COTOR Challenge 2 and Internal Review

Gary G Venter and John A Major – Guy Carpenter Instrat®

Abstract

We took the COTOR challenge as an opportunity to review internal methods for fitting severity distributions. These are to use MLE and the information matrix on the transformed beta family of distributions. They vary from part 4 methods only by using other information criteria for model selection and putting the covariance matrix from the information matrix into a lognormal distribution for the parameters, which seems to work better than the normal for small samples.

Alternatives we tried in this review include:

- Using the delta method to compute confidence intervals around the survival function as part of model selection. We typically look at intervals around the parameters as indications of how well the data has been able to identify parameters, but when parameters are correlated this is difficult to interpret, so the intervals around the survival function should help.
- A graphical analysis based on log-log plots of the survival function.
- Fitting a mixture of inverse exponential distributions.
- Testing the delta method's asymptotic results against the normal distribution of parameters from the information matrix. The delta method runs into problems with Jensen's inequality, and at least for the mean does not look very good for small samples.
- Testing the asymptotic normal distribution of the parameters against a Bayesian method that uses a non-informative prior. The gamma appears to give a better approximation than does either the normal or the lognormal that we have been using.
- Roughly quantifying model risk through a Bayesian approach to model distribution.

The simple Pareto above 5000 provides a reasonably good fit. For this, the distribution of the MLE given the parameter is well known. We studied a few more aspects of the Pareto.

- Only one of us (Venter) is a Bayesian. We explored Bayesian vs. frequentist approaches and found that adopting a particular plausible improper but informative prior produced agreement between the approaches on the parameter, but actually increased the difference between them on the layer mean. We did not manage to solve the Bayes vs. frequentist controversy.
- We tried a robust alternative to MLE. This produced an estimate similar to the frequentist one for the parameter and the Bayesian one for the layer mean.

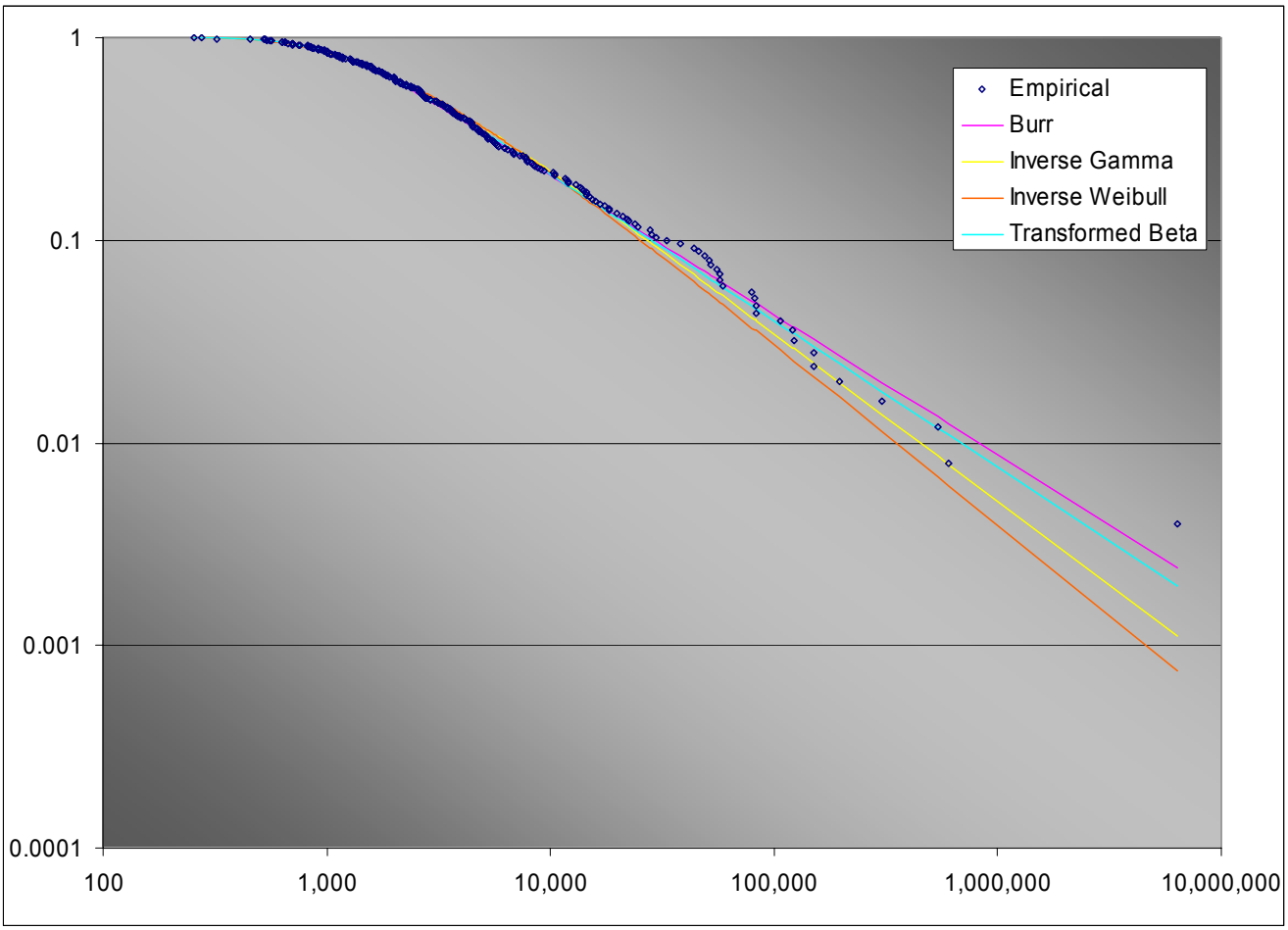
We also consider an alternative to counting parameters that adapts a method that works for nonlinear regression. It construes the fitted distribution as a fit of the sample, with x_i fit by $F^{-1}(p_i)$, where p_i is the sample probability of x_i. This is speculative however and more work would be needed to use this approach for evaluating distribution fits.

COTOR Challenge 2

250 claims were provided with the problem being the estimation of the expected cost of the layer 5M x 5M, with a 95% confidence interval. Since no frequency data was provided – in fact it was not even specified that the 250 claims represent a single year – the estimation below was done per ground-up claim in a sample of 250.

