Hunter College Master's Project in Statistics

Increased Limit Factor Determination

An Approach Using A Piecewise Distribution

Sean Ammirati
Fall 2017

I) Introduction:

Increased Limit Factors (ILF's) are used by actuaries to determine the appropriate increase or decrease in a policy's premium due to a change in the limit of liability. The determination of these factors influences the pricing and expected losses that arise from policies of different specifications of limits.

The limit of an insurance policy is some number, *L*, after which the insurance company will no longer pay for a loss. That is, it is an upper bound on the amount paid by an insurance company on a loss, known as the severity of the loss. Interesting problems arise when trying to determine the best way to go about pricing based on limits.

The desired end result is to produce some multiplicative constants for common limits that refer to the multiplier that should be applied to a priced premium after all other risks are considered. In order to do this, one uses a *base premium* at a certain limit (often the most commonly used limit, in the case of this project it is 1,000,000) to use as a reference point to determine the other factors.

This means that when we are pricing, we calculate the expected value of the severity given that we have a limit of one million dollars. If we then consider increasing the limit to, say, two million dollars, we will multiply the price by a multiplicative constant, the increased limit factor for two million dollars. This factor would simply be what we have determined as $\frac{E(Loss|Limit = 2M)}{E(Loss|Limit = 1M)}$.

It is then common practice to use interpolation between the points which are determined for any obscure or uncommon limits. Although we do lose some information at uncommon limits which must be interpolated to determine, in practice this makes determining the effect of a limit much easier to implement, understand, and interpret. Since limits do not often fluctuate greatly in practice and tend to be common, well defined numbers, it is useful to have a single multiplier which can account for the changes in risk.

However, to determine these numbers a distribution for losses must be determined. This distribution must take into consideration the limits. Limits follow the law of diminishing returns – that is, an increase of a limit from 1,000,000 to 2,000,000 will have a larger effect on the expected loss than an increase from 10,000,000 to 11,000,000. The ILF curve therefore increases at a decreasing rate, and will nearly stabilize for an adequately large limit.

The first approach to solve this problem is to consider the distribution of losses at each of the historical limits. We can then easily determine the expected loss and the ILFs would simply be the ratios of the expected losses to whatever loss is determined to be the base. Thus, on the surface the solution to this problem appears to be relatively simple.

Practically, however, this is nearly impossible to do, since there is not enough historical data of losses for each limit band. While certain limits may be relatively frequent enough to perform this kind of analysis on, many of the larger limits are not. We also would expect a larger amount of

variation in the losses of policies with higher limits. High variability and small sample sizes in these larger limits make this method unreliable for determining the ILF curve.

If we can find a generalized distribution for the losses, we can calculate the expectations given a piecewise function of severity. That is, letting x be the loss amount and L be the limit:

$$S|L = \begin{cases} x, & 0 \le x \le L \\ L, & x \ge L \end{cases}$$

where S is the severity of the claim to the insurance company.

By the definition of expectation,

$$E(S|L) = \int_{-\infty}^{\infty} s\, f(s|L) ds$$

$$E(S|L) = \int_{0}^{L} x\, f(x|L) dx + \int_{L}^{\infty} L\, f(x|L) dx$$

$$E(S|L) = E(X|0 \le x \le L) + L \cdot P(X \ge L)$$

It follows necessarily that if we can find some probability density function $f(x)$ for the loss amount, the above is directly calculable. Letting $\bar{F}(x)$ be the survival function (i.e. $1 - F(x)$, where $F(x)$ is the cumulative distribution function), the expectation we are trying to determine is:

(a) $$E(S|L) = E(X|0 \le x \le L) + L \cdot \bar{F}(L)$$

The goal of this project is to determine an appropriate probability density function for the loss amounts to determine the above expectation. Once such a distribution function is found, we can estimate the expected severity for each of a predetermined set of k limits $L_1, L_2 \ldots L_k$ and use these numbers to approximate the ILF curve.

This has been explored in numerous different papers and to varying complexity, such as Lee's (1977) famed graphical approach to finding ILFs, Halliwell's (2012) use of a mixed exponential distribution, and Wang's (1995) hazard function transforms for reinsurance, among many others. Probabilistic, parametric approaches are favored in general in the industry and much attention has been put towards creating better estimates of these factors.

I have attempted to fit a piecewise distribution to the losses, such that one portion of the losses is follows one distribution, and the remaining portion follows another. This result could be generalized to any k splits of the distribution and the theoretical results would be similar– however, determining thresholds would then be a multivariate problem. This will drastically increase the complexity of the model, and will almost certainly cause overfitting issues. Since the two-parted split fits our intuitive ideas of how parts of the distribution are separated, it appears to be best course of action in ILF determination.

My methodology can be easily generalized – using this procedure for different belly and tail distributions should have no different implementation besides those which are used in R functions (as the *fitdistrplus* package allows use of user-defined functions).

II) The Data

The data I worked with is primary data from an insurance company on various losses in different business units in their Professional Liability insurance unit. The three primary groups of losses are for Professional Lines, Financial Institutions, and Commercial Management Services. Within these larger groups, we have sub coverages pertaining to more specific types of insurance, i.e. within Professional Lines there is Cyber, Small Film, etc.

When an underwriter wishes to write a policy, they select from these sub-coverages, and the limit factor which is applied to determine the final premium amount is particular to that sub-coverage.

The goal of this project is to determine, if possible, ILFs by fitting distributions to these different groups. I started by considering the overall groups of ProLines, FI, and CMS, and then attempted to fit distributions to the smaller sub-coverages. As discussed later, this came with some issues that are mainly manifested through lack of data.

Since much of the data used in the actual application of this project is proprietary, I cannot disclose results which were extrapolated from actual losses. This paper outlines the general approach to solving the problem using a piecewise distribution function, and posits results from a simulated distribution based on actual losses. No losses indicated in this paper are true losses experienced by Axis Capital.

III) Considerations for Choosing a Distribution

The challenges in finding an appropriate distribution are due in part to the fact that we are dealing with a limited amount of information. Another consideration is that the tails of the distribution are often affected by catastrophic events. It is often desired in the insurance industry to take special care in modelling these catastrophic events, as they can ultimately have a very large impact on the company. Using the same distribution for the center and the tail of the distribution may be very misguided, especially based on the sparse amount of data we have in the tails (luckily for the company, but bad for such an analysis).

It is for this reason that I decided to split the distribution up into two parts – one for the vast majority of claims, and one for the outliers on the far right of the tail.

For the center or "belly" of the distribution, it is not immediately clear how to determine which distribution will fit best. There are, however, some clear choices of which commonly considered distributions could *not* work for this data.

Looking at a histogram of the losses, we can tell that the loss amounts are highly skewed. This indicates that certain distributions are inappropriate immediately – the exponential distribution, normal, or gamma distributions, for instance, will have far too little skewness for this to work correctly. This is explored later in the discussion of the *fitdistrplus* package in the R programming language, which makes it quite easy and intuitive to determine which distributions are appropriate for a given set of observations.

The nature of the data itself ($x>0$ and the desire to find expectations) excludes distributions like the normal, beta, and other, more obscure distributions, like the Cauchy. Expectedly, when trying to fit any of these distributions by maximum likelihood estimation, the distributions were either unable to fit entirely or were considerably bad fits when compared to distributions like the lognormal and Weibull) which more adequately fit the data and our preconceived notions of their distributions.

For the tail of the distribution, I have selected the Pareto Type II (also known as the *Lomax* distribution.) From here forward, the Pareto distribution will refer to the Pareto Type II distribution. There is a lot of evidence of the efficiency of the Pareto distribution in the determination of tails of loss curves, as explored by Goovaerts, Kaas, Laeven, Tang, and Vernic (2005) and Guillen and Sarabia (2011).

The Pareto distribution is an appropriate choice, in short, due to the fact that it is a long-tailed distribution. This lends itself to many properties that are useful in this case – it has an increasing conditional expectation in differences for values over a given threshold, it has a tail which decays slower than a power function, and exhibits the single big jump or "catastrophe principle".

Practically speaking, since the Pareto will decay more slowly than other distributions, it is the ideal candidate for the tail of the curve. Since there may be some issues with censored observations

(observations which themselves hit the limit in our historical data), this has an additional benefit in that it will provide some relative conservatism to our estimates, which is preferable. This discussion in the consideration of censored observations will be explored more in depth later in the paper.

IV) Definitions of Distributions

To be clear about the distributions used in this paper, I will define them formally here. The distributions in consideration are the log-normal and Weibull for the "belly" of the distribution and the Pareto distribution for the tails. To begin with, I will discuss the importance and nomenclature of so-called "heavy", "fat" and "long" tailed distributions. This motivates the reasoning behind choosing these distributions.

It can be somewhat confusing to differentiate between these three descriptive qualities of distributions. Although they are often used interchangeably when describing a distribution, these descriptions mean very specific, and separate, things. To be specific, heavy tailed distributions are a superset of the others.

A distribution is heavy-tailed, as described by Rolski and Scmidt (1999), if it goes to zero more slowly than an exponential function. In formal terms, if *f(x)* is the pdf of a heavy tailed distribution, then the moment generating function of *f* is infinite for all t>0, or:

$$E(e^{tx}) = \int_{-\infty}^{\infty} e^{tx} f(x)dx = \infty, \text{for all } t > 0$$

Interestingly, this does not necessarily imply that there exist no finite moments, or that the distribution is limited to a countably high number of moments necessarily, although this is the case in many situations. A good example would be the lognormal distribution, which – because of its relationship to the normal distribution – does have finite moments for each moment *k,* but does not have a finite moment generating function when t>0.

There are subtle differences between the other classifiers, and they are often interchanged. A random variable *X* is said to have a *long-tailed* distribution, as described by Asmussen (2003), if:

$$\lim_{x \to \infty} P(X > x + t | X > x) = 1$$

This can be interpreted as follows: as the random variable X approaches infinity, the probability of it being larger than some value x approaches the probability of it being larger than some other arbitrary value larger than x. All long-tailed distributions are, by definition, heavy-tailed, but the converse may not necessarily be true. Most commonly used distributions with these properties are both, such as those in consideration here: the lognormal, Weibull, and Pareto distributions.

Another important concept at work here is subexponentiality, as described by Embrechts, Klüppelberg and Mikosch (2013), which gives rise to the "catastrophe principle." That is, if $X_1, X_2, \ldots, X_n$ are *n* independent and identically distributed variables, then the distribution is subexponential if

$$P(X_1 + X_2 + \cdots + X_n > x) \sim P(\max(X_1, X_2, \ldots, X_n) > x)$$

This is a surprising result, and makes intuitive sense as to why this property can be useful. The probability of the sum of some *n* variables in a subexponential distribution being larger than some value x_0 is asymptotically equivalent to the maximum of the variables being larger than that value. This is what is known as the "catastrophe principle" or the "single big jump."

