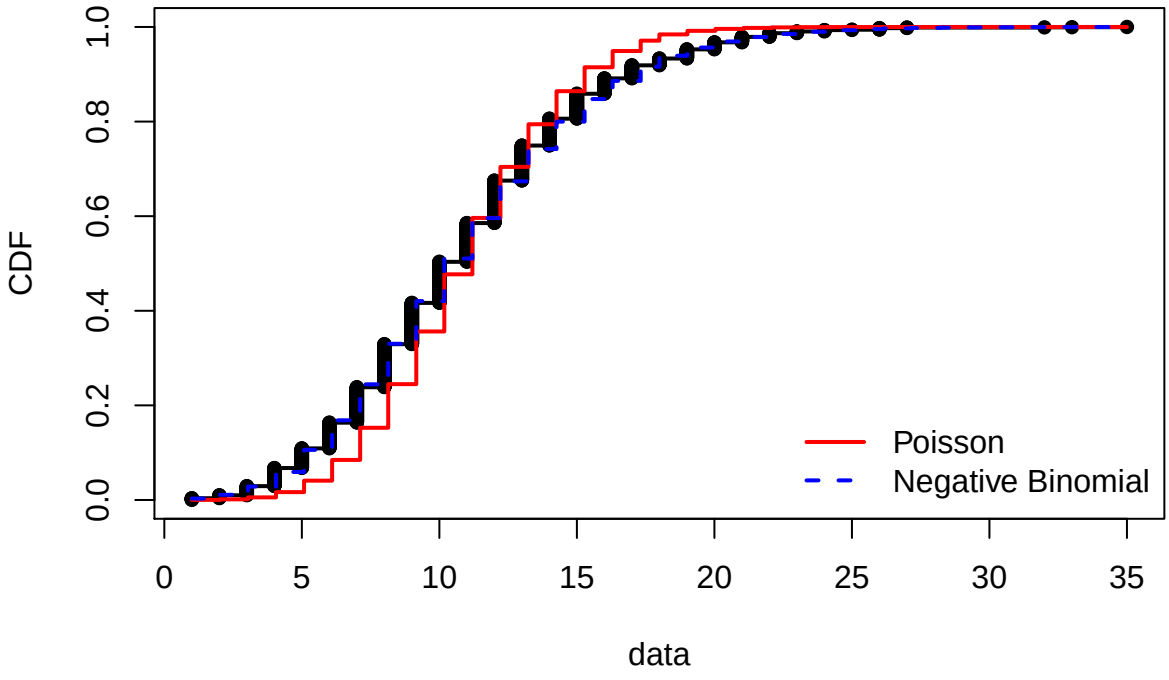


Figure A3: Q−Q plot for N_{11}

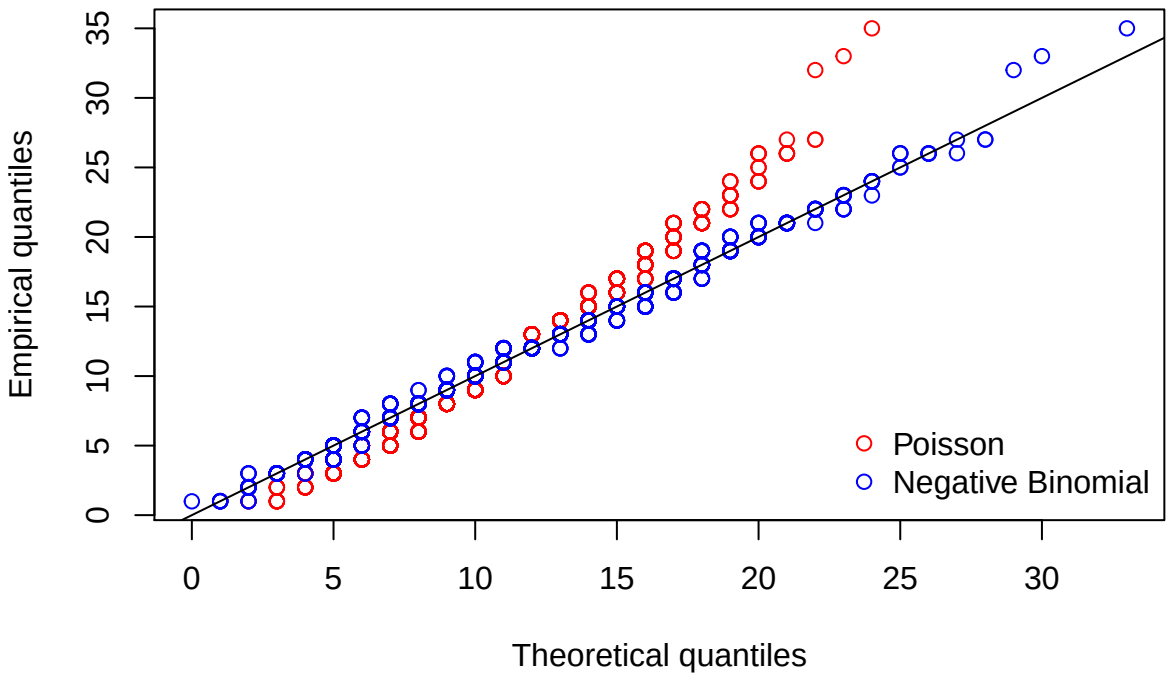


Figure A4: Histogram of N_{12}