Graphical Analysis

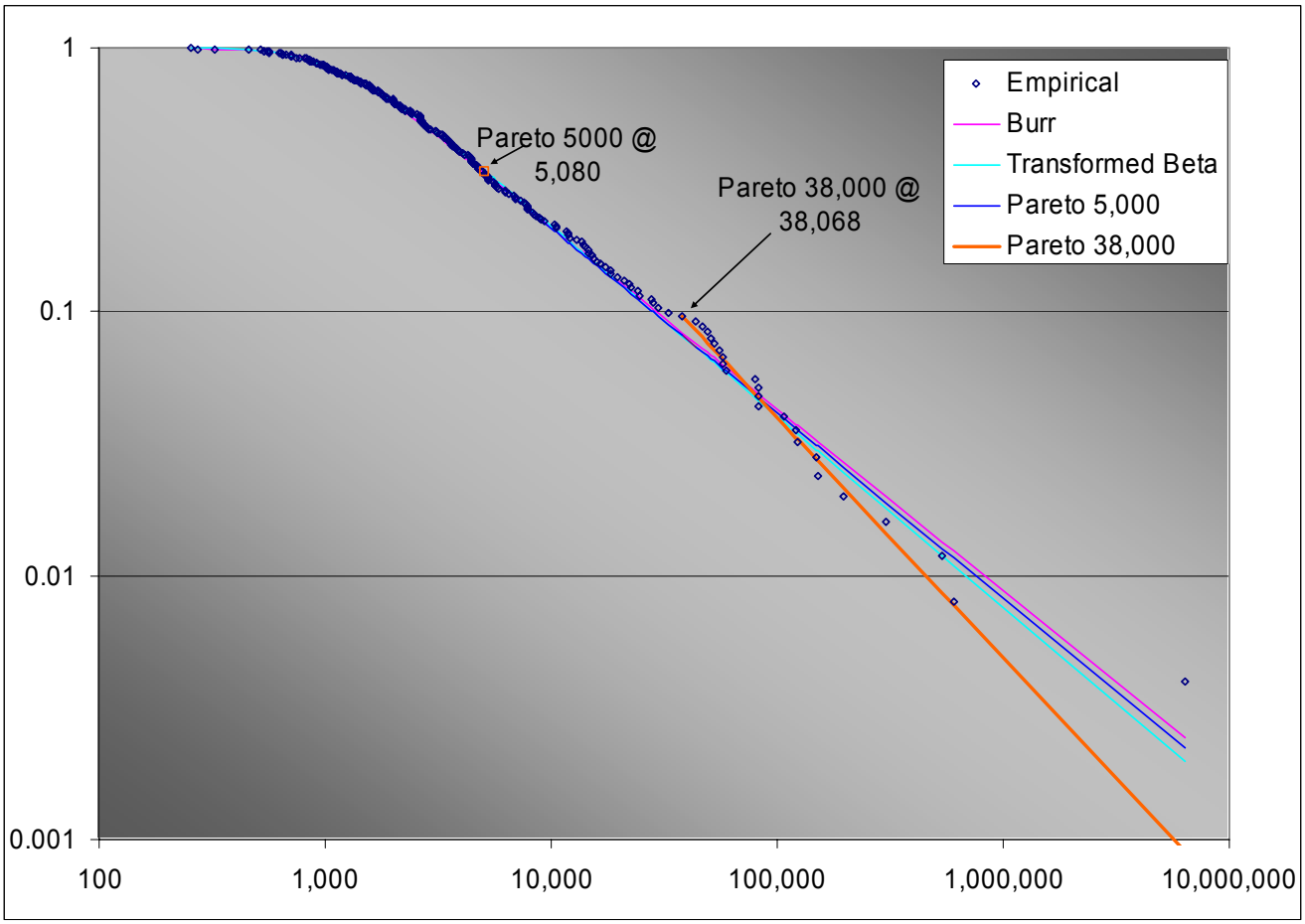
First a number of distributions from the transformed beta family were fit to the data. Some of the best fitting are graphed below. The graph shows the empirical and fitted survival functions.



The inverse Weibull and inverse gamma look too light at the very end. The transformed beta fits a bit better than the Burr through most of the tail, but is lighter at the very end.

For comparison, simple Paretos were fit above 5000 and 38,000. These are graphed along with the two heavier-tailed distributions below. The Pareto above 5000 falls right in between the Burr and transformed beta, whereas the Pareto above 38,000 has a much lighter tail. However from 38,000 to 600,000 it appears to fit the data fairly well. This raises the question as to whether or not pick-

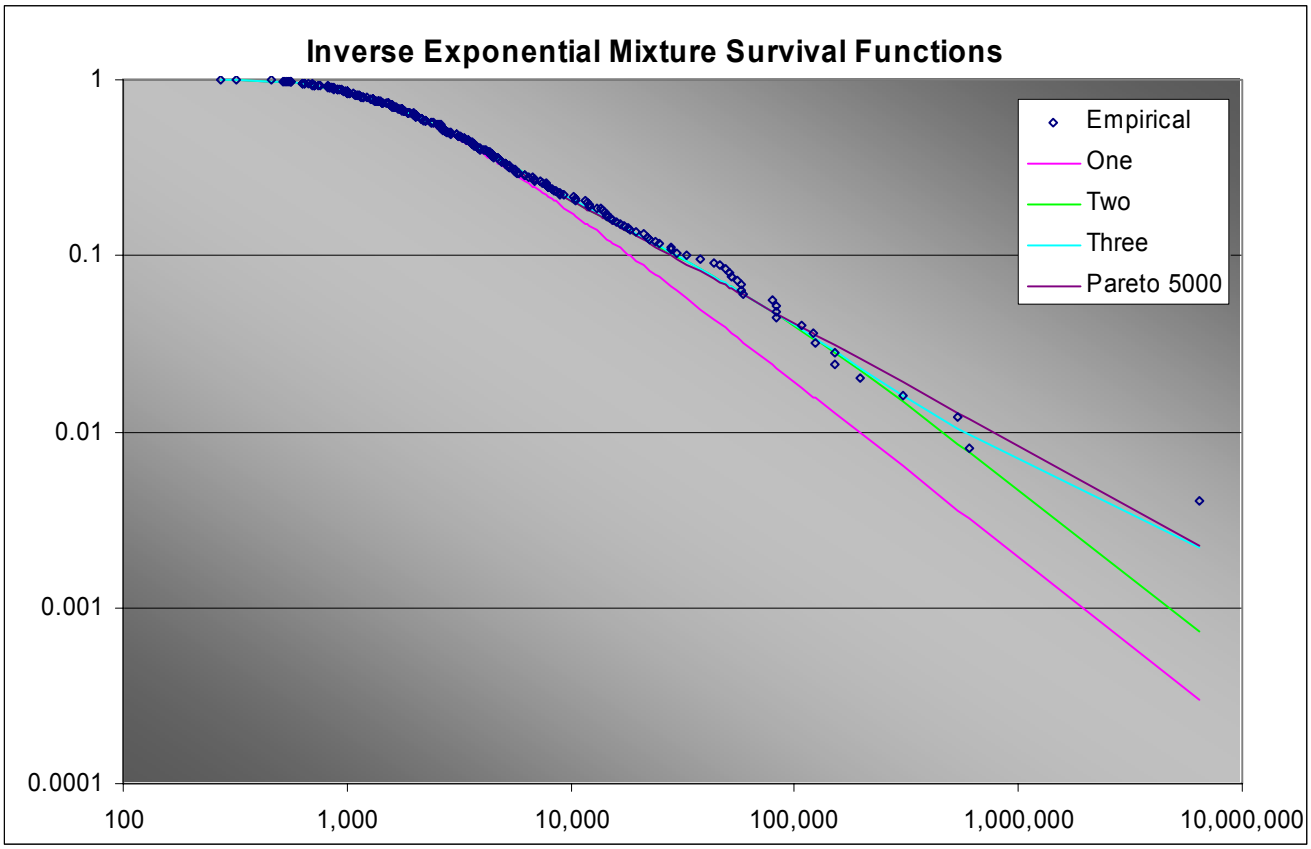
ing a starting point of 5000 is biasing the Pareto tail.