Again, all subexponential distributions are, by definition, long tailed, but the converse may not necessarily be true. In the case of the lognormal, Pareto and Weibull distributions, all three of these characteristics are satisfied. This indicates why these distributions are ideal for modelling our losses – they allow for some pretty large outliers, with a high variance, to be accounted for in our distributions.

Log-Normal

The log-normal is so named because it has a direct tie to the well-known normal distribution. If X is a normally distributed random variable, e^X has a log-normal distribution. Conversely, if X has a log-normal distribution, ln(X) has a normal distribution.

The distribution for the lognormal is then easily calculated by doing a change of variables. The probability density function (pdf) of the lognormal distribution is:

$$f(x|\mu,\sigma^2) = \frac{1}{x}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}$$

The cumulative distribution function (cdf) of the lognormal distribution is:

$$F(x|\mu,\sigma^2) = \Phi\left[\frac{\ln(x)-\mu}{\sigma}\right]$$

where Φ is the cdf of the standard normal distribution.

Here, the parameters μ and σ are the mean and standard deviation of the variable's natural logarithm – the normal distribution associated with X.

For our purposes, what is most important about this distribution is that there is a closed solution to the maximum likelihood estimators which makes them convenient to use. The maximum likelihood estimators are:

$$\hat{\mu} = \frac{\sum \ln(x)}{n} \quad , \quad \hat{\sigma}^2 = \frac{\sum(\ln(x)-\hat{\mu})^2}{n}$$

This is a familiar form since it is directly relatable to the maximum likelihood estimators of the normal distribution.

Though there are many interesting and useful properties about this distribution, I will focus on those that are directly applicable to the problem at hand. The log-normal distribution is

frequently used in many applications, especially in finance. For example, there is evidence that the first 97%-99% of the income distribution is distributed lognormally, as described in Clementi and Gallegati (2005). The log-normal is often used to describe exchange rates, price indices and stock price indices as well, as per the Black-Scholes model. Also, noting that the natural logarithm, *ln(x)*, is only valid for X > 0, it is convenient to use for heavy-tailed distributions which are strictly greater than zero, which loss amounts are.

It is worth noting that the log-normal distribution has a closed-form conditional expectation. This will be approximated later by a numerical integration technique, but it is also possible to solve directly in the case of the lognormal distribution.

The closed form conditional expectation is:

$$E(X|X < k) = e^{\mu+\frac{\sigma^2}{2}} \frac{\Phi\left[\frac{\ln(k) - \mu - \sigma^2}{\sigma}\right]}{\Phi\left[\frac{\ln(k) - \mu}{\sigma}\right]}$$

$$E(X|X \geq k) = e^{\mu+\frac{\sigma^2}{2}} \frac{\Phi\left[\frac{\mu + \sigma^2 - \ln(k)}{\sigma}\right]}{1 - \Phi\left[\frac{\ln(k) - \mu}{\sigma}\right]}$$

where Φ is the cdf of the standard normal distribution.

Weibull

The probability density function of the Weibull distribution is:

$$f(x|\lambda, k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-\left(\frac{x}{\lambda}\right)^k}$$

Notice that when k = 1, this becomes the exponential distribution. The parameters k and λ are referred to as the *shape* and *scale* parameters respectively.

The cumulative distribution function of the Weibull distribution is:

$$F(x|\lambda, k) = \begin{cases} 1 - e^{-\left(\frac{x}{\lambda}\right)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

As mentioned before, the Weibull distribution has the three mentioned desired properties of tails. It is quite a variant distribution, and can be used to model many different phenomena. There is no closed form for the likelihood estimates for the Weibull distribution, and they must be estimated by iteration.

Pareto (Type II or Lomax Distribution)

The Pareto Type II distribution, referred to simply as the Pareto distribution in this paper, has probability density function:

$$f(x|\lambda,\alpha) = \frac{\alpha}{\lambda}\left(1+\frac{x}{\lambda}\right)^{\alpha-1}$$

The parameters α and λ are referred to as the *shape* and *scale* parameters respectively.

The cumulative distribution function of the Pareto distribution is:

$$F(x|\lambda,\alpha) = 1-[1+\frac{x}{\lambda}]^{-\alpha}$$

There are many important applications and useful properties of the Pareto distribution. When considering the tails of the distribution, of all three distributions described, it has the heaviest tail, meaning its pdf goes to zero slower than the lognormal or Weibull.

It was original developed by Pareto (1964) to describe the allocation of wealth among individuals, which is historically well-known to have a very heavy tail. This is often related to the "Pareto principle", or the "80-20" rule, which posits that 80% of the wealth of a population is owned by 20% of the population. This is seen approximately many times in the real world, although this is simply a ballpark estimate and refers to the selection of a particular alpha above. Still, the Pareto distribution excels at describing situations where there are extreme outliers.

It is also "heavier" than a power function, not simply an exponential function as with the two earlier distributions, which makes it an ideal candidate for the fat tail associated to losses.

Although the selection of these distributions seems to be well-tuned to the problem at hand, especially when looked at for the dataset I was working with, other distributions could potentially be used with the same methodology as described in this paper. Other potential distributional selections include the Gamma, generalized Pareto, Burr and Lévy distributions.

V) Methodology for Selecting a Split Distribution

In order for this methodology to work, I must select some threshold *T* where the shift in distribution occurs. Therefore, we have a piecewise density function, such that:

(b)
$$f(x|T) = \begin{cases} \alpha b(x)/B(T), & 0 \le x \le T \\ (1-\alpha)z(x)/(1-Z(T)), & x \ge T \end{cases}$$

where *b(x)* and *z(x)* are the belly and tail distributions, respectively. These densities are divided by the probabilities of the conditions occurring (where *B* and *Z* are the cumulative distribution functions) to ensure the individual pdf integrate to 1. It is then multiplied by constants based on where the split is made in order to ensure that the overarching pdf will integrate to 1.

We can also determine a cumulative distribution function (cdf) from this information. It follows that the cdf is the following

(c)
$$F(x|T) = \begin{cases} \alpha\, B(x)/B(T), & 0 \le x \le T \\ \alpha + (1-\alpha)(Z(x) - Z(T))/(1-Z(T)), & x \ge T \end{cases}$$

where *B(x)* and *Z(x)* are the cdfs of the above-mentioned distributions.

Although it may seem that these distributions will be hard to manage, since we are only interested in point estimates of the expectation above (at point **(a)**), this allows us to use this more complex, less intuitive model to account for our differing initial inclinations about the belly and tail of the distribution.

The next step is two-fold: we must determine which distribution functions to use for the *b(x)* above (as we are assuming that *z(x)* follows a Pareto for convenience), and then we must determine the optimal split, α, on which to divide the data.

For the distributions of *b(x),* after consideration of a number of distributions, it appears that the Weibull and Lognormal distributions, as described earlier, are most apt to describe the loss data observed – they all have long tails, are right-skewed, and make sense with the nature of the data (they are positive and have finite expectation).

For each coverage type (ProLines, FI and CMS), and then for each subcoverage within them, all three distributions are used to determine the optimal fit. The optimal fit is determined by whichever has the largest log likelihood. Therefore, differently split distributions have different distributions fit to the belly. In most of the cases, Lognormal was the best fit, followed by Weibull.

Useful Packages

One advantage of using the R programming language for tasks such as these are that there are many open source packages which have been created to solve similar problems. In this project, two such packages which proved to be extremely useful were the *fitdistrplus*, and *actuar* packages.

The *fitdistrplus* package, by Delignette-Muller and Dutang (2015), has useful features for distribution fitting that are ultimately missed in the base version it expanded on (*fitdist*). *fitdistrplus* still uses optimization methods when there is no closed form solution to the likelihood determination, but also provides way of determining which fits are most appropriate as well. For each fitted distribution, *fitdistrplus* provides a comparison of the density to a histogram of the data, a CDF plot, Q-Q and P-P plots. This makes the process of determining how well a distribution can fit the data using graphical interfaces much simpler.

The *actuar* package, by Dutang, Goulet and Pigeon (2008), was specifically created by actuaries and for actuaries to automate industry-specific tasks. Of particular use in this case is, of course, the distribution function of the Pareto distribution. The package serves to add many uncommon distribution functions which are used by actuaries and integrates them seamlessly with the base R environment, so they may be used essentially identically to the base functions.

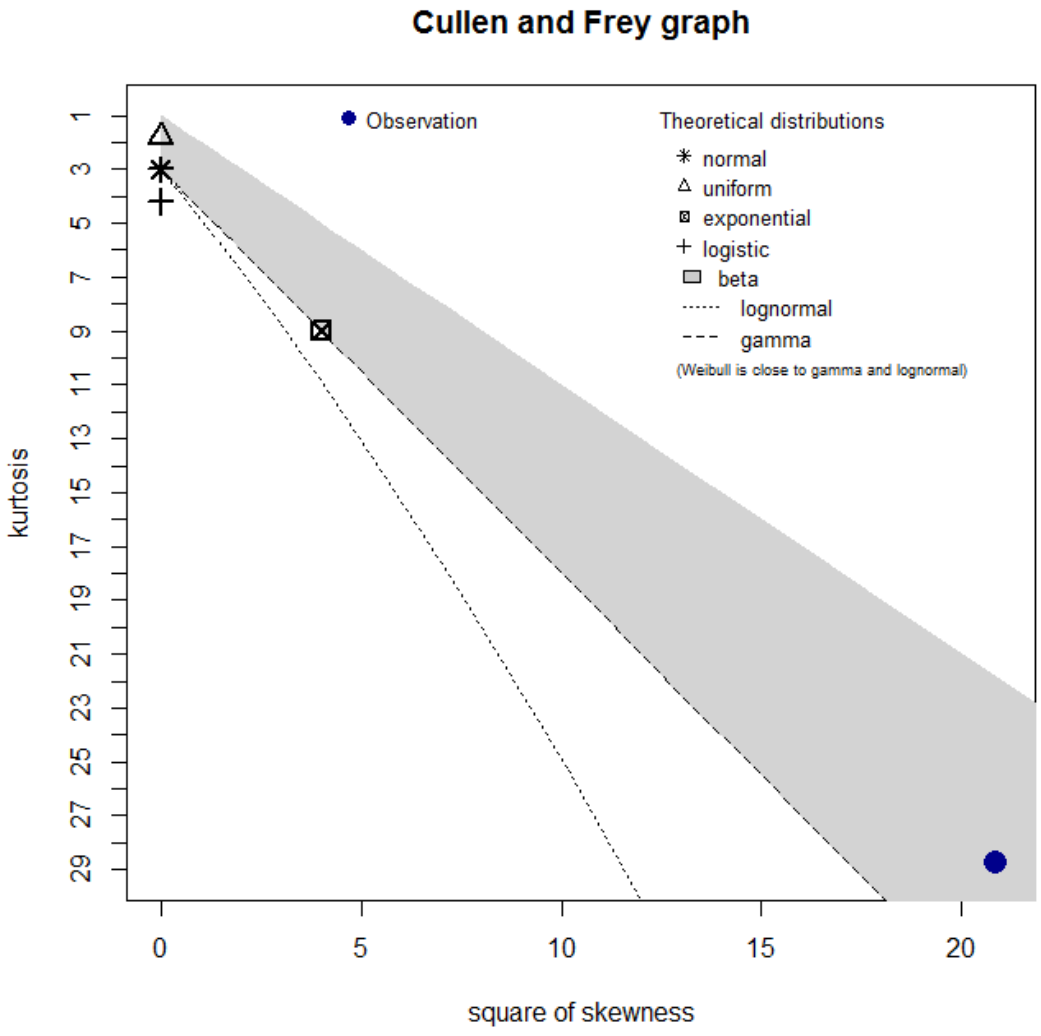
VI) Using *fitdistrplus* To Find Best Fitting Distributions

For the remaining analyses, I will be using a simulated version of the AxisPro dataset. That is, I will simulate from a distribution which was the end result of this project, with a lognormal belly and a Pareto tail. This is for illustrative purposes only, and this does not relate to any actual losses.