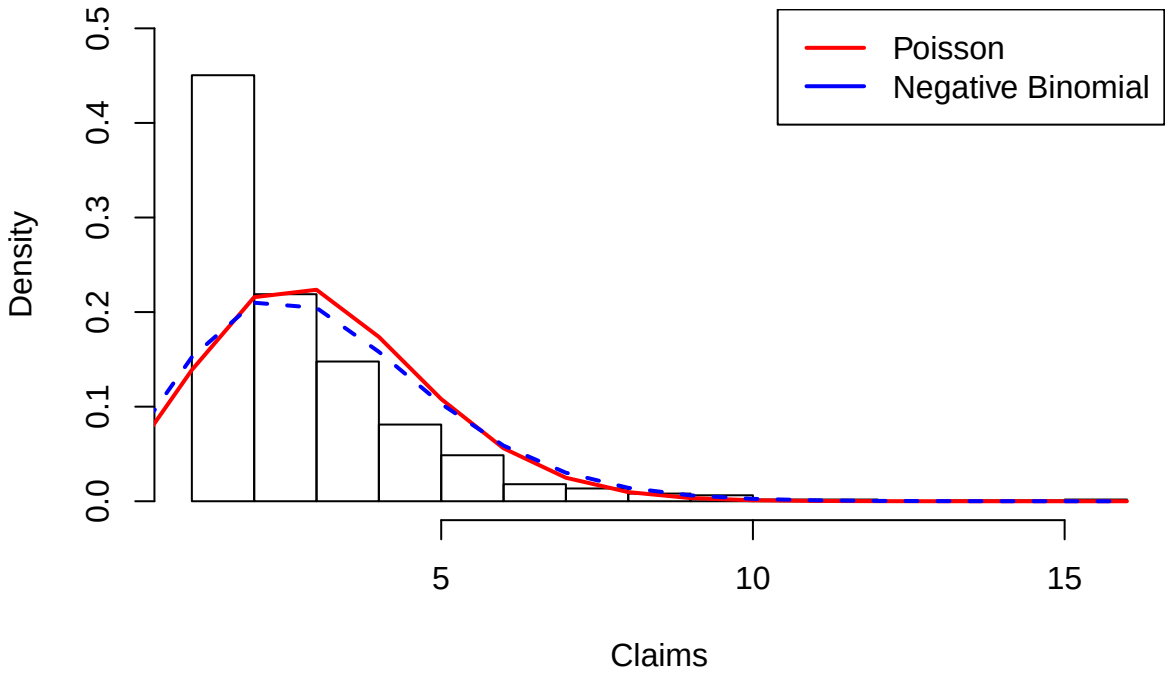


Figure A5: Empirical and theoretical CDFs for N_{12}

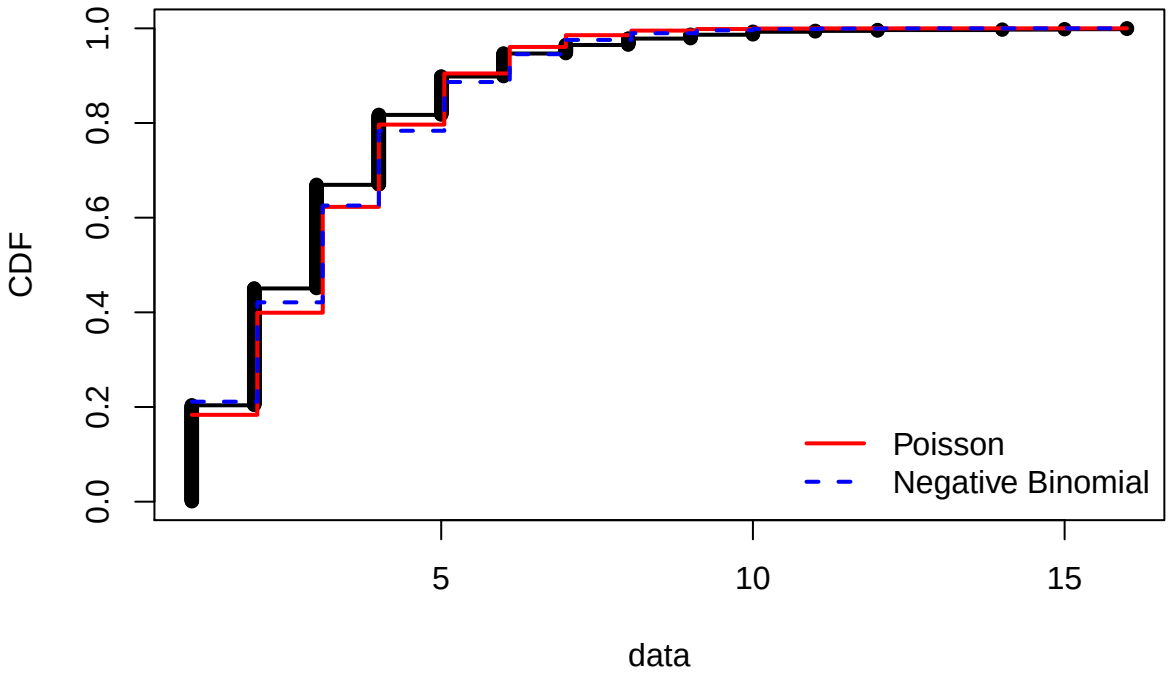