Some perspective on this can be provided by the Anderson-Darling test. This is a goodness-of-fit test that gives particular weight to the tail fits. The statistic for the Pareto 5000 is 0.292. You can reject the fit at the 95th percentile if the statistic is greater than 1.321. Thus the tail deviation around this distribution is well within the range of normal statistical fluctuation.

The errors are however correlated, which creates a possibility that this distribution is actually a mixture of some sort. To test this, a series of mixed inverse-exponential distributions were fit. The inverse exponential has $F(x) = e^{-\theta/x}$. For mixing more than one of these, the weights were established in advance and MLE found the θ's. ISO uses the opposite procedure for mixing exponentials. The rule used for setting the weights is that the second weight is 5% of the first, and the third 5% of the second, etc. Thus for instance for weighting 3 inverse exponentials the weights are 400/421, 20/421, and 1/421. The resulting survival functions for the first 3 mixtures and for the Pareto 5000 are graphed below.

All three mixtures fit well below about 6000, but the single distribution is too light in the tail (despite having an infinite mean!). The mixture of two fits well except at the end, and the mixture of three fits quite well all along. It also agrees with the Pareto at the last point.



The graphical analysis overall probably favors the Pareto 5000. It is not as heavy at the end as the Burr, more so than the transformed beta, and the same as the 3-mixture of inverse exponentials. Since the problem is one of large loss potential, the lighter distributions are disindicated, and the slightly better fits of the transformed beta and 3-mixture for some larger losses seem irrelevant. In most insurance applications there is no reason to believe that the smaller and larger losses are parametrically related, so only using the losses above 5000 does not appear to lose information.

Statistical Analysis

Some statistics of the fits are shown below.

Distribution	HQ/2	AIC/2	Pr >500K	CV	Pr >5M	CV	Pr ∃>6.4M
Empirical			1.20%		0.40%		100%
Inverse Gamma	2498.9	2497.5	0.92%	32.9%	0.14%	47.5%	24%
Inverse Weibull	2499.3	2497.9	0.73%	20.8%	0.09%	29.8%	17%
Burr	2500.2	2498.1	1.41%	32.5%	0.29%	46.5%	51%
Transformed Beta	2501.3	2498.9	1.25%	49.3%	0.24%	64.0%	45%
1-Mixure	2500.6	2499.9	0.39%		0.04%		7%
2-Mixture	2498.2	2496.8	0.91%		0.09%		17%
3-Mixture	2499.7	2497.6	1.09%		0.26%		42%
Simple Pareto > 5000			1.34%	36.8%	0.27%	55.3%	43%
Simple Pareto > 38,000			0.11%		0.09%		20%

The HQ/2 information criterion is to compare –lnL + number of parameters*log[log(n)], with n observations, where ln L is the maximal value of the log likelihood function. Lower values indicate a better fit. The HQ[1] is a compromise between the Akaike information criterion (AIC) and the Schwartz Bayesian criterion (BIC). In this case HQ/2 penalizes each additional parameter by 1.71. AIC/2 is half of the Akaike measure and penalizes each parameter only by 1. BIC/2 has a penalty of $\log\sqrt{n}$ per parameter, which is 2.76 in this case.

One problem with such information criteria is that the number of parameters can be deceptive. For instance, if two parameters are ±100% correlated, there is really only 1 parameter, as one determines the other. If two parameters are highly but not perfectly correlated, there could still be effectively less than two whole parameters. A suggestion about how parameters might be counted is in the Appendix, but for now it is enough to note that the information criteria do not give the final word on goodness of fit.

If parameters are closely linked, the resulting distribution function might have less variability than the parameters themselves do. Thus intervals around the distribution function can help evaluate goodness of fit, with tighter intervals indicating a better fit. This would happen for instance if the parameters are highly correlated. A tighter interval in general shows that the data has determined the distribution function more closely. In this case the CV (standard deviation divided by estimated value) is shown for two points on the survival function. The inverse Weibull has the tightest intervals of the distributions compared by this measure.

The CV's shown were calculated by the delta method from the Part 4 study note. This starts by expressing the survival function as a function g($\boldsymbol{\theta}$) of the parameter vector $\boldsymbol{\theta}$. Also needed is the covariance matrix $\boldsymbol{\Sigma}$ of the parameters from the MLE fitting procedure. Then the delta method gives the variance of g($\boldsymbol{\theta}$) as $g'^{T}\Sigma g'$, where g' is the (vertical) vector of partial derivatives of g wrt the parameters. The parameters and their CV's and correlations for some of the distributions are:

Parameter	Burr	Inv. Gamma	Transformed Beta	Inv. Weibull
Alpha	0.69	0.83	0.72	0.89
Theta	1,038.52	1,602.52	698.38	2,040.09
Beta	2.92		4.44	
Tau			1.94	
Parameter 1-2 Corr	69.65%	74.45%	(40.23%)	(30.70%)
Parameter 1-3 Corr	(46.13%)		38.60%	
Parameter 2-3 Corr	(73.49%)		(97.96%)	
Parameter 1-4 Corr			(64.32%)	
Parameter 2-4 Corr			89.64%	
Parameter 3-4 Corr			(83.05%)	
Parameter 1 CV	8.89%	7.8%	11.03%	5.0%
Parameter 2 CV	12.32%	10.4%	64.87%	7.4%
Parameter 3 CV	15.45%		64.16%	
Parameter 4 CV			44.22%	

[1] Hannan, E. and B. Quinn (1979). The determination of the order of an autoregression. *Journal of the Royal Statistical Society B*(41), 190–195.

These distributions have been parameterized so that moments exist in $(-\beta,\alpha)$. For instance the Burr df is $F(x) = 1 - (1+(x/\theta)^{\beta})^{-\alpha/\beta}$.

Even though the mixture of two inverse exponentials might be considered the best fit overall, the interest in this case is in the tail, so the fit in the tail is also examined. The best matches to the empirical survivor function at 500K and 5M are the transformed beta, 3-mixture, Pareto 5000, and Burr. The Burr is also better than the inverse gamma in the CV of the survival probabilities. The transformed beta is substantially worse, however, which suggests that it is truly over parameterized. The simple Pareto has CV's a bit higher than the Burr but comes closer to the empirical probabilities at these points. In addition, the last column of the first table shows the probability that at least one claim of 6.4M or greater would appear in a sample of this size from each distribution. For the single inverse exponential this is only 7% which would seem to reject this distribution at the 7% level. Although it is not low enough for statistical rejection at traditional levels for any of the other distributions, it is uncomfortably low for the inverse Weibull, Pareto 38,000, 2-mixture, and inverse gamma. Since fit for large losses is important, these will be eliminated.

In the end, the 3-mixture and Pareto 5000 are selected for further analysis. They are both in the middle of group in tail measures. The transformed beta appears to be over parameterized, and the Burr does not appear to give a better fit and is the most extreme in reacting to the largest point.

