Approximations of the Aggregate Loss Distribution

Dmitry E. Papush, Ph.D., FCAS,
Gary S. Patrik, FCAS, and Felix Podgaits

Approximations of the Aggregate Loss Distribution

Dmitry Papush, Ph.D., FCAS,
Gary Patrik, FCAS
Felix Podgaits

Abstract

Aggregate Loss Distributions are used extensively in actuarial practice, both in ratemaking and reserving. A number of approaches have been developed to calculate aggregate loss distributions, including the Heckman-Meyers method, Panjer method, Fast Fourier transform, and stochastic simulations. All these methods are based on the assumption that separate loss frequency and loss severity distributions are available.

Sometimes, however, it is not practical to obtain frequency and severity distributions separately, and only aggregate information is available for analysis. In this case the assumption about the shape of aggregate loss distribution becomes very important, especially in the "tail" of the distribution.

This paper will address the question of what type of probability distribution is the most appropriate to use to approximate an aggregate loss distribution.

Introduction

Aggregate loss distributions are used extensively in actuarial practice, both in ratemaking and reserving. A number of approaches have been developed to calculate aggregate loss distribution, including the Heckman-Meyers method, Panjer method, Fast Fourier transform, and stochastic simulations. All these methods are based on the assumption that separate loss frequency and loss severity distributions are available.

Sometimes, however, it is not practical to obtain frequency and severity distributions separately, and only aggregate information is available for analysis. In this case, the assumption about the shape of aggregate loss distribution becomes very important, especially in the "tail" of the distribution.

This paper will address the question what type of probability distribution is the most appropriate to use to approximate an aggregate loss distribution. We start with a brief summary of some important results that have been published about the approximations to the aggregate loss distribution.

Dropkin [3] and Bickerstaff [1] have shown that the Lognormal distribution closely approximates certain types of homogeneous loss data. Hewitt, in [6], [7], showed that two other positive distributions, the gamma and log-gamma, also provide a good fit.

Pentikainen [8] noticed that the Normal approximation gives acceptable accuracy only when the volume of risk business is fairly large and the distribution of the amounts of the individual claims is not too heterogeneous. To improve the results of Normal approximation, the NP-method was suggested. Pentikainen also compared the NP-method with the Gamma approximation. He concluded that both methods give good accuracy when the skewness of the aggregate losses is less than 1, and neither Gamma nor NP is safe when the skewness of the aggregate losses is greater than 1.

Seal [9] has compared the NP method with the Gamma approximation. He concluded that the Gamma provides a generally better approximation than NP method. He also noted that the superiority of the Gamma approximation is even more transparent in the "tail" of the distribution.

Sundt [11] in 1977 published a paper on the asymptotic behavior of the compound claim distribution. He showed that under some special conditions, if the distribution of the number of claims is Negative Binomial, then the distribution of the aggregate claims behaves asymptotically as a gamma-type distribution in its tail. A similar result is described in [2] (Lundberg Theorem, 1940). The theorem states that under certain conditions, a negative binomial frequency leads to an aggregate distribution, which is approximately Gamma.

The skewness of the Gamma distribution is always twice its coefficient of variation. Since the aggregate loss distribution is usually positively skewed, but does not always have skewness double its coefficient of variation, adding a third parameter to the Gamma was suggested by Seal [9]. However, this procedure may give positive probability to negative losses. Gendron and Crepeau [4] found that, if severity is Inverse Gaussian and frequency is Poisson, the Gamma approximation produce reasonably accurate results and is superior to the Normal, N-P and Escher approximations when the skewness is large.

In 1983, Venter [12] suggested the Transformed Gamma and Transformed Beta distributions to approximate the aggregate loss distributions. These gamma-type distributions, allowing some deviation from the Gamma, are thus appealing candidates.

This paper continues the research into the accuracy of different approximations of the aggregate loss distribution. However, there are two aspects that differentiate it from previous investigations.

First, we have restricted our consideration to two-parameter probability distributions. While adding the third parameter generally improves accuracy of approximation, observed samples are usually not large enough to warrant a reliable estimate of an extra, third, parameter.