Figure A6: Q−Q plot for N_{12}

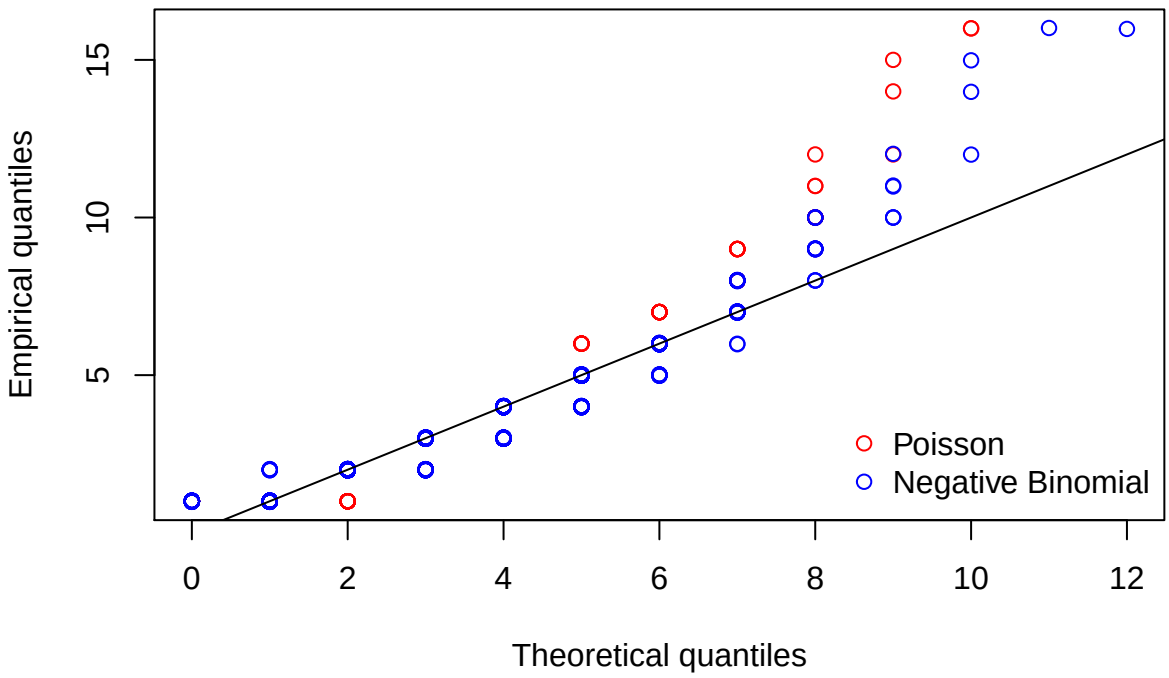


Figure A7: Histogram of N_{21}