To illustrate the power of using some of the utilities included it the *fitdistrplus* package, I will show the results of the *descdist* function. This function estimates the skewness and kurtosis of the given dataset and provides a simple graphical representation of how they compare to common distributions, like the normal, exponential, gamma, log-normal, and beta distributions. This proves extremely useful in determining which distributions are appropriate to fit to a given dataset.

Setting boot = TRUE will also use bootstrapping to estimate a range where the possible distributions can be.

Below is an example of the output:



This result is not very satisfying: it is suggesting a Beta distribution, but a Beta would simply be inappropriate for this particular problem (as it has a range between 0 and 1). This is another reason for the approach I have taken -- splitting the data was a way to account for the high skewness of the empirical distribution. We can see from the graph above that both the skewness and kurtosis of the empirical distributions is estimated to be quite a bit higher than our suspected candidates: the Weibull and lognormal distributions. Using the Pareto for the tails as the end result is an attempt to mitigate this issue. As we will see later on, it was partly but not entirely successful at fixing the skewness issue.

Another useful feature of the *fitdistrplus* package is the ability to bootstrap parameter values to get a confidence interval around predicted values. This can give you a sense of how variable your fitted distribution would be to new data points, as well as indicating how convincing the parameter values estimated by the *fitdist* function are.

Here is an example of the output:

```
1. boots <- bootdist(simexLN$dist$belly)
2. summary(boots)
```

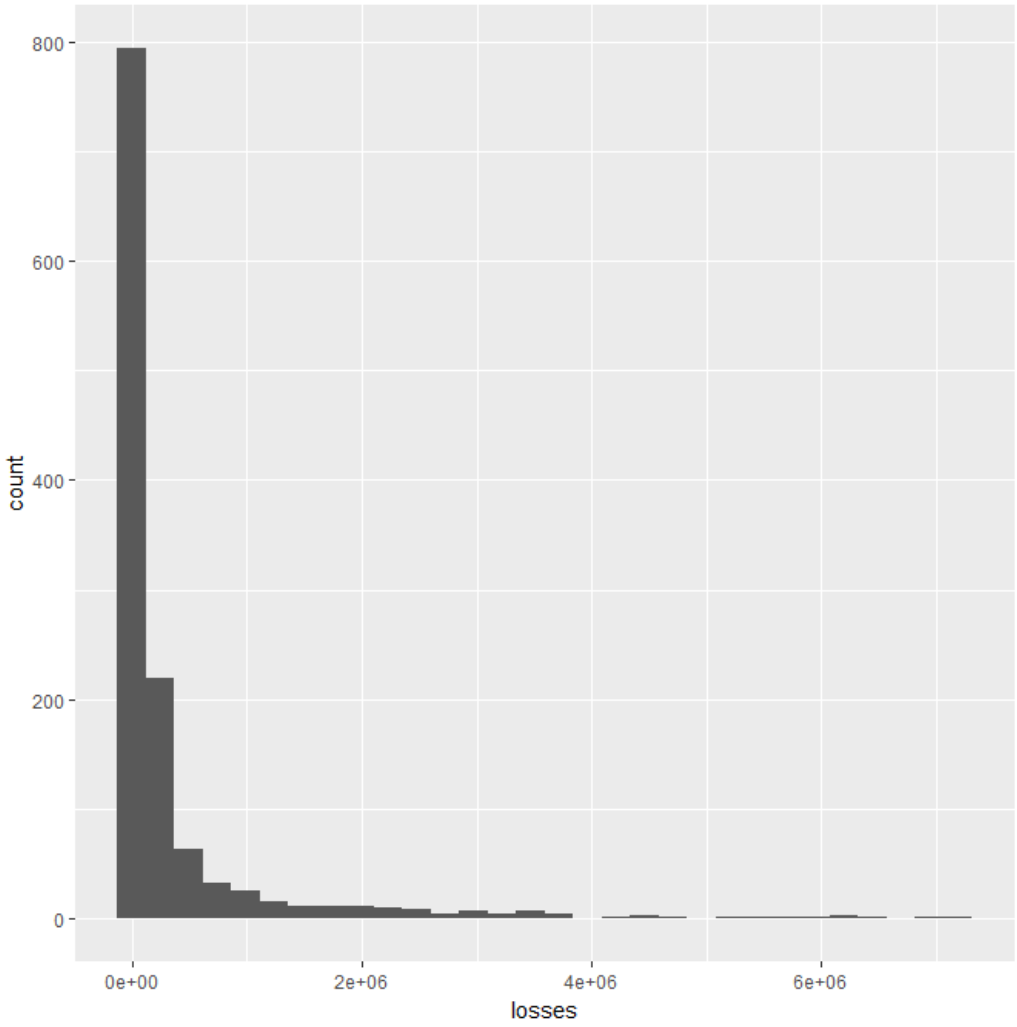
```
Parametric bootstrap medians and 95% percentile CI
           Median       2.5%     97.5%
meanlog 11.030514 10.9674162 11.084363
sdlog    0.967039  0.9246484  1.010675
```

This tells us, by bootstrapping, we can have 95% confidence that the meanlog parameter of the lognormal distribution for this dataset is between 10.967 and 11.084. This is a relatively good result – and it is further indicated by later results in looking at the distributions.

It should be noted that there are several options available to determine the best parameters using *fitdistrplus*, including moment matching, quantile matching and maximum goodness-of-fit estimation. For the purposes of this paper, maximum likelihood was the preferred method, but it is worthwhile to know other such methods are available to be used.

VII) Shape of the Empirical Data

To get an idea of how the distributions of losses look, I have produced a histogram of the losses below:



As you can see, the data is heavily skewed to the right. Most of the observed losses are in the first few bins, with a few sporadically at the very far right of the curve. This is more skewed than any common distribution, and so it makes splitting the distribution seem like an attractive alternative to try to model the tail of the distribution separately from the center. .

VIII) Determining the Best Split by Cross Validation

Since we are considering a rather strange distribution in **(b)**, it is not clear on how to determine the best split. Since the threshold itself is also related to α in that the $\alpha * 100$-th percentile of the data is the threshold, attempting to take a derivative here with respect to the likelihood function would be time-intensive and not very useful – indeed, intuition tells us that an exact maximum here is not truly the goal, but rather it would be more useful to simply determine a ballpark estimate of a good cutting point. Also, since this function is a piecewise function, it will likely be impossible to carry on in this way. For instance, the difference between the 90^{th} and 85^{th} percentile may be huge, but not so much between the 90^{th} and 90.5^{th} percentile.

Luckily, we have algorithmic approaches to finding the optimal point at which to split the dataset in order to achieve the maximum likelihood for the data.

My initial approach was simply to find the likelihoods of each observation based on whether it is above or below the thresholds for minute changes in the threshold. The intention here was to determine where it was that the likelihoods peaked, and therefore where it was most optimal to make a split. However, this approach had several problems – since the vast majority of the data was below the threshold, we would disproportionally have smaller increases in likelihood for those below than those above. This made the results incomparable; the larger the α in this method, the higher the likelihood.

The approach used in this project was iterated cross-validation. I proceeded as follows:

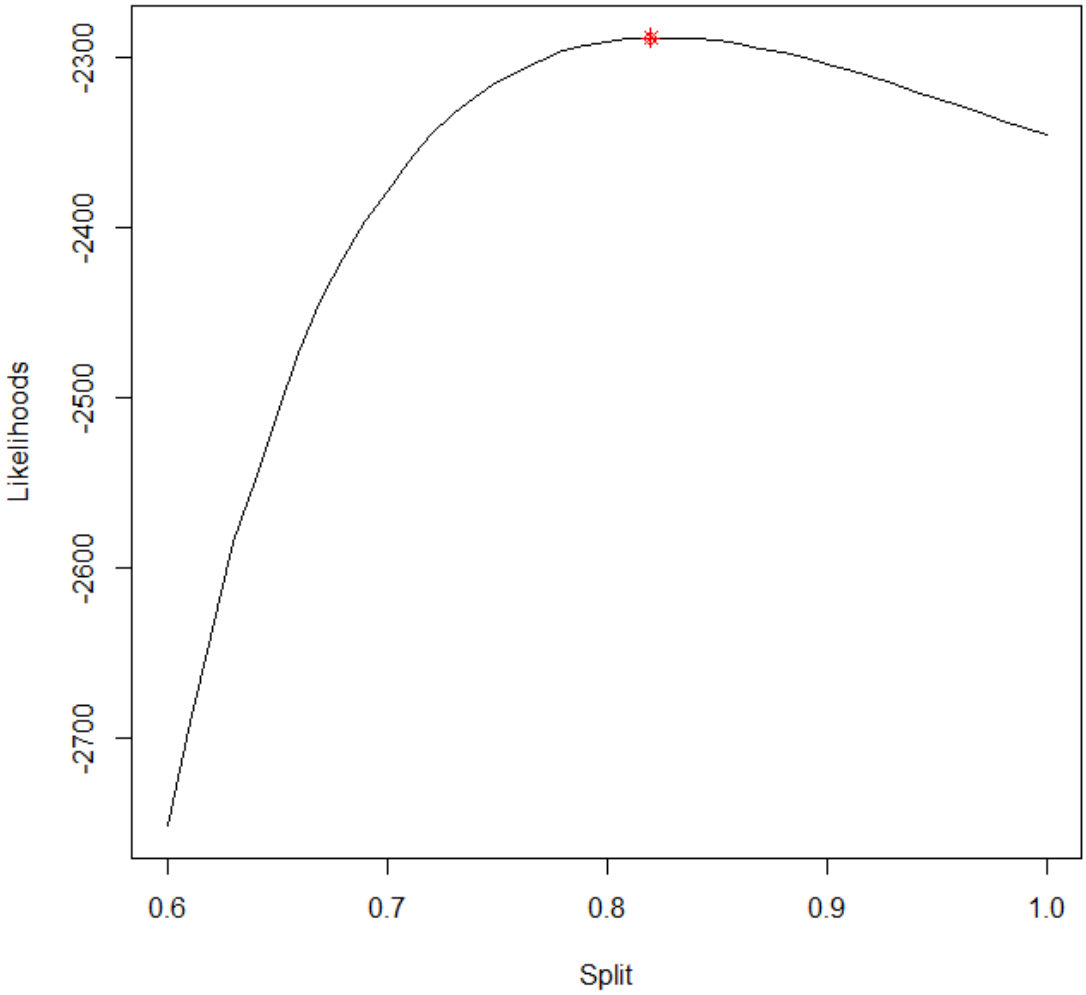
Split the data randomly in half into Test and Training sets. Then, sample a certain number of observations from the test set (in this case I sampled 100). For each threshold at percentiles from .6 to 1 in increments of .005, first fit the appropriate distribution to the observations above and below the threshold by using the training set. Then, calculate the likelihood of each test observation based on which side of the threshold it belongs.