Layer Mean

The expected layer loss per claim is the limited average severity at 10M less that at 5M. For the Pareto this comes to 10,282 vs. 9977 for the 3-mixture. These are close enough that the Pareto will be used from here on to estimate ranges. The limited average severity is a function of the parameters, so the delta method also gives its standard deviation. In this case the standard deviations are more than half the mean, so a normal approximation would give some probability to a negative cost. Thus a gamma approximation to the confidence interval is used. The gamma was selected instead of the lognormal because it has lower skewness, which makes it more like the normal of the asymptotic case.

	Layer Mean	Standard Deviation	2.5th Percentile	97.5th Percentile
Pareto	30,601	17,788	6,217	74,038
Scaled Pareto	10,282	6,070	2,089	24,877

The Pareto was fit above 5000, so only had 84 claims to fit, compared to 250 for the Burr. Thus the layer price per claim is per claim above 5000. This was 33.6% of the claims, so the mean can be expressed per ground up claim by scaling by the factor 0.336. The standard deviation cannot be so easily scaled, however, as there is uncertainty about the 0.336 factor. If the number of claims over 5000 out of the sample is considered binomial in 250 and 0.336, (0.336)(0.664)250 is its variance, so the variance of its ratio to 250 is (0.336)(0.664)/250. This binomial can be considered independent of the Pareto α, so the variance formula for product of independent variables can be used to calculate the variance of the scaled Pareto mean, and that gives the values shown.

Testing Asymptotic Assumptions

The distribution of the parameters from MLE is asymptotically multi-variate normal. In addition the expected value of a function of the parameters (here the layer mean) is asymptotically the function applied to the MLE-estimated parameters. However since there are not a whole lot of claims here, the asymptotic results might not apply.

This is relatively easy to test for the Pareto since there is only one parameter. The standard deviation of the parameter is about 11% of the parameter, so a normal approximation would not give any meaningful probability to negative parameters. However the standard deviation of the layer mean using the delta method is about 60% of the mean, which would allow negative expected values if a normal were applied. Apparently some results are more asymptotic than others.

In fact Jensen's inequalities would say that the mean of a function of the parameters would not equal the function of the mean for concave or convex functions. Thus the asymptotic rule for applying functions to the parameters might implicitly assume that the normal distribution is approaching a point mass.
First the mean of the function of the parameters can be tested by computing the conditional layer mean given α for the range of possible α's – here taken to be those within 5 standard deviations of the MLE estimate. Assuming that the distribution of the α's is normal, the mean of the conditional layer means can be computed by integration over the normal probabilities. The percentiles of the layer mean can also be computed from the conditional probabilities of the layer mean given the parameter.

Numerical integration of the postulated normal distribution of α finds the mean, 2.5th, and 97.5th percentiles of the layer loss per claim over 5000 to be approximately 36,200, 10,200, and 99,300. These are 18%, 63%, and 34%, respectively, above the values from the delta method. These differences make the straight application of functions to the MLE parameters questionable for both the mean and the distribution.

The normal assumption for α itself can be tested with a Bayesian approach suggested by Rodney Kreps. First a non-informative prior can be postulated for α – in this case $f(\alpha) = 1/\alpha$ on $(0,\infty)$. The probability mass diverges at both ends of the interval, so this f has infinite pulls both upward and downward, and so it should have minimal impact on the resulting posterior distribution. In contrast, a prior of 1 has more weight excess of any amount than below it on the positive reals, so exerts an upward pull.

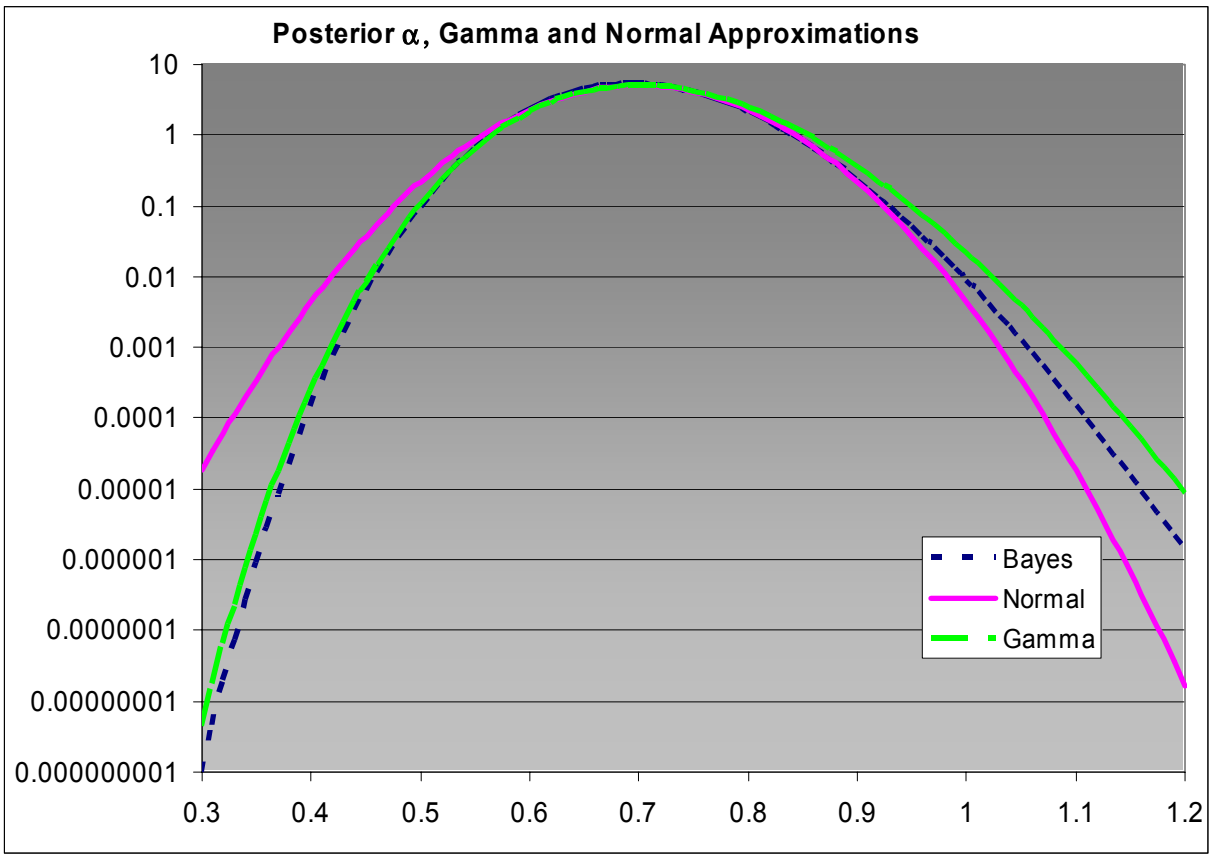
The likelihood function $L(\alpha)$ is the conditional distribution of the data given α. So the distribution of α given the data must be proportional to $L(\alpha)/\alpha$. This has to integrate to 1, which specifies the constant of proportionality. The resulting density is compared to the normal below, on a log scale.