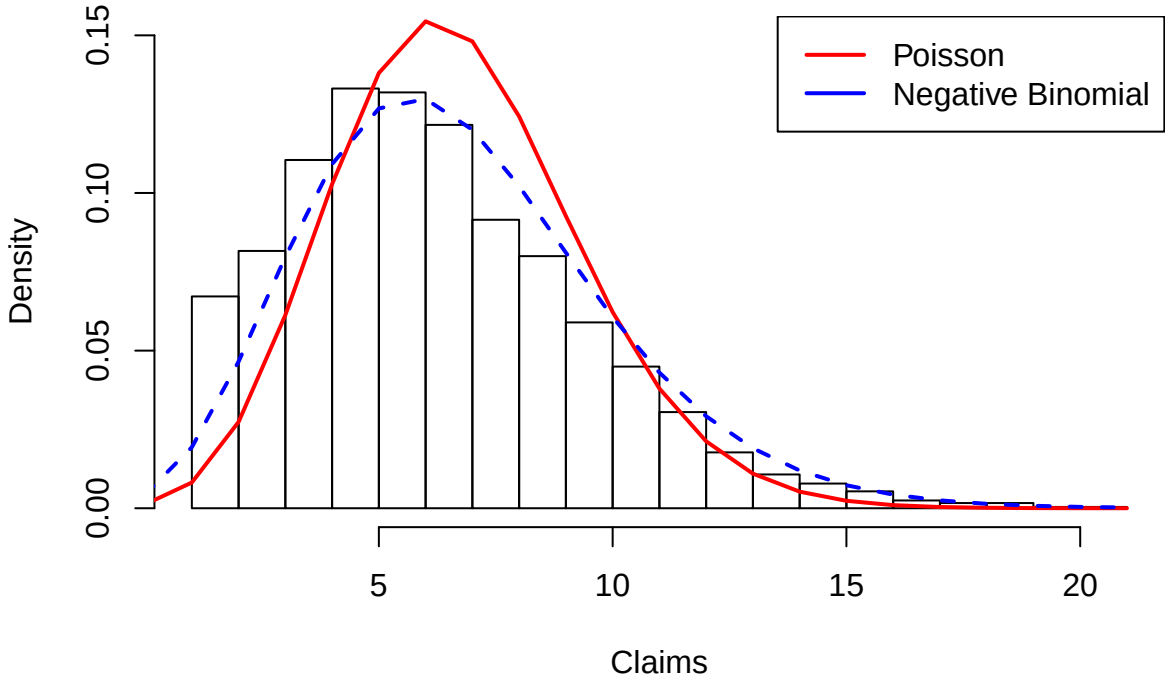


Figure A8: Empirical and theoretical CDFs for N_{21}

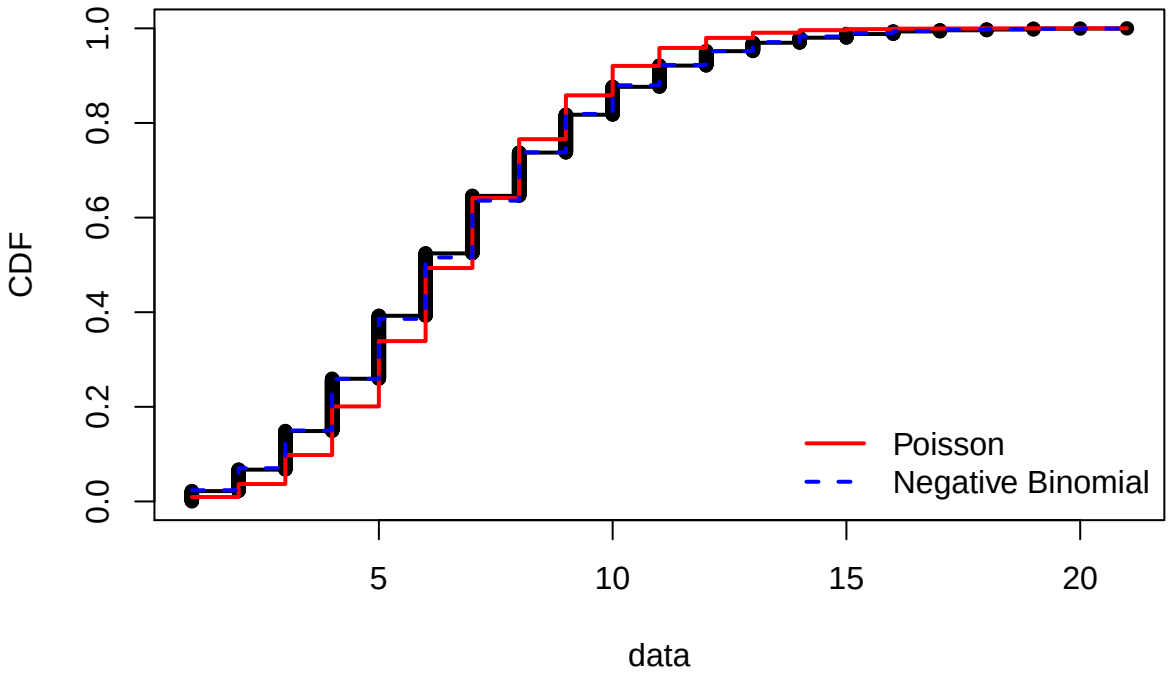