Second, all prior research was based upon theoretical considerations, and did not consider directly the goodness of fit of various approximations. We are using a different approach, building a large simulated sample of aggregate losses, and then directly testing the goodness of fit of various approximations to this simulated sample.

Description of the Method Used

The ideal method to test the fit of a theoretical distribution to a distribution of aggregate losses would be to compare the theoretical distribution with an actual, statistically representative, sample of observed values of the aggregate loss distribution. Unfortunately, there is no such sample available: no one insurance company operates in an unchanged economic environment long enough to observe a representative sample of aggregate (annual) losses. Economic trend, demography, judicial environment, even global warming, all impact the insurance marketplace and cause the changes in insurance losses. Considering periods shorter than a year does not work either because of seasonal variations.

Even though there is no historical sample of aggregate losses available, it is possible to create samples of values that could be aggregate insurance losses under reasonable frequency and severity assumptions. Frequency and severity of insurance losses for major lines of business are being constantly analyzed by individual insurance companies and rating agencies. The results of these analyses are easily available, and of a good quality. Using these data we can simulate as many aggregate insurance losses as necessary and then use these simulated losses as if they were actually observed: fit a probability distribution to the sample and test the goodness of fit. The idea of this method is similar to the one described by Stanard [10]: to simulate results using reasonable underlying distributions, and then use the simulated sample for analysis.

Our analysis involved the following formal steps:

1. Choose severity and number of claims distributions;
2. Simulate the number of claims and individual claim amounts, and calculate the corresponding aggregate loss;
3. Repeat many times (5,000) to obtain a sample of aggregate losses;
4. For different probability distributions, estimate their parameters, using the simulated sample of aggregate losses;
5. Test the goodness of fit for the various probability distributions.

Selection of Frequency and Severity Distributions

Conducting our study, we kept in mind that the aggregate loss distribution could potentially behave very differently, depending on the book of business covered. Primary insurers usually face massive frequency (large number of claims), with limited fluctuation in severity (buying per occurrence excess reinsurance). To the contrary, an excess reinsurer often deals with low frequency, but a very volatile severity of losses. To reflect possible differences, we tested several scenarios that are summarized in the following table.

Scenario #	Type of Book of Business	Expected Number of Claims	Per Occurrence Limit	Type of Severity Distribution
1	Small Primary, Low Retention	50	\$0 – 250K	5 Parameter Pareto
2	Large Primary, Low Retention	500	\$0 – 250K	5 Parameter Pareto
3	Small Primary, High Retention	50	\$0 – 1000K	5 Parameter Pareto
4	Large Primary, High Retention	500	\$0 – 1000K	5 Parameter Pareto
5	Working Excess	20	\$750K xs \$250K	5 Parameter Pareto
6	High Excess	10	\$4M xs \$1M	5 Parameter Pareto
7	High Excess	10	\$4M xs \$1M	Lognormal

Number of claims distribution for all scenarios was assumed to be Negative Binomial. Also, we used Pareto for the severity distribution in both primary and working excess layers. In these (relatively) narrow layers, the shape of the severity distribution selected has a very limited influence on the shape of the aggregate distribution. In a high excess layer, where the type of severity distribution can make a material difference, we tested two severity distributions: Pareto and Lognormal. More details on parameter selection for the frequency and severity distribution can be found in the exhibits that summarize our findings for each scenario.

Distributions Used for the Approximation of Aggregate Losses

As we discussed before, we concentrated our study on two-parameter distributions. Basically, we tested three widely used two-parameter distributions, to test their fits to the aggregate loss distributions constructed in each of the seven scenarios. Each of these three distributions was an appealing candidate to provide a good approximation. The following table lists the three distributions used.