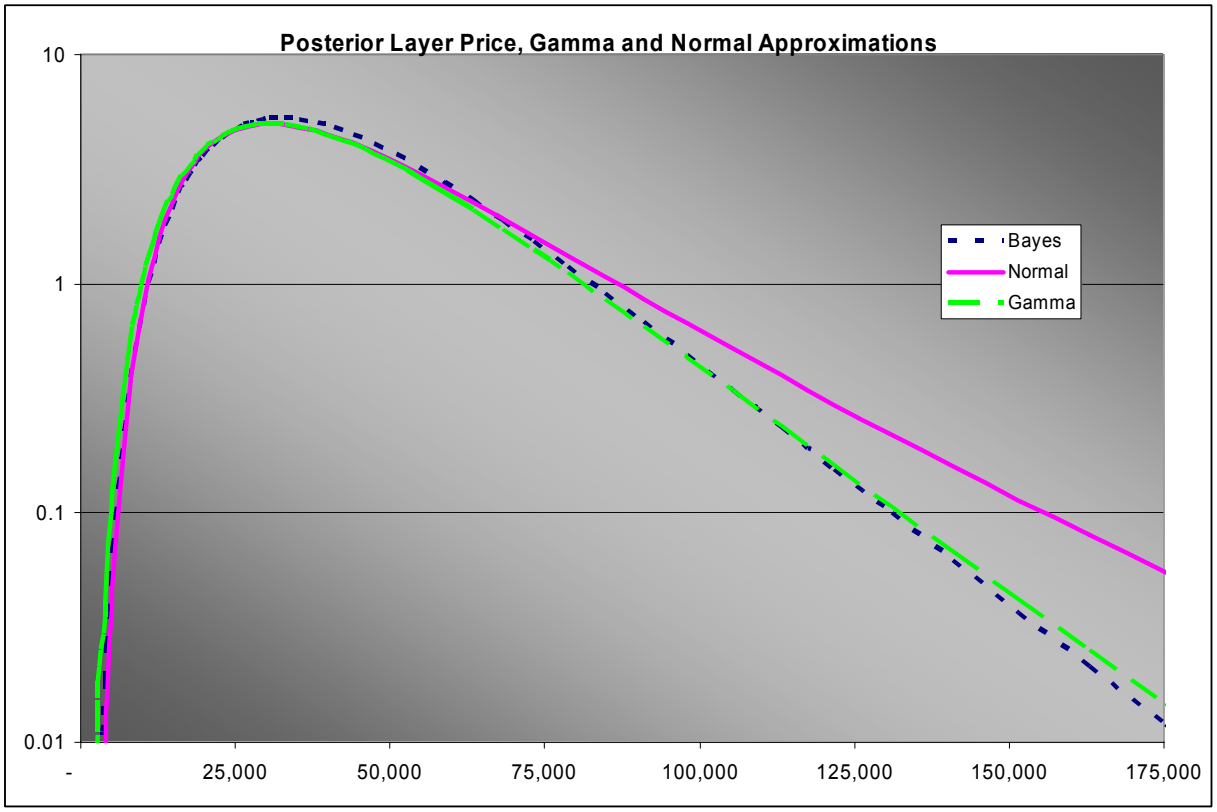
The Bayesian posterior is positively skewed. Since the layer mean is a decreasing function of α, this gives less weight to higher values of the mean than the normal does. The expected layer cost

then comes out approximately 35,300 with 2.5th and 97.5th percentiles of 10,000 and 88,600.

Scaling this to 250 claims gives a per claim expected layer cost of about 12,000 with a range of (3000, 30,000). This range again scales the divergences by more than the mean (the distances from the percentiles to the mean have been increased by the ratio of the delta method standard deviations before and after being mixed by the binomial).

The graphs also put in a gamma approximation to the distributions of α and the layer mean. These were set by matching the CV and mode to the MLE estimates, as MLE goes for the maximum value, and the mode might also avoid the Jensen problem. The gamma for α is actually heavier than the posterior distribution in the right tail. The lognormal (not shown) is heavier still, but the Weibull is even lighter than the normal (with shape parameter above about 3.6 the Weibull is negatively skewed). Thus the gamma is probably the best bet for a distributional assumption. For the layer mean the tails are reversed and the gamma matches the posterior fairly well. The easiest way to apply this in the multivariate case is probably with the normal copula.