Figure A9: Q−Q plot for N_{21}

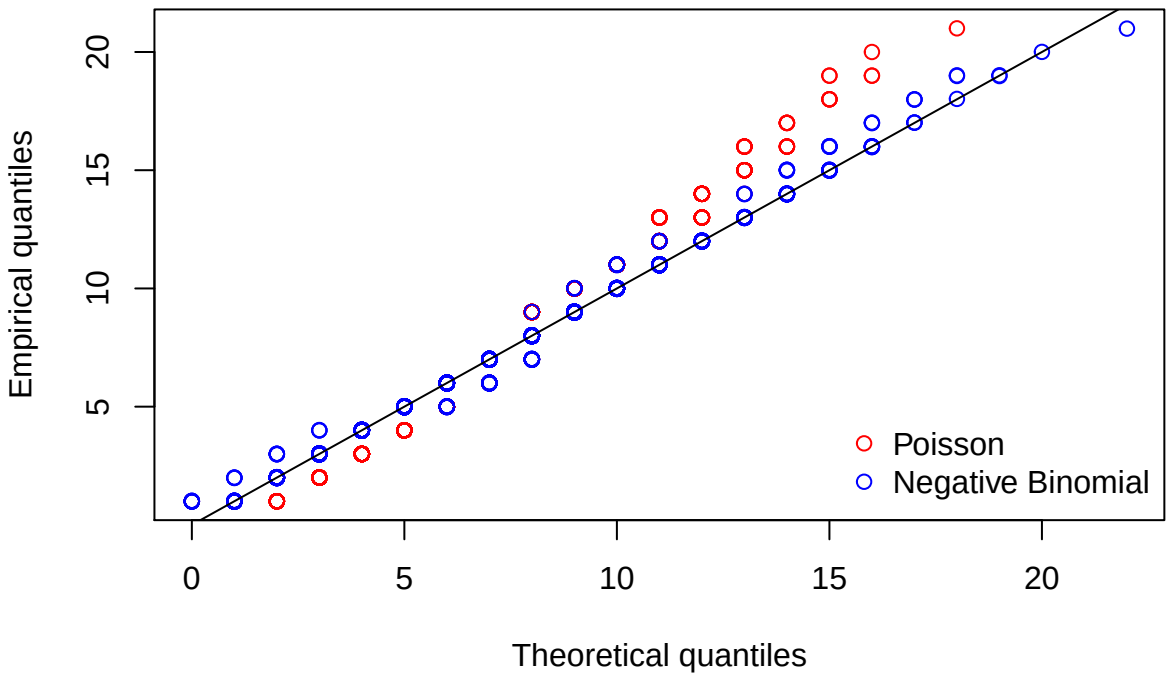


Figure A10: Histogram of N_{22}

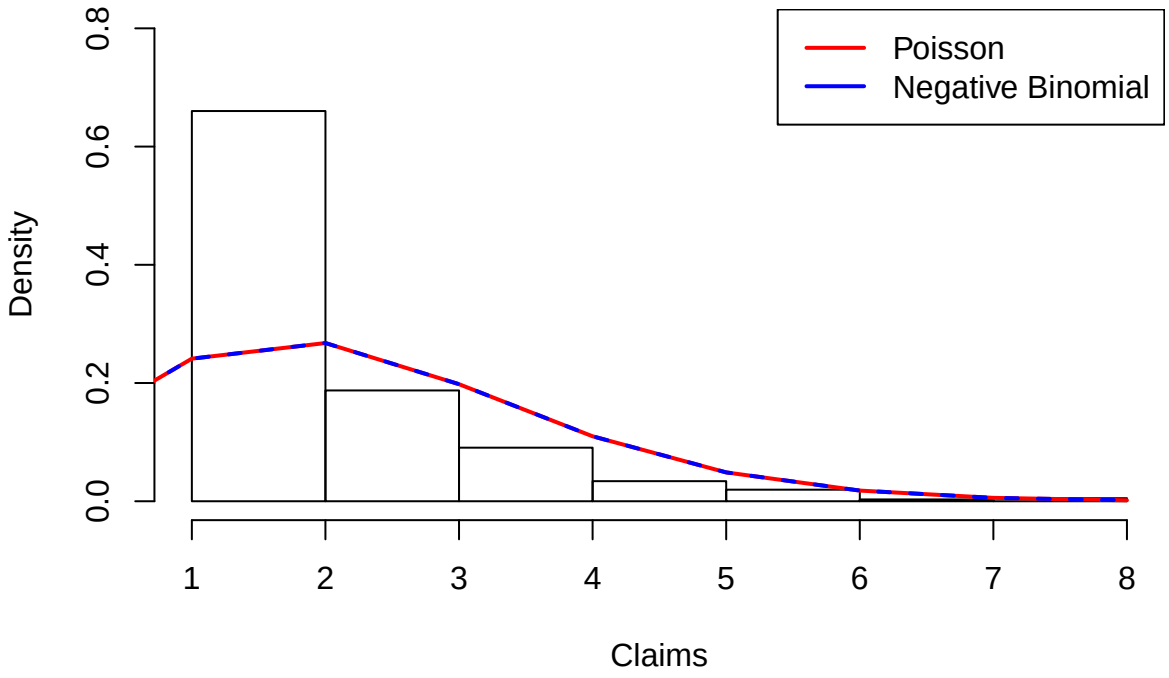