Type of Distribution	Parameters	Probability Density Function	Mean	Variance
Normal	μ $\sigma > 0$	$f(x) = 1/(\sigma\sqrt{2\pi})$ * $\exp(-(x - \mu)^2/(2\sigma^2))$	μ	σ^2
Lognormal	μ $\sigma > 0$	$f(x) = 1/(\sigma x\sqrt{2\pi})$ * $\exp(-(\ln x - \mu)^2/(2\sigma^2))$	$\exp(\mu + \sigma^2/2)$	$\exp(2\mu + \sigma^2)$ * $[\exp(\sigma^2) - 1]$
Gamma	$\alpha > 0$ $\beta > 0$	$f(x) = 1/(\Gamma(x))$ * $\beta^{-\alpha}x^{\alpha-1}\exp(-x/\beta)$	$\alpha\beta$	$\alpha\beta^2$

A Normal distribution appears to be a reasonable choice, at least when the expected number of claims is sufficiently large. One would expect a Normal approximation to work in this case because of the Central Limit Theorem (or, more precisely, its generalization for random sums; see, for instance, [5]). As we shall see, however, to make this happen, the expected number of claims must be extremely large.

A Lognormal distribution has been used extensively in actuarial practice to approximate both individual loss severity and aggregate loss distributions ([1], [3]). A Gamma distribution also has been claimed by some authors ([6], [9]) to provide a good fit to aggregate losses.

Parameter Estimates and Tests of Goodness of Fit

Initially we used both the Maximum Likelihood Method and the Method of Moments to estimate parameters for the approximating distributions. The parameter estimates obtained by the two methods were reasonably close to each other. Also, the distribution based on the parameters obtained by the Method of Moments provided a better fit than the one based on the parameters obtained by the Maximum Likelihood Method. For these reasons we have decided to use the Method of Moments for parameter estimates.

Once the simulated sample of aggregate losses and the approximating distributions were constructed, we tested the goodness of fit. While the usual "deviation" tests (Kolmogorov – Smirnov and χ^2-test) provide a general measurement of how close two distributions are, they can not help to determine if the distributions in question systematically differ from each other for a broad range of values, especially in the "tail". To pick up such differences, we used two tests that compare two distributions on their full range.

The Percentile Matching Test compares the values of distribution functions for two distributions at various values of the argument up to the point when the distribution functions effectively vanish. This test is the most transparent indication of where two distributions are different and by how much.

The Excess Expected Loss Cost Test compares the conditional means of two distributions in excess of different points. It tests values $E[X - x \mid X > x] * \text{Prob}\{X > x\}$. These values represent the loss cost of the layer in excess of x if X is the aggregate loss variable. The excess loss cost is the most important variable for both the ceding company and reinsurance carrier, when considering stop loss coverage, aggregate deductible coverage, and other types of aggregate reinsurance transactions.

Results and Conclusions

The four exhibits at the end of the paper document the results of our study for each of the seven scenarios described above. The exhibits show the characteristics of the frequency and severity distributions selected for each scenario, estimators for the parameters of the three approximating distributions, and the results of the two goodness-of-fit tests.

The results of the study are quite uniform: for all seven scenarios the Gamma distribution provides a much better fit than the Normal and Lognormal. In fact, both Normal and Lognormal distributions show unacceptably poor fits, but in different directions.

The Normal distribution has zero skewness and, therefore, is too light in the tail. It could probably provide a good approximation for a book of business with an extremely large expected number of claims. We have not considered such a scenario however.

In contrast, the Lognormal distribution is overskewed to the right and puts too much weight in the tail. The Lognormal approximation significantly misallocates the expected losses between excess layers. For the Lognormal approximation, the estimated loss cost for a high excess layer could be as much as 1500% of its true value.

On the other hand, the Gamma approximation performs quite well for all seven scenarios. It still is a little conservative in the tail, but not as conservative as the Lognormal. This level of conservatism varies with the skewness of the underlying severity distribution, and reaches its highest level for scenario 2 (Large Book of Business with Low Retention). When dealing with this type of aggregate distribution, one might try other alternatives.

As the general conclusion of this study, we can state that the Gamma distribution gives the best fit to aggregate losses out of the three considered alternatives for the cases considered. It can be recommended to use the Gamma as a reasonable approximation when there is no separate frequency and severity information available.

Bibliography.

1. Bickerstaff, D. R. *Automobile Collision Deductibles and Repair Cost Groups: The Lognormal Model*, PCAS LIX (1972), p. 68.