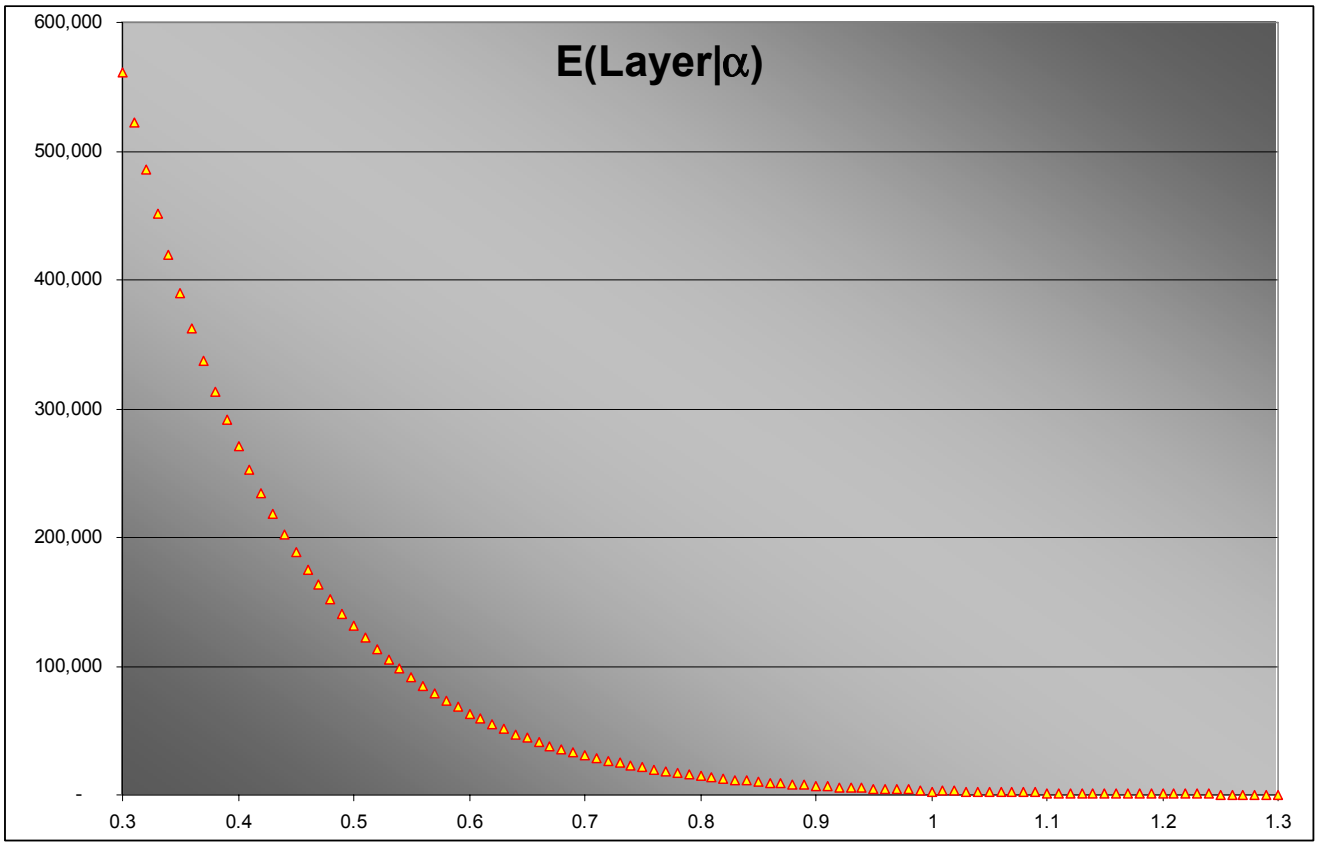
Bias

If a sample $(x_1, \ldots, x_n)$ of size n is drawn from a population distributed Pareto in α with minimum c, the MLE *a* of α is[2] $n/\Sigma\ln(x_j/c)$. The distribution of *a* can be derived from the Pareto assumption, and it is an inverse gamma: $f(a|\alpha) = (\alpha n/a)^n/[a\Gamma(n)e^{-\alpha n/a}]$ with an expected value of $n\alpha/(n-1)$. Thus $(n-1)a/n$ is an unbiased estimate of α. In this case the MLE is biased upward, which is biased towards being a less heavy-tailed distribution. Nonetheless, using *a* to calculate the expected layer cost will result in an upwardly biased cost. The values of the parameter less than *a* produce rapidly increasing layer costs, so much so that even though the lower parameters are a bit less likely than the higher parameters, the expected layer cost averaged over all the possible parameters from the inverse gamma is above the cost that would be produced by the true parameter α. Simulation and numerical integration both show that for α around 0.7 with a sample of n=84, the bias in the layer cost from the MLE estimator is about +9%.

Thus adjusting *a* for bias would decrease it, but that would lead to a higher layer cost, which would already be biased upward if it were based on *a*. This ambiguous situation in adjusting for bias was discussed in Major (1999) "Taking Uncertainty Into Account: Bias Issues Arising from Parameter Uncertainty in Risk Models," *CAS Forum* Summer. His conclusion is that there is no single adjustment for bias – the adjustment depends on what you are trying to estimate.

[2] E.g., see Rytgaard (1990) "Estimation in the Pareto Distribution" ASTIN 20 #2 for this and the distribution that follows.

This can be clarified by looking at the graph of the layer cost by α. This is shown per claim above 5000 so is about triple what it would be per ground-up claim. However it is clear that the cost increases sharply with lower α. Thus when you end up estimating an α that is a little too small, the estimated layer cost is considerably too high. Thus even though a lot of samples from a distribution with a given α will produce estimates of the parameter that are on the average a bit too light-tailed, the average of the layer costs will be somewhat higher than the cost from the real α.



From a Bayesian viewpoint, you would like to know what the distribution of α is given a. The inverse gamma above is $f(a|\alpha)$. With an opinion of how α might be distributed, the posterior could be calculated. A general thought for insurance risk is that α could be anything, but is more likely to be small. Specifying $f(\alpha) \propto 1/\alpha^2$ would express this. Its integral from zero to any positive number diverges, but from there to infinity is finite. Thus it has somewhat of a downward pull on the posterior distribution. With this prior, $f(\alpha|a)$ is gamma distributed with scale parameter a/n and shape parameter $n - 1$. The expected value of α is then $(n - 1)a/n$, which is the unbiased estimate.

This is not a bias adjustment, but is rather a model of some properties of the universe that is generating the sample. The model says that there is a set of α's that could produce insurance losses from any line, with smaller values being more likely but any value possible. Then given some data the distribution changes as a result of seeing the data. The resulting expected value of α is somewhat below the MLE from the data and is a number such that the expected value of the MLE

from that α is the observed MLE. In that model of the universe, the expected value of the layer is higher than the conditional expected value given either *a* or α. This is because that model is always anticipating lower values of α.

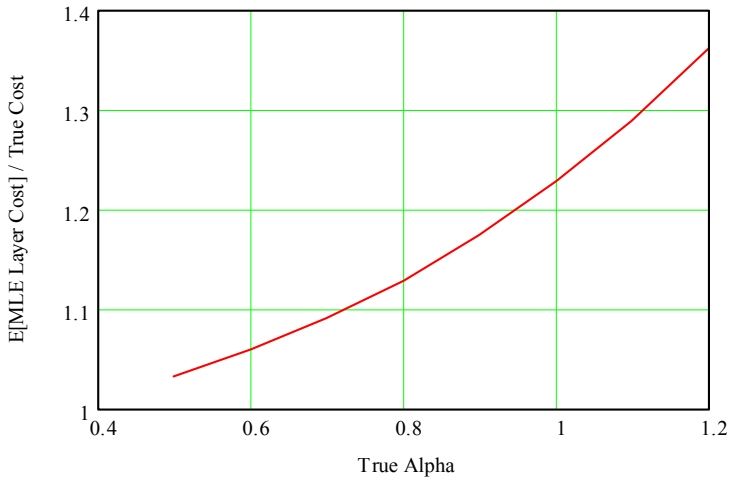
In this case, the MLE *a* = 0.70 produces an estimate of α of 0.691667. The layer cost from an actual α of 0.691667 is 10,900 per ground-up claim, but from the posterior gamma distribution of α's it would have an unconditional mean of 12,600. If a lot of random samples of size 84 were taken from a true Pareto with an α of 0.691667, the mean of the implied layer means from the MLE parameters would be about 11,900, which would be biased upward from the true mean by 9%. But since we do not know the true α and in fact are only estimating a distribution of it given the data, there is nothing inconsistent about estimating a mean layer cost that does not correspond to the mean parameter.

The posterior gamma has a (.025,.975) range of (3600 to 31600) around the mean of 12,600 for the layer expected cost.