To be exact, if the test observation is greater than the threshold, use the best fitting Pareto distribution based on the training set to calculate the likelihood. Else, if it is below the threshold, use the best fitting distribution of choice (Lognormal or Weibull in this case) on the training set to calculate the likelihood.

We then sum these likelihoods together and record the total likelihood for each split in the distribution. This will give us a total likelihood on the testing observations for each split of the randomly selected test data. We then reiterate (for my purposes, I did it 100 times) and divide by the total number of iterations to obtain an average likelihood for each of the splits.

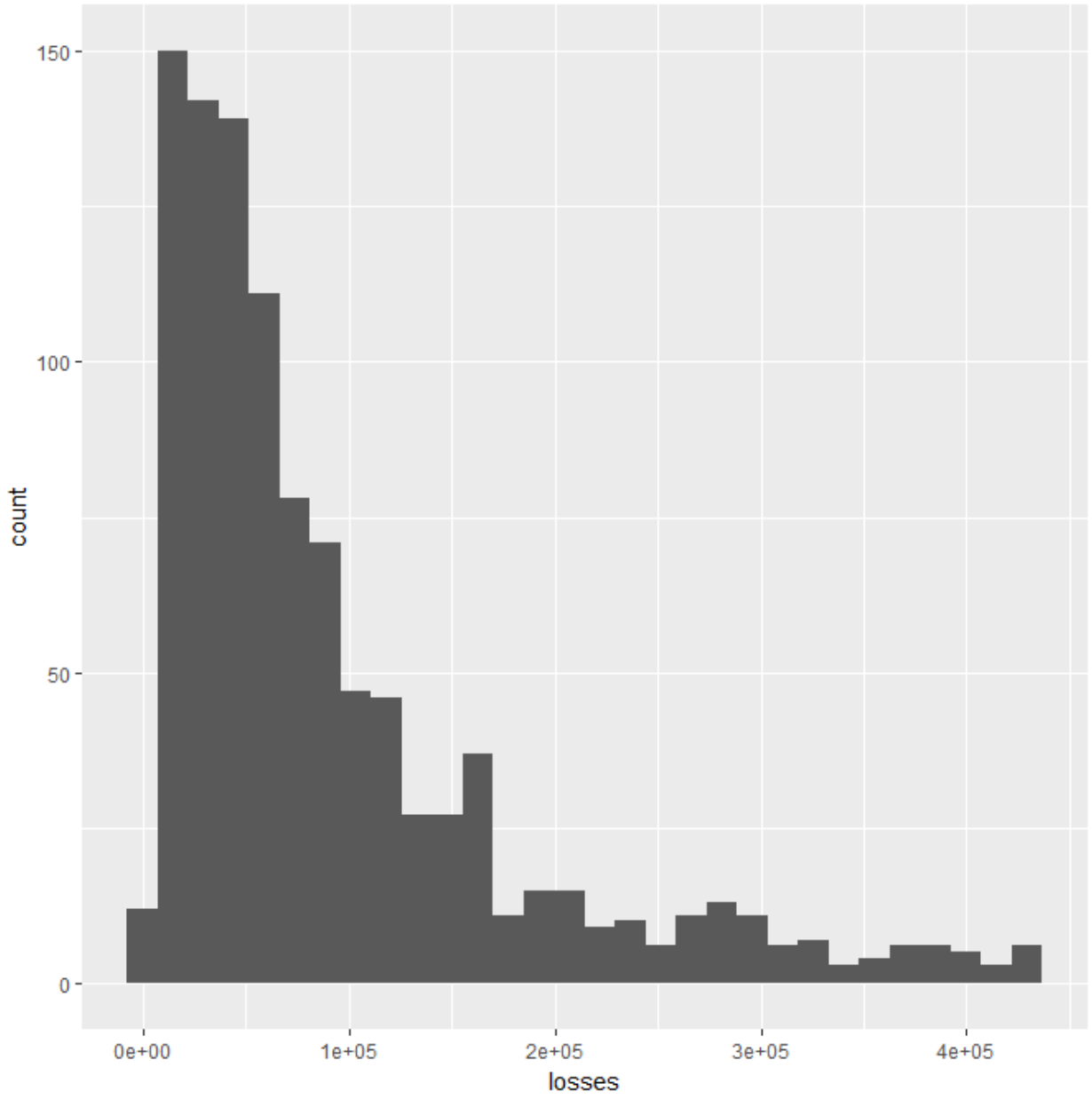
To get an idea of how this was coded, I have included a snippet from the R code which was used to determine the likelihoods. This was then iterated for numerous randomized test and training sets to make the final choice.

```
test_piecewise <- function(dist,Train,Test,low=.6,high=1,step=.01,samp=100){

  #!# Determines crossvalidation likelihoods based on a sample of the test set.

  samp <- sample(Test,samp,replace=FALSE)
  likl <- vector()
  b<-list()
  pars <- lapply(X= returnparam(dist,Train,low=low,high=high,step=step)[[1]],FUN=function(x) try(x$estimate))
  for(i in 1:(((high-low)/step)+1)){
    threshold <- fnd_Short_Range(Train,(high-(i-1)*step))[2]
    b[[i]]<- suppressWarnings(tryCatch(fitdistr((Train[Train>=threshold]) - threshold,dpareto,start=list(shape=.8,scale=max(Train)-threshold),control=list(maxit = 10000,reltol = 1e-9)),silent=TRUE,error = function(e) -Inf))
    b[is.na(b)] <- 0

    bool <- samp < threshold
    a <- samp[bool]
    d <- samp[!bool]
    if(!is.atomic(b[[i]])){

      params <- list()
      params[[1]] <- a
      for(j in 1:length(pars[[i]])){
        params[[1+j]] <- pars[[i]][j]
      }
      names(params) <- names(formals(paste("d",dist,sep="")))[1:(length(pars[[i]]) + 1)]

      lik <- try(sum(log(dpareto(d,shape = b[[i]]$estimate[1],scale = b[[i]]$estimate[2]))) + sum(log(do.call(paste("d",dist,sep=""),params))))
      likl[[i]]<- as.numeric(lik)
    }
  }
  likl
}
```



Above I have posted an example of a plot of these determined CV Splits and their likelihood for the simulated dataset. We can see that the highlighted point is the maximal point of the splits, so we select based on this. This was .83 in this case, which makes sense – we would expect the optimal split to be somewhere between 80-95%, which this graph supports. If it is too low, the Pareto distribution would likely be unable to be fit well to the tail end of the data, and if it's too high, there would be too little data to fit a distribution to the tail with much accuracy.

To illustrate graphically why this is a good idea and how this benefits our analysis, we can look at a histogram of the simulated dataset when this cross-validation technique is applied. Here, we can see the losses up to and including the 83rd percentile, the split point determined to be optimal by the previously mentioned technique.



While this distribution is still highly skewed, it no longer suffers from the extreme outliers of the full dataset (see above for reference). By splitting in this way, we can account for the extreme outliers separately, and this approach will give us the best possible split for a predetermined set of theoretical distributions for the belly and tail. This also indicates reasoning for using a long-tailed distribution for the belly of the distribution as well, as even with the large outliers removed the distribution remains skewed.

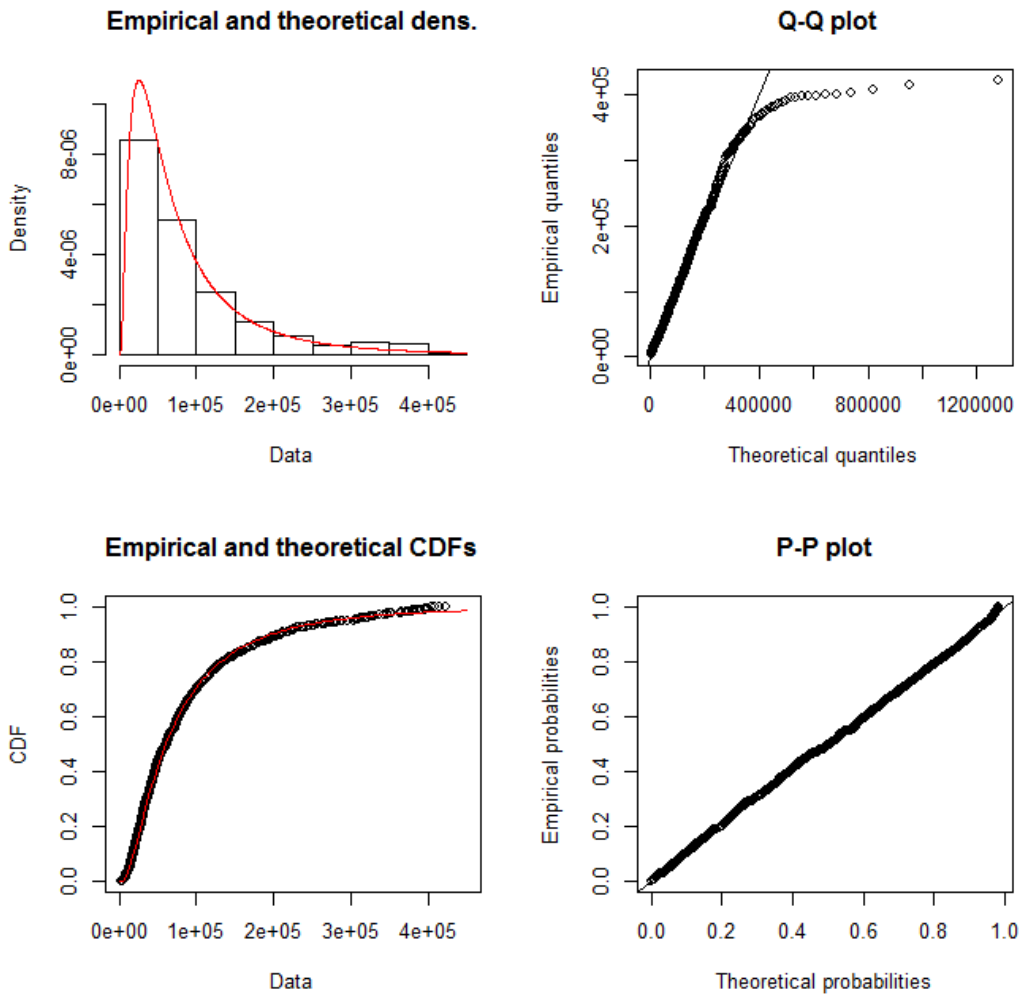
IX) Diagnostic Plots of the Fitted Distributions

The next step, after finding these distributions, is to see how well of a fit they are for the observed data. This is possible directly from the *fitdistrplus* package again, by calling plot on a *fitdistr* object. Doing this will plot a distribution function over a histogram, an empirical vs theoretical cumulative distribution function, a Q-Q plot and a P-P plot.