2. Cramer, H. *Collective Risk Theory*, The Jubilee Volume of Forsakringsaktiebolaget Skandia, 1955.

3. Dropkin, L. B. *Size of Loss Distributions in Workmen's Compensation Insurance*, PCAS LI (1964), p. 68.

4. Gendron, M., Crepeau H. *On the computation of the aggregate claim distribution when individual claims are Inverse Gaussian*, Insurance: Mathematics and Economics, 8:3, 1989, p. 251.

5. Gnedenko B.V., Korolev V.Yu. *Random Summation: Limit Theorems and Applications*. CRC Press, 1996.

6. Hewitt, C.C. *Distribution by Size of Risk – A Model*, PCAS LIII (1966), p. 106.

7. Hewitt, C.C. *Loss Ratio Distributions – A Model*, PCAS LIV (1967), p. 70.

8. Pentikainen, T. *On the Approximation of the total amount of claims*, ASTIN Bulletin, 9:3, 1977, p. 281.

9. Seal, H. *Approximations to risk theory's F(X, t) by means of the gamma distribution*, ASTIN Bulletin, 1977.

10. Stanard, J.N., *A Simulation Test of Prediction Errors of Loss Reserve Estimation Techniques*, PCAS LXXII (1985), p. 124.

11. Sundt, B. *Asymptotic behavior of Compound Distributions and Stop-Loss Premiums*, ASTIN Bulletin, 13:2, 1982, p. 89.

12. Venter, G. *Transformed Beta and Gamma Distributions and aggregate losses*, PCAS LXX (1983).

Scenario 1.

Frequency:	Negative Binomial	
Expected Number of Claims		50
Severity:	5 Parameter Truncated Pareto	
Expected Severity		13,511
Per Occurrence Limit		250,000

Method of Moments estimated parameters for:

Lognormal		Normal		Gamma	
Mu	13.347	Mu	691,563	Alpha	4.521
Sigma	0.447	Sigma	325,246	Beta	152,965
Mean	691,563			Mean	691,563

Percentile matching

			P(X>x)	
x	**Empirical**	Lognormal	Normal	**Gamma**
500,000	**69.36%**	69.22%	72.21%	**68.90%**
750,000	**38.06%**	34.27%	42.87%	**37.02%**
1,000,000	**16.48%**	14.72%	17.15%	**16.16%**
1,250,000	**6.16%**	6.08%	4.30%	**6.11%**
1,500,000	**1.94%**	2.53%	0.65%	**2.09%**
1,750,000	**0.62%**	1.07%	0.06%	**0.66%**
2,000,000	**0.06%**	0.47%	0.00%	**0.20%**

Expected Loss costs

		E[X-x| X>x] * P(X>x)	
Empirical	Lognormal	Normal	**Gamma**
237,751	227,011	178,648	**234,823**
104,504	100,316	43,996	**103,823**
38,636	42,118	5,123	**40,019**
12,293	17,660	245	**13,924**
3,518	7,553	4	**4,483**
870	3,323	0	**1,359**
111	1,507	0	**393**

Scenario 2.

Frequency:	Negative Binomial	
Expected Number of Claims		500
Severity:	5 Parameter Truncated Pareto	
Expected Severity		13,511
Per Occurrence Limit		250,000

Method of Moments estimated parameters for:

Lognormal		Normal		Gamma	
Mu	15.740	Mu	6,922,204	Alpha	47.462
Sigma	0.144	Sigma	1,004,786	Beta	145,849
Mean	6,922,204			Mean	6,922,204

Percentile matching

			P(X>x)	
x	**Empirical**	Lognormal	Normal	**Gamma**
6,000,000	**82.48%**	82.07%	82.06%	**81.94%**
7,000,000	**44.92%**	44.05%	46.91%	**45.01%**
8,000,000	**13.74%**	14.13%	14.17%	**14.25%**
9,000,000	**2.64%**	2.94%	1.93%	**2.63%**
9,500,000	**1.02%**	1.18%	0.52%	**0.94%**
10,000,000	**0.28%**	0.44%	0.11%	**0.30%**
10,500,000	**0.02%**	0.16%	0.02%	**0.09%**

Expected Loss costs

		E[X-x| X>x] * P(X>x)	
Empirical	Lognormal	Normal	**Gamma**
1,009,130	1,001,072	836,942	**1,007,562**
362,107	362,956	170,371	**363,947**
83,509	89,015	10,310	**83,937**
11,978	15,524	137	**12,315**
3,521	5,838	8	**4,024**
586	2,072	0	**1,192**
16	699	0	**322**

Scenario 3.