Frequentist Alternative

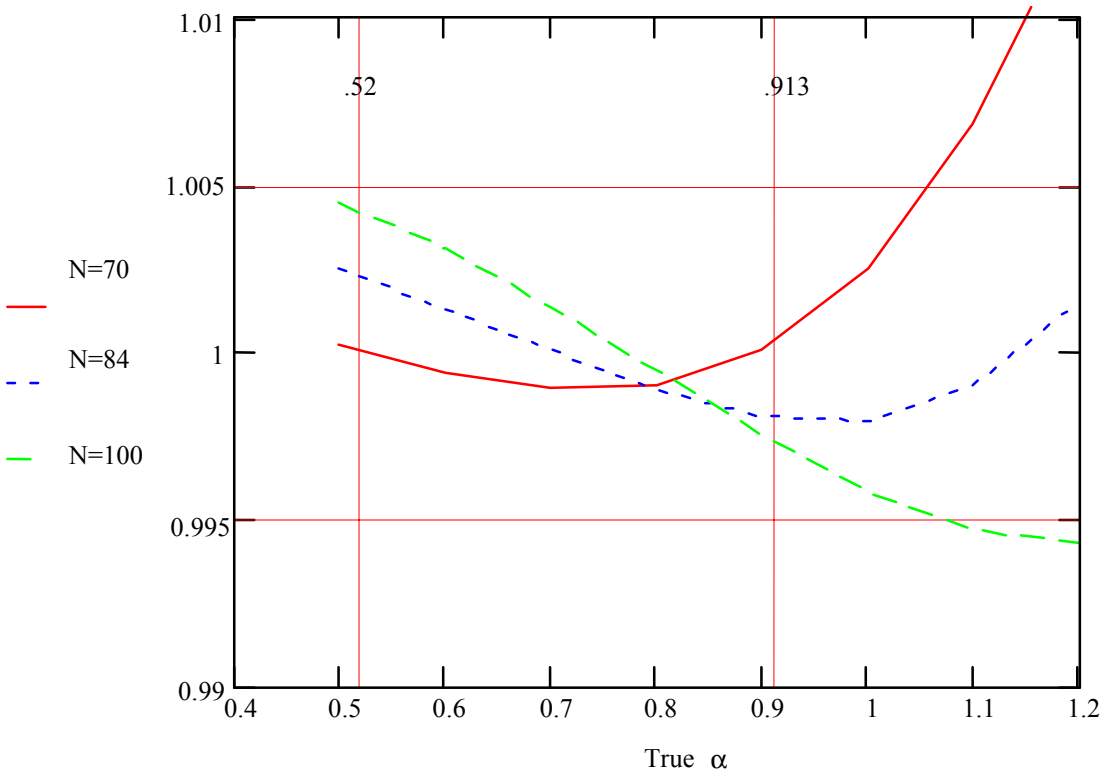
The authors do not both fully embrace the Bayesian viewpoint. While the Bayesians are looking for a single model that represents the data and the parameters as random variables, the frequentists seek estimators with certain desirable properties regardless of the source of the parameters. They find it hard to accept that the layer cost of 10,900 from the fitted α, which is biased upward, should be further increased. Any α in the vicinity of 0.70 would produce, over a large number of samples, a layer cost that is about 9% high when calculated from the MLE *a*. Thus to get an unbiased estimate, the layer cost from *a* = .70 should be reduced by about 9% to get an unbiased estimator. With a bit of refinement, the unbiased layer estimate is thus 9265 with a 95% confidence interval of (2654, 29248). The analysis that produces this is as follows.

This chart shows the ratio of layer expected value to actual value for samples of 84 at various values of true α based on numerical integration of the sampling distribution of the MLE *a*:



An approximately unbiased estimator is given by $\frac{LayerCost(a)}{0.983+\frac{24.5}{N}\cdot a^{2.5}}$.

The following graph, again based on numerical integration, shows it is within +/-0.5% of being unbiased for 70<N<100 and 0.5<α<1 (which covers the 99% two-tail CI for α for the COTOR sample of 0.520 to 0.913):



For our sample, this estimator gives a layer cost of 9265 versus the MLE of 10282. Notice this is lower than the result of simply applying the 9% reduction that would be appropriate if we knew the true α were 0.7.

A Bayesian Rejoinder

The layer cost of 9265 would arise from an α of about 0.714. With such a Pareto there is a 41% probability that the MLE estimator of a sample of 84 would be 0.70 or less, which is not so unreasonable. However the frequentist approach over-emphasizes expected values. "Bias" is a charged term that may lead to this emphasis, but with these highly nonlinear processes means are meaningless, or at least misleading. With a true α of 0.70, the median MLE estimate from samples of 84 is 0.7028, and the layer cost from this value is the median layer cost of 10,075. Thus a frequentist analysis using medians would at least adjust the α and the layer cost consistently. Adjusting 0.70 by

the ratio 7000/7028 gives an α of about 0.697 and a layer cost of about 10,500. However from a Bayesian viewpoint this still does not consider the full uncertainty to and opinion of the decision maker. The frequentists' viewpoint is that with any true α, the process of generating a sample, taking the MLE estimate from that sample, and calculating the layer mean from that will on the average give a high layer mean. But it will also be low more often than it is high. The Bayesian viewpoint admits that all is right, but emphasizes that the estimation has not pinned down the possible parameters. Since it is still possible that the actual parameter is different than your estimate, and the layer cost grows more with smaller parameters than it shrinks with larger parameters, the possible layer means you could be facing average to a higher number than the MLE estimate.

Robust Estimation

So far MLE has been emphasized. However it has two problems. First of all its good properties are asymptotic, so it is not clear how well it works for small samples. Also its efficiency (minimum variance property) holds only if the sample is actually from the distribution being fit. It is not robust to deviations from this assumption[3]. Robustness can be measured by the breakdown point (BP) which is basically the degree to which the estimated parameter remains uninfluenced by the presence of outlying observations, which possibly (but not with certainty) could be due to contamination of the dataset rather than being properly representative of the target parametric model. It is estimated by looking at the effect on the fitted parameter of replacing some of the largest observations by very large values. The outliers could be points that are too high or too low compared with what the model would generate so a robust procedure does not necessarily weaken the tail.

MLE has a BP of zero. Brazauskas and Serfling discuss an estimator they call the generalized median (GM) which is not much less efficient than MLE but is robust. It is essentially the median of the MLE taken over all subsets of a fixed size, like 3 or 4, of the sample. For this data the GM with 3 points produces an estimate of 0.674 for the Pareto 5000, with an expected layer mean of 12,400 and 95% confidence interval of (2300, 28000). This is in close agreement with the posterior distribution adjusting the MLE fit. The GM estimator is probably unbiased for the parameter, but a simulation study of drawing samples from a known Pareto and looking at the resulting GM-estimated layer cost still shows an upward bias for the layer mean. Overall however the agreement of the GM result with the posterior gamma is encouraging and does not move us off of the gamma values.

Model Risk

The results so far have built in a bit of typical actuarial conservatism. Because an upper layer was of interest, distributions that did not give a high enough (over 20%) probability of actually observing the largest claim in the sample were discarded. In addition a prior was assumed that gave a downward pull to the Pareto parameter, albeit only to the extent that MLE over-estimates this parameter. This prior does increase the layer mean over what plugging in the actual estimated parameter would produce, however, even though the layer mean from the parameter is already bi-

[3] For instance, see Brazauskas and Serfling (2000) "Robust and Efficient Estimation of the Tail Index of a Single-Parameter Pareto Distribution," NAAJ 4 4 from which most of this section derives.

ased upward from what a true Pareto layer cost would be (assuming that the sample is drawn from a true Pareto the layer cost from the sample MLE parameter is on the average higher than the true layer cost).

Some of these conservative elements would be readily picked up by a reinsurance underwriter who used to be an actuary and pressure would be applied to have a more balanced view. Especially discarding better overall fits might be resisted, as the large loss could be a rare observation from a lighter-tailed but better fitting distribution. So a method of combining estimates from different distributions could be needed in a competitive situation.