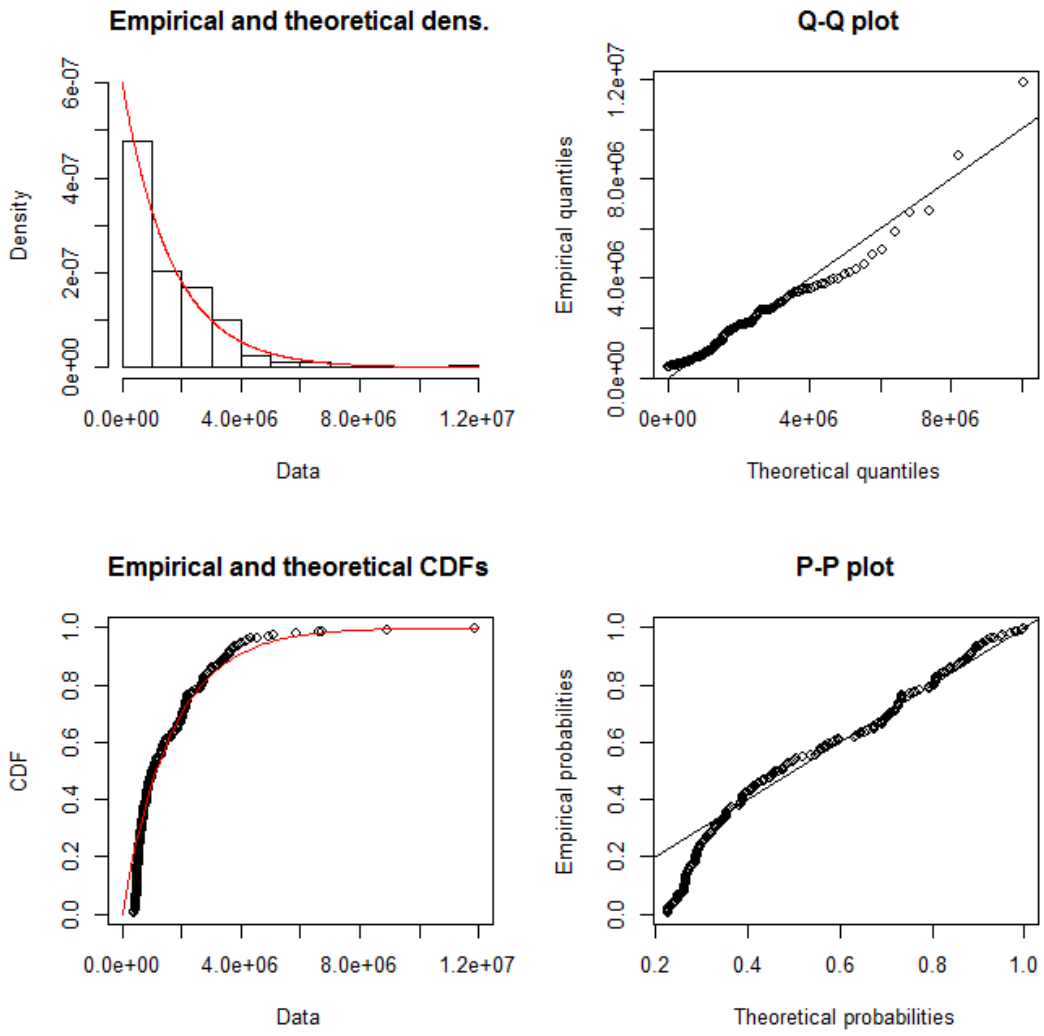
Although the likelihoods are useful in determining how the distributions fit relatively to one another, they cannot give us an objective idea of how they perform overall. This is especially true if there are parts of the curve you are more concerned with being accurate than others – even if the likelihood is greater in one curve compared to another, for the purposes you are using them for, the other curve may still be more appropriate, if not as well of a fit.

I have produced plots of the lognormal fit on the simulated dataset that we have been using below to give an idea of how these plots look.

First, for the belly of the distribution:



For the tail:



As this dataset has been simulated from this very distribution, it expectantly follows the density function and CDF very closely. In my experience with true loss data, there was of course larger discrepancies, but far less than any one single distribution alone. Both the cdf and pdf appeared to be well-fitting in most cases. The diagnostics here are derived from those of Nolan (2001).

The Q-Q Plots tended to be the most off, while the P-P plots followed quite closely in most situations. This indicates that the centers of the distributions were good fits (which a P-P plot is an indicator of) while the tails (particularly the right tails) were poor fits. Luckily, in these cases the theoretical distribution tended to be an underestimation, which leads to a more conservative estimate, which is preferable in this case. Also, by having a methodology where we split the tail and belly into separate parts, this is less of an issue.

By looking at these plots, the theoretical distributions appeared to fit surprisingly well to both ends of the distribution, which is promising.

X) Dealing with Censored Data

So far in the discussion of losses, we have considered them to be independent of the limits themselves. An issue arises immediately in trying to fit a distribution to this historical data: what if the actual loss amount was far larger than the loss amount recorded? That is, if there was a $10M limit, and the recorded loss is $10M, it is very likely that the actual loss amount, independent of the limit, was far larger. The insurance company is not informed of any losses larger than those which they are liable for. In this way, our data is censored, and this can have a significant impact on how well our fitted theoretical distributions fit with the true distributions of losses.

This can be of great detriment to our calculations if it is not accounted for, especially if there are a good number of these instances where the data is censored. Consider if there are some recorded losses where the loss is censored by the limits. That is, for each case, the true loss is larger than the observed loss. It follows that the expectation of a distribution which is fit to the censored data will be lower than the true expectation. This is not a good thing for the purposes of determining an ILF curve because if the expectation is an underestimate, our curve will be too flat.

In this project I have considered two approaches to this problem: dividing the losses into those which are close to the limit and far from the limit, and using the *fitdistcens* function in the *fitdistrplus* package. I will describe both methods briefly below.

First Approach: Separating the Data by Percent of Limit

One way of dealing with this problem would be to make the distributions themselves dependent on the limit. More precisely, on the percentage of the limit on which the loss accumulated. We can select some threshold β, where $0<\beta<1$, and if the loss-to-limit ratio is above this threshold it will be considered separately from those which are below it.

One would hope that when β is set to a high enough proportion, the distribution of losses with loss-to-limit ratios above β would have some relatively simple distribution. Then, we wouldn't have to even consider whether the loss amounts are greater than the limit – for any given limit, we would have some idea of where the loss will lie within the limit. We would be shifting our lens from the idea of all losses (which are censored) to those which are of concern to us (the uncensored, observed data). On the surface, this would appear to be one possible solution to this problem.

In practice, this solution seemed to be a bit of a dead-end for two reasons: the distribution of losses above the loss-to-limit ratio was not simple and highly skewed, and there was no simple way of adjusting this ratio to make it behave in this manner. Because of this, any more complex distribution would require more data, and thus this exercise had defeated the purpose. The second, and larger, problem was that the data had already been split twice – once for the limits themselves, and once for the thresholds in the mixed distribution. Splitting again based on the loss-to-limit ratio created datasets which were far too sparse – indeed, for some of the business units there were no losses which would accurately fit into.

This approach would work well if there was a much larger body of data to work with, but unfortunately most insurance companies on their own do not have enough data to use this approach. Similar to finding distributions for losses at each individual limit, this solution falls short when trying to determine a reasonable distribution with sparse data.

Second Approach: Using the fitdistcens Function in fitdistr

Luckily, the *fitdistplus* package has an alternative solution to this problem. The *fitdistcens* function is quite flexible and will allow one to maximize the likelihood based on a different approach for censored and uncensored observations. One can have left, right, or interval censored data. In this case, we have right censored data.

As is described in the vignette of the package, one can feed the data into the *fitdistcens* function as a two column data frame, with any censored observations denoted with an NA. Since we are dealing with right censored data, we will replace the right column with NA when the loss amount was at (or above) the limit. This data will be considered censored. For the uncensored data below the limit, we will simply duplicate the loss amount in both columns.

The algorithm here now simply attempts to maximize the following formula, as per the vignette:

$$L(\theta) = \prod_{x \notin C} f(x|\theta) \times \prod_{x \in C} 1 - F(x|\theta)$$

Here, C is the set of (right-) censored observations. The algorithm then attempts to maximize this function numerically, giving us a result which considers the censored observations. That is, for censored observations, we consider the survival function rather than the density for the censored observations, and take the product. We then would proceed as usual in likelihood estimation, and attempt to maximize the above function.

Luckily, the package takes care of it for us, and we are provided with a new distribution which will give us a more conservative estimate of the parameters. For simplification, in this project we have used the splits from the uncensored distribution as the splits in the new, censored ones. This is for convenience, considering it is not too likely they will be all that different. It is worth noting, however, that one could potentially do the same process as we did above with cross-validation with the *fitdistcens* function (as it provides likelihoods as well) to get a more exact estimate of the optimal point at which to split the distribution.

By using this package to deal with the censored data, we get a more conservative estimate of the cdf, and therefore we will have larger expected losses with the new curve.

Care must be taken in considering whether it is better to treat the data as though it were censored or uncensored, however. In this project, some of the datasets/subgroups within them had very little (or no) censored observations. While the *fitdistcens* function attempts to mitigate the inaccuracies caused by potential censoring, the statistical inferences which can be drawn from the

above "likelihood" function are murky. For instance, there is no way to calculate the variance of the parameters in this case, nor is there a way to straightforwardly plot the empirical distribution and compare it to the theoretical density. These things should be – and were – taken into consideration when considering whether or not to use the censored data approach in the first place. Some other considerations for censored data are explored in Turnbull (1976), which are reflected in the *fitdistcens* functions' estimations of the empirical cdf curve.

There are some ways to circumvent the lack of a measure of reliability on the estimates, which is by using the *bootdistcens* from the *fitdistr* package, which can produce similar results to those seen earlier but for censored data.