Figure A11: Empirical and theoretical CDFs for N_{22}

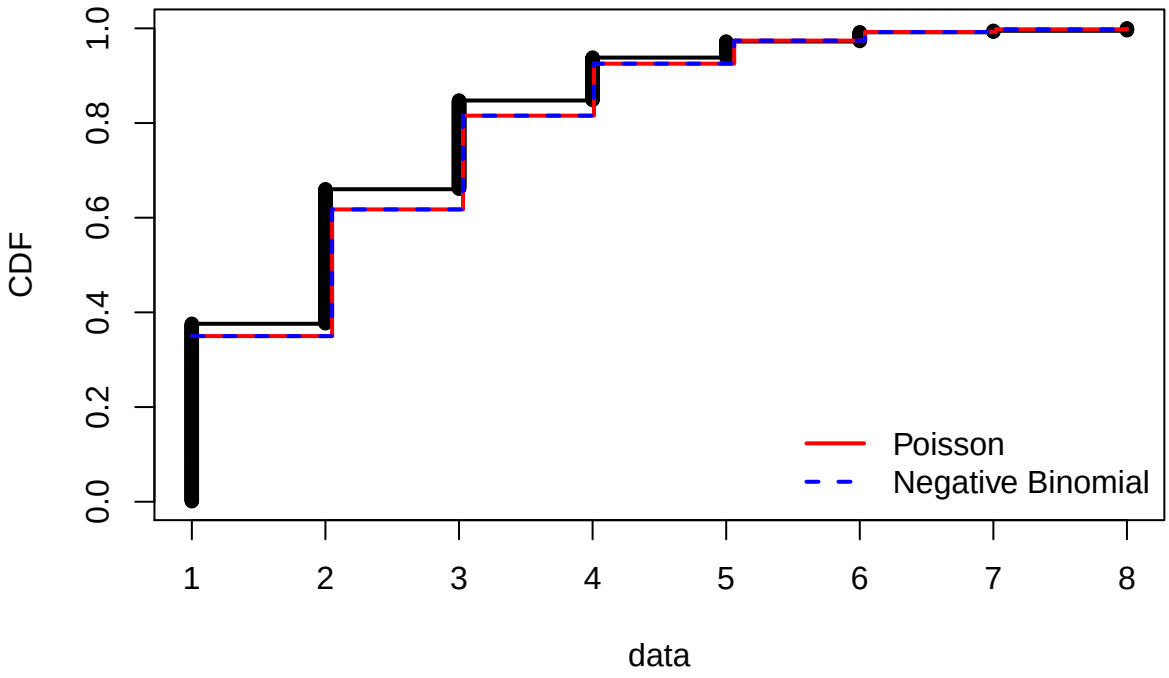


Figure A12: Q−Q plot for N_{22}