Since there are really only two types of fits competing here – good tail fits with overall worse fit and good overall fits with worse tail fits – and the tail probabilities were similar for the members within each group, it should suffice to weight together one of each. This is easiest for the Pareto 5000 and Pareto 38,000, with α's of 0.70 and 0.91 and probabilities of having the largest actual loss or one larger in their samples of 43% and 20%. These later probabilities will be taken as proxies for Pr(sample|parameter). To do a weighting a Bayesian procedure is proposed with the $1/\alpha$ prior, to be unbiased. The posterior probabilities of the two Paretos are then proportional to 43%./0.7 and 20%/0.91 = 0.6143 and 0.2198. Adjusting to add to 1, these are 74% and 26%. Applying the $1/\alpha^2$ prior to 0.91 produces a layer mean of 7360. Weighting this with 12,600 gives a weighted mean of 11,300. The combined confidence interval is approximately 1500 to 29,000.

This is the final answer from the Bayesian point of view. However some additional information gathered informally gives a different result. This is discussed in Appendix 3, which takes a game theory perspective.

Appendix 1 – Counting Parameters

This problem also comes up in fitting models to data when each data point can be expressed as the model for that point plus a random innovation. One suggestion in that case, for instance proposed by Jianming Ye (1998) "On measuring and correcting the effects of data mining and model selection" J Am Stat Assoc 93:120 - 31, is to define the generalized degrees of freedom used up (i.e., number of parameters in the model) as the sum over all the observations of the derivative of the fitted value at that observation with respect to the observation. This might be possible analytically or can be approximated numerically by making a small change at an observation and seeing how much the fitted point changes, and repeating for all observations. This approach is discussed further in Efron, Bradley (2004) "The Estimation of Prediction Error: Covariance Penalties and Cross-Validation," JASA vol 99 #467, September.

As an example, suppose you fit a cubic polynomial to 4 points. The polynomial will go through all four. Changing any of the points by a small amount will change the corresponding fitted value by the same amount, so the sum of the derivatives will be 4. Thus all degrees of freedom are used up. One way to think of this is that each data point has a degree of freedom initially, which gives it power to pull the model towards itself. If it can completely control the model, so any change in the point changes the fitted value by the same amount, the model has used up that entire degree of freedom. If it can only pull the fitted value by half of the change, the model has only used ½ of that degree of freedom, etc.

This could be applied to fitting distributions by considering every data point to be modeled as a quantile of the fitted distribution. In this case with 250 points, the 5th point (sorted ascending), for example, could be modeled by the x that has $F(x) = 1/500 + 4/250 = 1.8\%$. Then the derivatives of the fitted points wrt the corresponding data points can be summed as above to get the number of parameters.

As a test of this, the MLE estimate of the inverse exponential θ is $n/\Sigma x_j^{-1}$, and $F^{-1}(p) = -\theta/\ln(p)$. If $G(p_j)$ denotes the fitted loss at the j^{th} actual loss (order statistic) then $\partial G(p_j)/\partial x_j = [G(p_j)/x_j]^2/[n\ln(p_j)]$. For this sample these sum to 1.01, which is pretty close to the 1 parameter it should be.

This method actually aims to count the degrees of freedom used up by the fitting. Thus it might be possible for the same distribution to have a different number of degrees of freedom depending on the fitting method. However trying different fitting methods may give some insight. For instance, if a Burr is fit by forcing it through 3 points, that would use up 3 degrees of freedom by this method. However if this is not possible, that would suggest that the Burr actually has fewer than 3 parameters, although this number may vary somewhat by fitting method. Trying to match the Burr losses at $(2j - 1)/500$ for the 1st, 125th, and 250th largest observations gets close but is not able to fit exactly (minimum SSE for the quantiles appears to be about 95), so maybe there are intrinsically a tad fewer than 3 parameters.

Appendix 2 – Inverse Exponential Distribution

$F(x) = e^{-\theta/x}$
$f(x) = \theta e^{-\theta/x}/x^2$
$E(X^y)$ exists for $y < 1$
$LAS_x = \theta E_1(\theta/x) + x(1 - e^{-\theta/x})$ where E_1 is the exponential integral:

$$E_1(z) \equiv \int_z^\infty \frac{e^{-t}}{t} dt = -\gamma - \ln z - \sum_{n=1}^{\infty} \frac{(-z)^n}{n\Gamma(n+1)}$$ where $\gamma = 0.57721566490153286060651209$ is the Euler-Mascheroni constant

$\partial f(x)/\partial\theta = (x - \theta)e^{-\theta/x}/x^3$
$\partial^2 f(x)/\partial\theta^2 = -(2x - \theta)e^{-\theta/x}/x^4$

Appendix 3 – Game Theory Perspective

This author learned that the layer to be priced was selected after the sample was drawn: "We wanted to select a part of the distribution where there was not a lot of data - so that it would truly be a challenge."[4]

The frequentist approach outlined in the main body of the paper seeks an estimation methodology which will give, on average, the correct answer in a series of hypothetical[5] games with the following structure:

(1) Opponent chooses a distribution
(2) Opponent chooses attachment and limits for the layer
(3) Opponent draws a random sample and presents it to Player, asking for an estimate of layer cost.

Call this "Game A." It appears this is not quite accurate. Instead, we have Game B:

(1) Opponent chooses a distribution
(2) Opponent draws a random sample
(3) Opponent observes maximum of the sample and chooses attachment and limits
(4) Opponent presents results to Player, asking for an estimate of layer cost

Unfortunately, the specification of step (3) is somewhat vague. We can conceptualize it as

(3) Opponent chooses attachment = A(Xmax) and limit = L(Xmax)

but we need suitable functions A and L to operationalize this. For the following analysis, I will choose L(X) = A(X) = 0.78X.

Note the Game B specification is only an approximation to reality. It certainly is not likely that the challenge would involve layers defined in anything other than round numbers. It is likely that a Pareto model for the tail is not exactly correct. More importantly, it might be that several samples were examined, or that calculations or estimates of the true probability of exceeding $5mm might have been done. f the maximum of the sample had been $4.5mm, or $10.5mm, the layer still might have been 5XS5. Nonetheless it is instructive to examine the long-run frequency behavior of the MLE in this context.

Simulation shows that for a sequence of games of type B the MLE produces layer cost estimates that are on average 19% higher than actual, and this factor appears to be constant across the relevant range of alpha values. This makes it easy to create an unbiased estimator – simply divide the MLE by 1.19. The result for this problem is $8587.

Percentage points for confidence intervals were created by parametric bootstrap and "BCa" (bias

[4] Louise Francis, personal communication.
[5] This is perhaps the philosophical Achille's Heel of the Frequentist methodology, but it is vital from a Game Theory perspective. See the sidebar on the Three Doors problem in Major (2002) "Advanced Techniques for Modeling Terrorism Risk."

corrected accelerated) method with 100,000 replications.[6] This took into account both the variability in estimating α and the variability in estimating the fraction of claims over $5000. Specifically, the ratio of ML estimated layer cost to the actual layer cost was the random variable analyzed, because the layer itself is a function of the sample. Percentage points for this ratio were applied to the MLE for the particular sample given in the challenge. The results are shown in the following table:

2.5%	$	3,708
10.0%	$	5,349
25.0%	$	7,323
50.0%	$	10,441
75.0%	$	15,381
90.0%	$	22,802
97.5%	$	38,849

[6] Efron and Tibshirani (1993) An Introduction to the Bootstrap.