XI) Solving for Expectation:

Now that we have found a best fitting distribution through maximum likelihood, we would like to solve for the following expectation:

(a) $$E(S|L) = E(X|0 \le X \le L) + L\bar{F}(L)$$

where X has the pdf and cdf as described in **(b)** and **(c)**. By using the cross-validation approach described earlier, we were able to determine an appropriate split, α, and threshold T (the 100α-th percentile of the empirical data) with which we will be able to solve this expectation. From this information we can determine the best fitting parameters for the distributions by maximum likelihood. Now, it is simply a matter of how to get the expectation from this information.

Since both T and L are known, we can use this to determine the expectations. By the law of total expectation:

(d) $$E(X|0 \le X \le L) = E(X|0 \le X \le L, 0 \le X \le T)P(0 \le X \le T) + E(X|0 \le X \le L, X > T)P(X > T)$$

This holds because the events $0 \le X \le T$ and $X > T$ are disjoint and complete partitions of the sample space.

Using the determined splits, we will assume that these probabilities are exactly equal to the splits we've determined, that is $P(0 \le X \le T) = \alpha$, and so $P(X > T) = 1 - \alpha$. These probabilities exist by construction – it assumes that T is the *true* αth percentile of the population. Some interesting results may be found by attempting to estimate this population percentile directly, if one wanted to expand on this rather simple methodology. In this case, we will assume that this holds.

For the expectations, there may (or may not) be closed form expressions of this conditional expectation. It is therefore useful to get this into a form that can be easily solved by numerical integration techniques.

We can split **(d)** up into two scenarios.

If T<L, then:

$$E(X|0 \le X \le L, 0 \le X \le T) = E(X|0 \le X \le T) \text{ and}$$
$$E(X|0 \le X \le L, X > T) = E(X|T \le X \le L)$$

So

$$E(X|0 \le X \le L, T < L) = \alpha E(X|0 \le X \le T) + (1 - \alpha)E(X|T \le X \le L)$$

If T>L, then

$$E(X|0 \le X \le L, X > T) = 0 \text{ and}$$

$$E(X|0 \le X \le L, 0 \le X \le T) = E(X|0 \le X \le L)$$

So

$$E(X|0 \le X \le L, T > L) = \alpha E(X|0 \le X \le T)$$

Now, looking at the survival distribution, $\bar{F}(L|T) = 1 - F(L|T)$,

(e)
$$\bar{F}(L|T) = \begin{cases} 1 - \alpha\frac{B(L)}{B(T)}, \text{T} > \text{L} \\ 1 - \alpha - \frac{(1-\alpha)(Z(L)-Z(T))}{1-Z(T)}, \text{T} < \text{L} \end{cases}$$

So we can now determine the full expectation in **(a)** in pieces.

If T<L:

$$E(S|L, \text{T} < \text{L}) = E(X|0 \le x \le L, \text{T} < \text{L}) + L\bar{F}(L|\text{T} < \text{L})$$

(f) $= \alpha E(X|0 \le X \le T) + (1 - \alpha)E(X|T \le X \le L) + L\left(1 - \alpha - \frac{(1-\alpha)(Z(L)-Z(T))}{1-Z(T)}\right)$

If T>L:

$$E(S|L, T > L) = E(X|0 \le X \le L, T > L) + L\bar{F}(L|T > L)$$

(g) $= \alpha E(X|0 \le X \le L) + L(1 - \alpha\frac{B(L)}{B(T)})$

Note that when L = T, both expectations give the expected result $\alpha E(X|0 \le X \le T) + L(1 - \alpha)$. This means there will be no jumps, and our expectation is continuous, as $\lim_{L \to T+} E(S|L,T) = \lim_{L \to T-} E(S|L,T) = \alpha E(X|0 \le X \le T) + L(1 - \alpha)$.

Similarly, since we provide L and T as constants, there is no need to consider probabilities in the total expectation – therefore, we can define an indicator function:

$$= I(L,T) = \begin{cases} 1, L < T \\ 0, L > T \end{cases}$$

Using this indicator function, we can define the expectation:

$$E(S|L,T) = I(L,T)\,E(S|L, L < T) + \big(1 - I(L,T)\big)\,E(S|L, L > T)$$

where $E(S|L, L < T)$ and $E(S|L, L > T)$ are defined as in points **(f)** and **(g).**

XII) Determining the Expectations of Truncated Distributions

The above expectations $E(S|L, L > T)$ and $E(S|L, L < T)$ contain distributions which are truncated. If *f(x)* is the probability density function of our losses, we can use Bayes' rule:

$$f(x|a \leq X \leq b) = \frac{P(a \leq X \leq b|X = x)f(x)}{P(a \leq X \leq b)}$$

Note that $P(a \leq X \leq b|X = x) = 1$ if the instance of X, *x*, is within the limit and zero elsewhere. So this truncated density is non-zero only within the bounds. Since we are treating $P(0 \leq X \leq T) = \alpha$ we can use this for values of the cdf. We can then rewrite this truncated density as:

$$f(x|a \leq X \leq b) = \frac{f(x)}{F(b) - F(a)}, x \in [a, b]$$

where *F(x)* is the cumulative distribution function of *X* as described in point **(c)**.

From here we can solve for the expectation of a random variable conditioned on its values being truncated. The expectation of this truncated variable is then:

$$E(X|a \leq X \leq b) = \frac{\int_a^b xf(x)dx}{F(b) - F(a)}$$

Although this is a relatively simple result, the integral in the numerator above may not be simple to solve analytically. For this reason, we can turn to numerical approximations to determine the integral in the numerator.

The expectations we would wish to solve are the following:

$$E(X|0 \leq X \leq L) = \frac{\int_0^L xf(x)dx}{F(L)}$$

$$E(X|T \leq X \leq L) = \frac{\int_T^L xf(x)dx}{F(L) - \alpha}$$

$$E(X|0 \leq X \leq T) = \frac{\int_0^T xf(x)dx}{\alpha}$$

for a given L and T, where *F* is the cdf given in **(c).**

It particular, for use in equations in points **(f)** and **(g)**, we are interested in:

$$\alpha E(X|0 \le X \le T) = \int_0^T xf(x)dx$$

$$(1-\alpha)E(X|T \le X \le L) = \frac{1-\alpha}{F(L)-\alpha}\int_T^L xf(x)dx$$

$$\alpha E(X|0 \le X \le L) = \frac{\alpha}{F(L)}\int_0^L xf(x)dx$$

Notice that these equations have multiplicative constants. When $\frac{\alpha}{F(L)} > 1, \alpha > F(L)$ so $F(T) > F(L)$ and, since a cdf is a monotonically increasing function, T>L. Also, since $1-\alpha > 0, \frac{1-\alpha}{F(L)-\alpha} < 0 \; iff \; F(L) - \alpha < 0$ or $F(L) < F(T)$ also when T>L. From this construction we can then define two variables:

$$\varphi(L,\alpha) = \max\left(1, \frac{\alpha}{F(L)}\right)$$

$$\gamma(L,\alpha) = \max\left(0, \frac{1-\alpha}{F(L)-\alpha}\right) \; if \; F(L) \neq \alpha$$

These will work as a more expansive substitute of our above indicator function.

Then we can define our expectation as:

(h) $$E(S|L,T) = \varphi(L,\alpha)\int_0^{\min(L,T)} xf(x)dx + \gamma(L,\alpha)\int_T^{\mathrm{L}} xf(x)dx \; + \; \mathrm{L} \cdot \bar{\mathrm{F}}(\mathrm{L})$$

It should be noted that in some distributions, there is a closed form for the truncated expectation, and therefore for the integral listed above. In these cases, it may be worthwhile to simply use the closed form formula, as it will be (marginally) more accurate and will not have issues close to the threshold as the numerical integration can have. For small differences between T and L, γ may "blow up" quicker than the integral diminishes. Theoretically, it would still approach the same values as T approaches L. Although this may seem to be an issue, since the limits we are interested in are not very close to one another numerically it does not make much of a difference.

XIV) R Code Implementation

Here, I will describe how to find the previous expectation using R. Once we have our two density functions and our threshold/split, we can calculate this directly. I have entered the coding for the implementation function below:

```
implement <- function(limit, threshold, split,p1,p2,p3,p4,dist = "lnorm"){
  cdf <- splitcdf(limit,threshold, split,p1,p2,p3,p4,dist)
  integrand <- function(x) x * splitpdf(x,threshold,split,p1,p2,p3,p4,dist)

  phi <- max(1, split/cdf)
  gam <- max(0, (1-split)/ (cdf - split))

  mn <- min(limit, threshold)

  a <- integrate(integrand, lower = 0, upper = mn)$value
  b <- integrate(integrand, lower = threshold, upper = limit)$value
  c <- limit * (1-cdf)

  res <- phi*a + gam*b + c
  return(res)
}
```

Not included in this code is the user created *splitcdf* and *splitcdf* functions that pertain to the pdf and cdf listed earlier at points **(b)** and **(c)**.

This code uses the integrate function to find a numerical approximation of the aforementioned integral for each of the distributions. It then estimates the formula in point **(h)**. Implementing this using the *phi* and *gam* functions makes this much easier to deal with and less error prone than using conditions.

After we have these expectations, determining the Increased Limit Factors (ILFs) at various limits is simple. Our ILF at a selected limit would then be:

$$ILF(L_s) = \frac{E(S|L_s)}{E(S|L_B)}$$

Where L_s is the selected limit and L_B is the limit at the base. For the purposes of this report, I have selected $L_B = \$1{,}000{,}000$ and the expectation by **(h)** in increments of 500,000 from 500,000 to 20,000,000. These numbers can be changed to better fit one's circumstances.

XV) Example on Simulated Data

The final results of this analysis will be shown through results which were based on a simulated dataset.