Frequency:	Negative Binomial	
Expected Number of Claims		50
Severity:	5 Parameter Truncated Pareto	
Expected Severity		18,991
Per Occurrence Limit		1,000,000

Method of Moments estimated parameters for:

Lognormal		Normal		Gamma	
Mu	13.590	Mu	958,349	Alpha	2.265
Sigma	0.605	Sigma	636,775	Beta	423,106
Mean	958,349			Mean	958,349

Percentile matching

	P(X>x)			
x	**Empirical**	Lognormal	Normal	**Gamma**
1,000,000	**38.68%**	35.47%	47.39%	**38.70%**
1,500,000	**18.28%**	14.84%	19.75%	**17.27%**
2,000,000	**6.92%**	6.44%	5.09%	**7.08%**
2,500,000	**2.82%**	2.95%	0.77%	**2.75%**
2,750,000	**1.54%**	2.04%	0.24%	**1.69%**
3,000,000	**0.92%**	1.43%	0.07%	**1.03%**
3,250,000	**0.42%**	1.01%	0.02%	**0.62%**
3,500,000	**0.28%**	0.73%	0.00%	**0.37%**

Expected Loss costs

E[X-x| X>x] * P(X>x)			
Empirical	Lognormal	Normal	**Gamma**
233,797	212,405	110,782	**228,287**
94,548	94,109	13,815	**94,254**
35,445	44,012	692	**36,798**
12,438	21,761	13	**13,826**
7,085	15,599	1	**8,382**
4,021	11,313	0	**5,062**
2,534	8,296	0	**3,029**
1,697	6,145	0	**1,807**

Scenario 4.

Frequency:	Negative Binomial	
Expected Number of Claims		500
Severity:	5 Parameter Truncated Pareto	
Expected Severity		18,991
Per Occurrence Limit		1,000,000

Method of Moments estimated parameters for:

Lognormal		Normal		Gamma	
Mu	16.065	Mu	9,685,425	Alpha	23.564
Sigma	0.204	Sigma	1,995,223	Beta	411,021
Mean	9,685,425			Mean	9,685,425

Percentile matching

	P(X>x)			
x	**Empirical**	Lognormal	Normal	**Gamma**
10,000,000	**40.96%**	39.79%	43.74%	**41.12%**
12,000,000	**12.50%**	12.44%	12.30%	**12.59%**
14,000,000	**2.18%**	2.81%	1.53%	**2.43%**
15,000,000	**0.88%**	1.23%	0.39%	**0.92%**
16,000,000	**0.36%**	0.52%	0.08%	**0.32%**
17,000,000	**0.12%**	0.21%	0.01%	**0.10%**
18,000,000	**0.06%**	0.08%	0.00%	**0.03%**

Expected Loss costs

E[X-x| X>x] * P(X>x)			
Empirical	Lognormal	Normal	**Gamma**
650,476	651,609	283,657	**654,235**
160,831	165,420	14,936	**151,977**
22,879	33,145	166	**24,231**
8,930	13,941	9	**8,544**
3,160	5,689	0	**2,799**
1,060	2,268	0	**857**
186	888	0	**247**

Scenario 5.

Frequency:	Negative Binomial	
Expected Number of Claims		20
Severity:	5 Parameter Truncated Pareto	
Expected Severity		315,640
Per Occurrence Excess Layer		$750K x $250K
Skewness		0.416

Method of Moments estimated parameters for:

Lognormal		Normal		Gamma	
Mu	15.571	Mu	6,306,951	Alpha	5.301
Sigma	0.416	Sigma	2,739,428	Beta	1,189,872
Mean	6,306,951			Mean	6,306,951

