METHODS FOR FITTING DISTRIBUTIONS TO INSURANCE LOSS DATA

CHARLES C. HEWITT, JR. AND BENJAMIN LEFKOWITZ

SUMMARY

The methods described in this paper can be used to fit five types of distribution to loss data: gamma, log-gamma, log-normal, gamma + log-gamma, and gamma + log-normal. The paper also discusses applications of the fitted distributions to estimation problems; e.g., computing the effects of inflation on the loss portion of deductible credits and increased limits charges, and determining changes to claim frequencies and severities brought about by changes in deductibles and limits. A computer program carries out all the calculations.

INTRODUCTION

Casualty actuaries frequently wish to extract information from insurance loss data. Generally, an actuary will group individual losses by size of loss and then fit a continuous positive distribution to the aggregated data. In this way, he can characterize the universe from which the sample was selected. For example, a distribution fitted to one month's losses could be used to characterize the distribution of annual losses.

Bickerstaff and Dropkin have shown that the log-normal distribution closely approximates certain types of homogeneous loss data [1], [2]. Hewitt showed that two other positive distributions, the gamma and the log-gamma, also give good fits [3], [4]. Used alone, each of these distributions assumes that the observed losses are generated by a single underlying process. This may not always be the case. For example, a sample of observed losses may contain some that involved litigation and others that did not. In this situation, a single distribution may not fit the aggregate data as well as a combination of two (or more) distributions added together.[1] Herein, such combinations are called compound distributions. This paper describes algorithms for fitting two particular compound distributions, gamma + log-gamma, and gamma + log-normal, and three simple distributions: gamma, log-gamma and log-normal.

[1] Each component distribution has its own form, i.e., gamma, log-gamma or log-normal, and its own parameters, e.g., mean, variance. If the proportion of losses (either claim count or amount) in one distribution is P, then the proportion in the second distribution is 1-P.

The authors do not claim that all insurance loss data can be fitted by the methods described below or, in fact, by any analytical methods. However, after many years' experience, we are convinced that these methods will produce useful results for most practical problems.

ALGORITHMS FOR FITTING DISTRIBUTIONS

The usual method for fitting a distribution to observations involves estimating the distribution's parameters or moments from a sample of actual loss frequencies, and then using those parameters to compute the distribution's densities, i.e., its theoretical loss frequencies. The normal distribution's parameters, for example, are the mean and variance. Given their values, one can obtain loss frequencies by consulting tables of the normal distribution.

This method cannot be applied to a compound distribution because its parameters are not directly computable from the sample observations. Instead, an iterative procedure must be used to approximate them. The procedure, described in Appendix A, repeats the following steps:

1. Split the data between the two distributions.
2. Estimate each distribution's parameters.
3. Fit the distributions.
4. Compare the computed frequencies to the actual frequencies.

Each iteration attempts to adjust the data split and distribution fits so as to improve the correspondence between actual and theoretical frequencies. There is no guarantee that the correspondence will improve each iteration, or that the best fit will be obtained after a finite number of trials. Generally, it takes fewer than ten iterations to reach stability, by which we mean that the mean of the fitted compound distribution changes little from one iteration to the next.

A problem common to fitting any distribution—single or compound—to aggregate loss data is the location of the "mass-point" of each loss interval.[2] A single value must represent all observations within an interval; very often, the interval midpoint is used for this purpose. The choice of mass-point influences the distribution's parameters and hence the quality of the resulting fit. In most distributions arising in casualty insurance applications, losses are skewed toward the upper boundary of their intervals. However, in the normal distribution and distributions like it, losses are skewed toward the lower boundary

[2] I.e., the first two moments, loss amount and loss amount squared weighted by frequency.

in intervals lying to the left of the mode, and towards the upper boundary in intervals lying to the right of the mode.

The algorithms described in Appendix A include calculations that correct the possible bias introduced by the exclusive use of the interval midpoint. The mass-points of the amount-of-loss and the square of the amount-of-loss must be adjusted at each successive iteration, because they are used to compute the moments needed to estimate the parameters of the individual distributions. In most instances, the correction substantially affects only the uppermost and lowermost intervals.

EXAMPLE

The quality of compound distribution fits can be illustrated by an example. Table 1 contains automobile bodily injury loss data along with log-normal and gamma + log-normal fits to the data.[3]

As can be seen, the compound distribution, gamma + log-gamma, is superior to the log-normal alone because it better approximates the frequency of low value and high value losses.