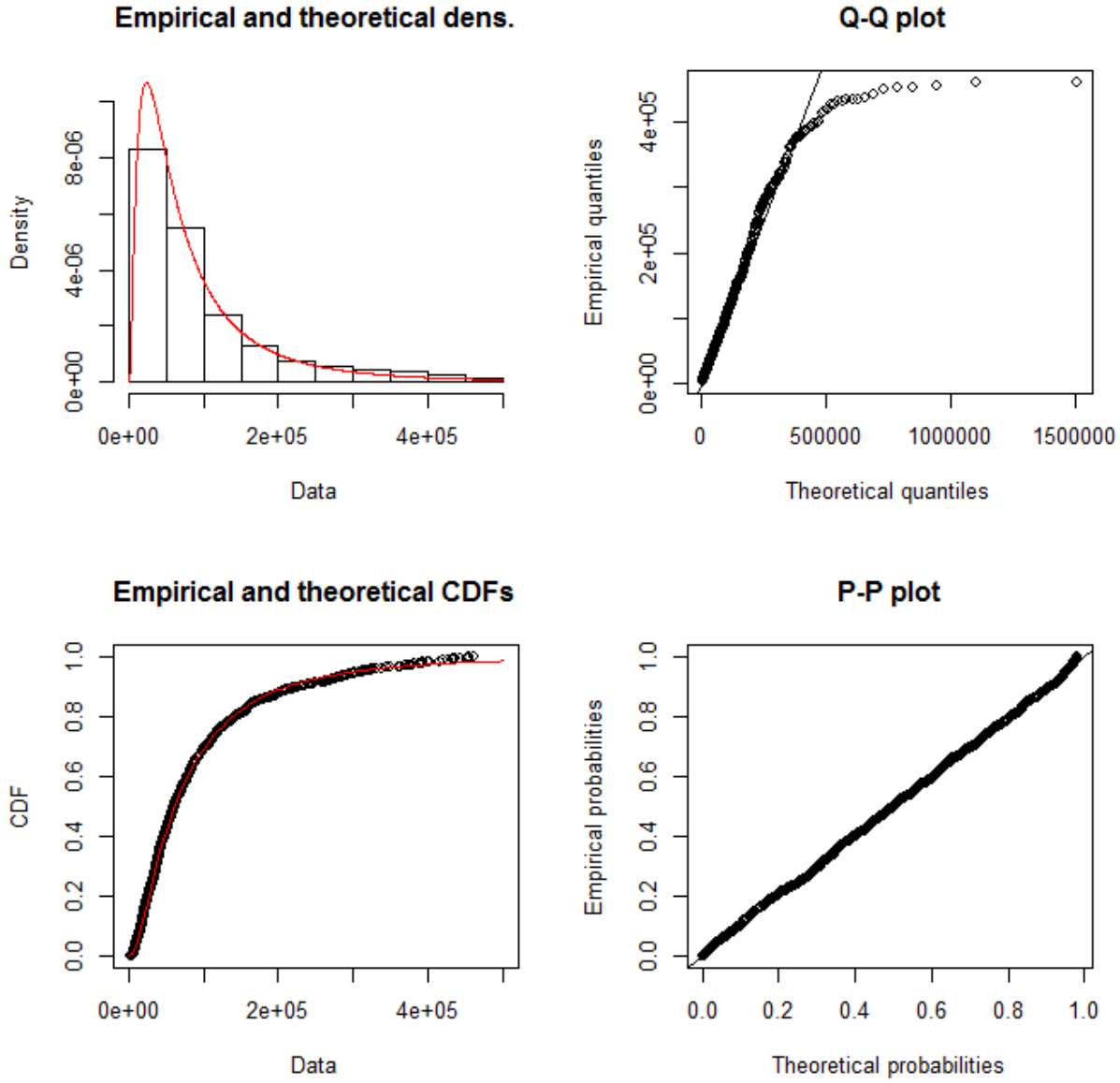
The following R code output describes the lognormal and Weibull fits which were selected based on the methodology outlined in this paper:

```
> simexLN

$dist
$dist$belly
Fitting of the distribution ' lnorm ' by maximum likelihood
Parameters:
          estimate Std. Error
meanlog 11.0270618 0.02996694
sdlog    0.9691885 0.02118973

$dist$tail
Fitting of the distribution ' pareto ' by maximum likelihood
Parameters:
          estimate Std. Error
shape 1.551506e+06         NA
scale 2.862260e+12         NA


$spl
[1] 0.84

$max_lik
[1] -2287.974

> simexW

$dist
$dist$belly
Fitting of the distribution ' weibull ' by maximum likelihood
Parameters:
          estimate Std. Error
shape     1.172093          0
scale 92927.436553        NaN

$dist$tail
Fitting of the distribution ' pareto ' by maximum likelihood
Parameters:
          estimate Std. Error
shape 1.284896e+05         NA
scale 2.178295e+11         NA


$spl
[1] 0.82

$max_lik
[1] -1764.501
```

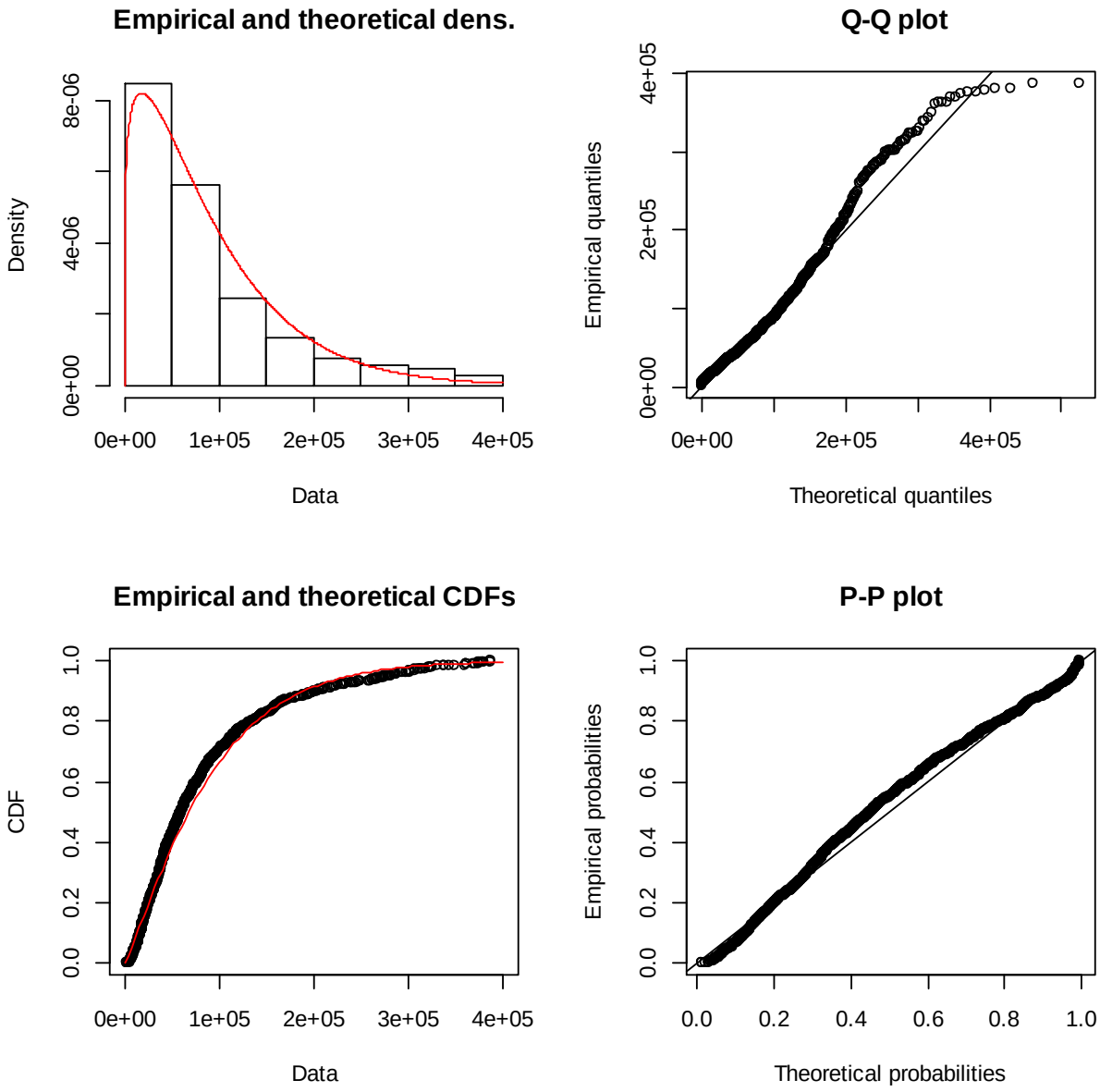
Thus, we have our four parameters for each distribution and our selected splits. In this case, it appears that by the maximum likelihoods, the Weibull-Pareto distribution is more appropriate than the LogNormal-Pareto.

To get a visual representation of how good of a fit it is in each segment, we can use the *plot* method of the *fitdistrplus* packages' *fitdist* object.

```
> plot(simexLN$dist$belly)
```



```
> plot(simexW$dist$belly)
```



It does appear from these graphs that the Weibull-Pareto is the most appropriate choice. One could also plot the tails to see how well they fit the end of the distribution. This has been omitted here for brevity.

Directly from here, we can implement our procedure for the expectation. As an example, let's use 1,000,000 (our base limit) as the limit we want the expectation of. Then, evaluating the formula in **(h)**:

```
> findExpFit(1000000, simexW)
```

```
317637.4
```

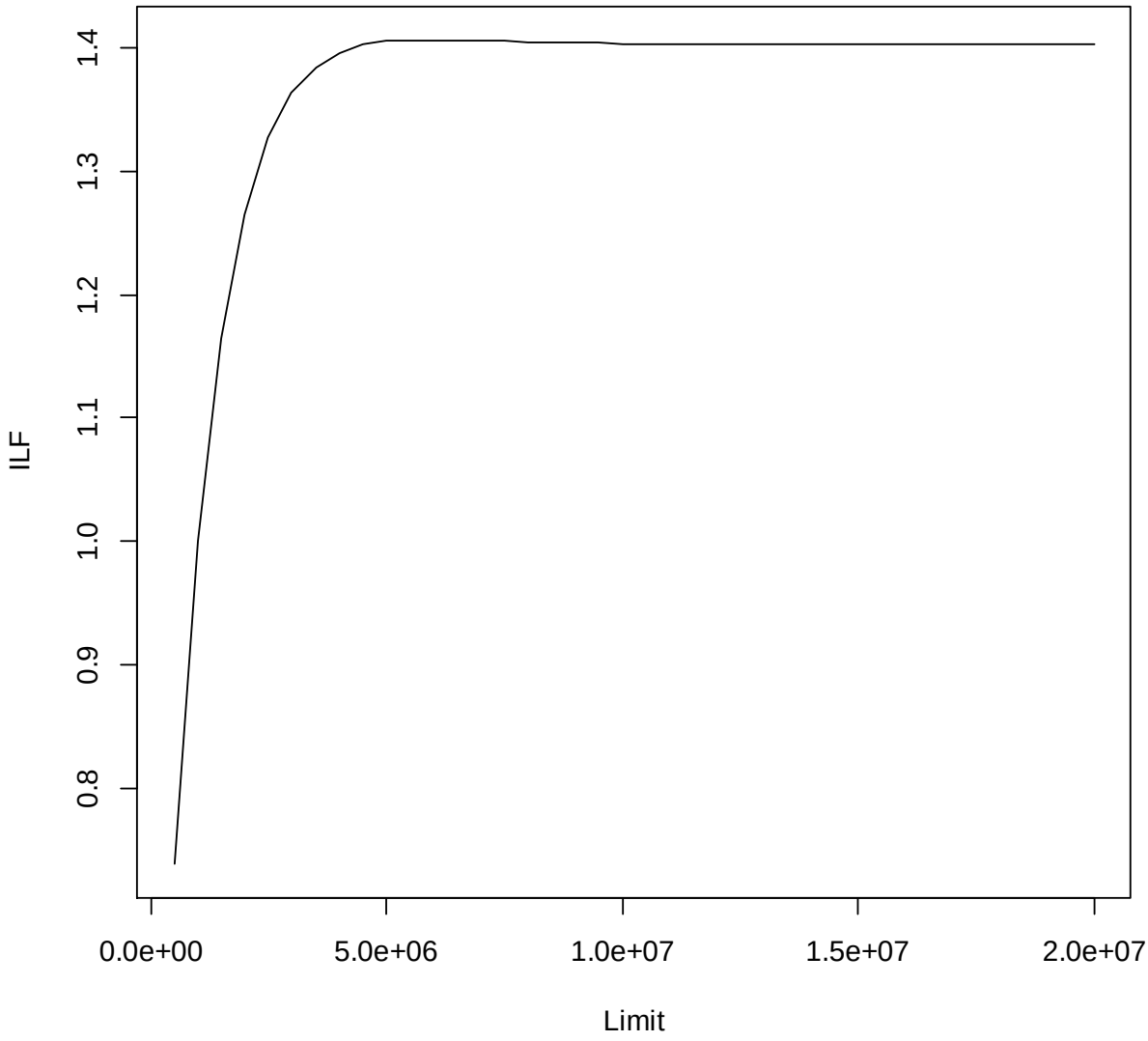
That is, we expect a loss of $317,637 on a limit of $1,000,000. This seems like a reasonable assertion.