Percentile matching

	P(X>x)			
x	Empirical	Lognormal	Normal	Gamma
6,000,000	**50.26%**	46.50%	54.46%	**48.72%**
8,000,000	**24.48%**	21.77%	26.83%	**23.81%**
10,000,000	**9.70%**	9.40%	8.88%	**9.87%**
12,000,000	**3.28%**	3.96%	1.88%	**3.63%**
14,000,000	**1.00%**	1.68%	0.25%	**1.22%**
16,000,000	**0.26%**	0.72%	0.02%	**0.38%**
20,000,000	**0.04%**	0.14%	0.00%	**0.03%**

Expected Loss costs

E[X-x| X>x] * P(X>x)			
Empirical	Lognormal	Normal	Gamma
1,225,433	1,173,911	682,504	**1,218,440**
503,151	515,230	120,368	**511,426**
174,623	219,525	9,991	**191,144**
54,155	93,588	355	**65,274**
14,274	40,508	5	**20,761**
3,491	17,921	0	**6,238**
772	3,779	0	**494**

Exhibit 4

Scenario 6.

Frequency:	Negative Binomial	
Expected Number of Claims		10
Severity:	5 Parameter Truncated Pareto	
Expected Severity		1,318,316
Per Occurrence Excess Layer		$4M x $1M
Skewness		1.887

Method of Moments estimated parameters for:

Lognormal		Normal		Gamma	
Mu	16.006	Mu	12,985,319	Alpha	0.901
Sigma	0.864	Sigma	13,683,648	Beta	14,419,533
Mean	12,985,319			Mean	12,985,319

Percentile matching

	P(X>x)			
x	**Empirical**	Lognormal	Normal	**Gamma**
15,000,000	**31.80%**	27.46%	44.15%	**31.13%**
20,000,000	**22.42%**	17.57%	30.41%	**21.61%**
25,000,000	**15.32%**	11.70%	19.00%	**15.05%**
30,000,000	**10.56%**	6.06%	10.69%	**10.50%**
40,000,000	**5.36%**	4.15%	2.42%	**5.14%**
50,000,000	**2.52%**	2.32%	0.34%	**2.52%**
60,000,000	**1.10%**	1.38%	0.03%	**1.24%**

Expected Loss costs

E[X-x| X>x] * P(X>x)			
Empirical	Lognormal	Normal	**Gamma**
4,359,267	3,731,938	1,991,361	**4,315,503**
3,026,841	2,628,509	806,980	**3,011,961**
2,094,021	1,908,864	271,738	**2,105,672**
1,455,987	1,421,737	74,986	**1,473,927**
689,160	837,602	3,009	**724,182**
324,128	525,046	49	**356,764**
152,034	345,040	0	**176,096**

Scenario 7.

Frequency:	Negative Binomial	
Expected Number of Claims		10
Severity:	Lognormal	
Expected Severity		2,166,003
Per Occurrence Excess Layer		$4M x $1M
Skewness		1.190

Method of Moments estimated parameters for:

Lognormal		Normal		Gamma	
Mu	16.601	Mu	20,233,595	Alpha	1.786
Sigma	0.667	Sigma	15,141,348	Beta	11,330,681
Mean	20,233,595			Mean	20,233,595

Percentile matching

	P(X>x)			
x	**Empirical**	Lognormal	Normal	**Gamma**
20,000,000	**42.16%**	37.60%	50.62%	**40.65%**
25,000,000	**30.66%**	25.76%	37.65%	**29.44%**
30,000,000	**21.52%**	17.77%	25.95%	**20.99%**
40,000,000	**10.64%**	8.76%	9.59%	**10.31%**
50,000,000	**5.24%**	4.55%	2.47%	**4.91%**
60,000,000	**2.06%**	2.48%	0.43%	**2.29%**
70,000,000	**0.76%**	1.41%	0.05%	**1.05%**

Expected Loss costs

E[X-x| X>x] * P(X>x)			
Empirical	Lognormal	Normal	**Gamma**
5,984,377	5,371,757	3,116,920	**5,861,977**
4,167,564	3,806,611	1,488,582	**4,122,201**
2,877,740	2,731,389	615,453	**2,872,036**
1,299,737	1,462,583	65,324	**1,364,938**
546,755	822,414	3,471	**635,179**
204,979	482,606	88	**291,051**
69,923	293,789	1	**131,796**