To find appropriate ILFs, we can take the expectation at specific values of the limit. For this example, I used evenly spaced increments from $500,000 to $20,000,000, but this was a somewhat arbitrary selection. Different companies and circumstances may require a different base premium or a different selection of limits. Typically, a base limit should be a limit where there is a sizable number of claims, as it will be used as a reference limit.

The results of this simulated dataset are output below:

```
> findExpFit(1000000, simexW)
```

	Limit	ILF
1	500000	0.7387231
2	1000000	1.0000000
3	1500000	1.1640052
4	2000000	1.2653577
5	2500000	1.3268526
6	3000000	1.3633394
7	3500000	1.3843790
8	4000000	1.3960491
9	4500000	1.4021592
10	5000000	1.4050599
11	5500000	1.4061752
12	6000000	1.4063499
13	6500000	1.4060715
14	7000000	1.4056119
15	7500000	1.4051147
16	8000000	1.4046502
17	8500000	1.4042470
18	9000000	1.4039124
19	9500000	1.4036425
20	10000000	1.4034293
21	10500000	1.4032631
22	11000000	1.4031351
23	11500000	1.4030372
24	12000000	1.4029628
25	12500000	1.4029066
26	13000000	1.4028642
27	13500000	1.4028323
28	14000000	1.4028084
29	14500000	1.4027905
30	15000000	1.4027771
31	15500000	1.4027671
32	16000000	1.4027596
33	16500000	1.4027540
34	17000000	1.4027499
35	17500000	1.4027468
36	18000000	1.4027445
37	18500000	1.4027427
38	19000000	1.4027414
39	19500000	1.4027405
40	20000000	1.4027398



There are a few things to notice here about these results extracted from the simulated dataset. These ILFs are very small in size – in practice, one would expect far higher premiums at a limit of $20,000,000 then at $10,000,000. This seems to indicate they should be nearly the same. Another glaring problem is the fact that these ILFs are not steadily increasing, they decrease at some level of limit size! This is against our notions that the ILF curve should be monotonically increasing and should not be this flat. The reasons for this will be explored in the next section.

XVI) Applied Results

Unfortunately, the case of the simulated dataset was very similar to the results of the actual datasets. In many cases, the ILF curve was far too flat and non-monotonically increasing. There are two indicators as to why this may be the case.

In the datasets I was using, losses were typically of small size and did not represent true losses very well. That is, there were many small losses, and a few very large losses, but nothing that even approached $20M. It is because of this that the curves are so flat – especially in the case of Axis Pro, which the simulated dataset was based on, there simply weren't enough large losses to model this correctly. It would seem that, according to our own historical loss history, we should not even consider the $20M risk to be any riskier than the $10 M risk, because according to the model the probability of the loss being between these two values is very low.

That is, there was a scaling issue when looking at the limits as compared to the losses. Many limits were $10M+, but losses stayed very low relative to this. Attempting to model the losses individually, even using the techniques for censored distributions as described above, was not enough in this case to get a reliable sense of where the true losses distribution lies. The expectations, then, were far too small in the upper ranges.

Since this is a parametric model, it will require quite a large number of losses to be accurate. This is especially true because of the probabilities of catastrophic events. Although using the Pareto in the tail was used as a way to circumvent this issue, ultimately the lack of data (and especially the lack of large losses) proved detrimental to getting accurate results.

Because of this issue, many companies do not rely on their own historical records to estimate ILFs, because the amount of data needed to estimate the curve accurately would be unsurmountable. With the exception of very large companies, most insurance companies rely on industry standards to determine these numbers.

One could consider that the ILFs should be some weighted average of the ILFs determined above based on historical data and on an industry standard. This depends on the outlook of the final result – if the ILFs based on industry standards were already too flat as it is, and the ILF determined by this procedure is even flatter, it will have an undesired effect. The selection of the weights is, too, arbitrary, and would indicate how confident one was in the historical records. Still, this may be something to consider in the case that the same issue occurs elsewhere.

What matters most is not necessarily the pure number of data points one has, but the variety in the data points available. When historical losses do not even approach limits where one is trying to determine the factor, it can be quite difficult to model. Indeed, in this project only 2%, 3% and 4% of the losses were full-limit losses, with a clear majority of them being less than 50% of the limit. This is a desirable case for the insurance company, but disadvantageous for this analysis.

Although distributions like the log-normal, Weibull and Pareto have heavy-tails, as described earlier, the probability of these large losses occurring based on maximum likelihood of previous observations is still very small. Without adequate support from the data, the truncated expectations will be very similar the further out into the tail you are.

While theoretically this approach will give reliable estimates of these truncated expectation, in practice one does not have enough data to model the loss amounts directly. I would suggest a weighted average approach for situations where there is a moderate amount of data. The more historical data one has, the more they can rely on the estimates provided by this approach.

As is clearly illustrated, with such problems at the primary coverage level (for AxisPro, CMS and FI), the sub-coverages are even less reliable at determining expectations! Although I did produce results for these as well, the result was unfulfilling due to a lack of data.

Another consideration is the multiplicative constants in **(h).** While these do hold true based on the previous analysis, the constant γ will decay as L approaches infinity. If the value of the integral attached to this is nearly constant, it can account for this strange behavior. In practical situations, once this occurs for some limit *L,* all limits above the maximal value should be held constant. These differences are likely due to the numerical approximations of the integral involved here.

XVII) Conclusion

In determining increased limit factors (ILFs), parametric approaches to approximating loss distributions with theoretical distributions are commonly used. This project examines a particular solution to this problem which accounts for a different distribution in the "belly" and "tail" of the distribution. This is to account for the large tails common in insurance loss data. To accomplish this, I have created a piecewise function and used various distributions that were sensible for the belly and tail. This approach can be generalized to an arbitrary number of splits, and an arbitrary selection of appropriate distributions. The R programming language will allow this implementation to be relatively simple.

In order to determine an appropriate split, a cross-validation approach was used. Using the *fitdistrplus* package, I was able to maximize the likelihood function for right censored data, which is useful when working with historical loss data. There are considerations as to whether this approach is necessary, and these are explored. After determining an appropriate threshold and split, I calculated expectations for various limit sizes based on the distribution. I developed the motivation and calculation of these expectations. In **(h)**, I provide a formula which is easy to compute using appropriate software which does not require a conditional function, which was useful for practical implementation of this methodology.

While the results of this project were unsuccessful at determining an accurate ILF distribution, due to a low volume of loss data to base this on, the methodology could be applied in various ways. One could consider a Bayesian approach, using industry ILFs as a prior and updating this prior with the new piecewise distribution function. One could also consider a simple weighted average of industry and historically-based data dependent on credibility. Since the determination of an appropriate distribution function can be useful for various actuarial studies, using this approach may be a more robust and efficient way of estimating the tails of the distribution than using one overarching distribution function. There is a tradeoff in that this approach will, in general, require a larger pool of data than simply fitting a common, theoretical distribution.

References

Arnold, B. C., Balakrishnan, N., & Nagaraja, H. N. (1998). Wiley Series in Probability and Statistics.

Asmussen, S. (2003). Applied probability and queues: Stochastic modelling and applied probability. *Applications of Mathematics (New York)*, *51*.

C. Dutang, V. Goulet and M. Pigeon (2008). actuar: An R Package for Actuarial Science. Journal of Statistical Software, vol. 25, no. 7, 1-37. URL http://www.jstatsoft.org/v25/i07

Clementi, F., & Gallegati, M. (2005). Pareto's law of income distribution: Evidence for Germany, the United Kingdom, and the United States. In *Econophysics of wealth distributions* (pp. 3-14). Springer Milan.

Embrechts, P., Klüppelberg, C., & Mikosch, T. (2013). *Modelling extremal events: for insurance and finance* (Vol. 33). Springer Science & Business Media.

Goovaerts, M. J., Kaas, R., Laeven, R. J., Tang, Q., & Vernic, R. (2005). The tail probability of discounted sums of Pareto-like losses in insurance. *Scandinavian Actuarial Journal*, *2005*(6), 446-461.

Guillen, M., Prieto, F., & Sarabia, J. M. (2011). Modelling losses and locating the tail with the Pareto Positive Stable distribution. *Insurance: Mathematics and Economics*, *49*(3), 454-461.

Halliwell, L. J. (2012). The mathematics of excess losses. *Variance*, *6*, 32-47.

Kohavi, R. (1995, August). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*(Vol. 14, No. 2, pp. 1137-1145).

Lee, Y. S. (1988). The mathematics of excess of loss coverages and retrospective rating-a graphical approach. *PCAS LXXV*, *49*.

Marie Laure Delignette-Muller, Christophe Dutang (2015). fitdistrplus: An R Package for Fitting Distributions. Journal of Statistical Software, 64(4), 1-34. URL http://www.jstatsoft.org/v64/i04/.

Nolan, J. P. (2001). Maximum likelihood estimation and diagnostics for stable distributions. *Lévy processes: theory and applications*, 379-400.

Pareto, V. (1964). *Cours d'économie politique* (Vol. 1). Librairie Droz.

Rolski, T., Schmidli, H., Schmidt, V., & Teugels, J. L. (2009). *Stochastic processes for insurance and finance* (Vol. 505). John Wiley & Sons.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 290-295.

Wang, S. (1995). Insurance pricing and increased limits ratemaking by proportional hazards transforms. *Insurance: Mathematics and Economics*, *17*(1), 43-54.