The goodness-of-fit can be measured by the Chi-square statistic, χ^2 [5]. The difference between actual and log-normal distribution has a $\chi^2 = 28.7$ with 15 degrees of freedom. This means that there is about a 2.5% chance the log-normal explains the data. The difference between actual and the gamma + log-gamma distribution has $\chi^2 = 3.5$ with 12 degrees of freedom. There is only about a 1% chance that the agreement could arise by chance alone.

DISTRIBUTION TABLE

A distribution fitted to the number of losses can be used to compute the cumulative dollars of loss and the deductible credit or "buy back." A deductible credit is the proportional loss reduction caused by imposing a deductible.[4] Readers may be more familiar with the term "loss elimination ratio" [6]. Notice that the limits in Table 2 are not the same as the limits in Table 1.

[3] The data is from a 1969 Department of Transportation study of automobile injuries. It shows general damages on serious injury cases in California.

[4] Purchasers of automobile collision insurance understand that they pay something less than full cost when they are willing to pay the first $50, $100, etc. of any loss.

To illustrate the use of the distribution table, refer to the limit of $10,000. The table shows that losses of $10,000 or less account for 82.15% of the losses by number but only 12.64% of the loss dollars. These loss dollars plus the first $10,000 on the 17.85% of the losses which exceed $10,000 account for 25.88% of all loss dollars. Put another way, for a policy with no limit but with a $10,000 deductible, the table indicates a (loss dollar) credit of 25.88%.

The distribution table also can be used to determine the loss portion of charges for increased limits. The formula is:

$$Charge = \frac{\begin{array}{c}(deductible\ credit\ for\ the\ increased\ limit\\ -\ deductible\ credit\ for\ the\ basic\ limit)\end{array}}{deductible\ credit\ for\ the\ basic\ limit}$$

For example, to determine the loss portion of the charge for a layer of coverage between $10,000 and $100,000, compute

$$\frac{.6029 - .2588}{.2588} = 1.33$$

i.e., the loss portion of the increased limits charge is 133% of basic limits losses. (The increased limits factor is, of course, 2.33 if expenses remain proportionate.)

EFFECT OF INFLATION

It is possible to construct distribution tables that take into account the effect of inflation on loss settlements, thereby allowing actuaries to answer questions of the following type: What would happen to loss costs if future losses were distributed in a manner similar to past losses, but settlement costs were 100% higher?

The answer is found in Table 3 which uses data from Table 2 but assumes a 100% inflation rate.[5] Note the contrast between the two tables. Only 71.09% of the inflated losses (Table 3) are less than $10,000 and they account for only 6.83% of the loss dollars, whereas 82.15% of the uninflated losses (Table 2) are less than $10,000 and they account for 12.64% of the loss dollars. Similarly, under 100% inflation, the deductible credit decreases from 25.88% to 17.54%. Inflation causes the loss portion of the increased limits charge for $100,000 limits to rise from 133% to 184%.

[5] I.e., a 100% increase in the value of each loss from the settlement date to the date for which losses are being used.

The algorithm for measuring the effect of inflation is shown in Appendix B.

FITTING TRUNCATED DISTRIBUTIONS

Because insureds use deductibles or retentions in many lines of casualty insurance, data collected for use by the actuary may be incomplete in that nothing is available for losses below some fixed dollar amount (such as $100). The flexibility of the gamma, log-gamma and log-normal distributions is such that their moments (where they exist) also are distributed respectively as gamma, log-gamma and log-normal [1], [3], [4]. The missing portion of a truncated distribution may contain many losses, but often the missing loss dollars do not amount to a great deal. Therefore, it is suggested that the actuary initially use the dollar amount of losses to fit the distribution. The number of omitted losses resulting from the use of the deductible can then be estimated and used to fit the distribution of the number of cases.

The method for obtaining the parameters of the distribution of cases after estimating the parameters of the distribution of loss dollars is shown in Appendix C.

For large losses, the data collected by the actuary may be inaccurate because of policy loss limits. Here there are no missing cases, but the arbitrary limit obscures the true (unbounded) value of these larger losses.

This problem can be solved by calculating the ''true'' value[6] for the mass-point of the uppermost interval in a manner that is independent of the reported values of the larger losses. The actuary selects the lower limit of the uppermost interval for the data to be fitted so that all cases which may have been arbitrarily valued fall into the uppermost interval. Then, by fitting the number of cases and not their dollar value, the effect of policy limits is ignored. The method for calculating an interval mass-point is explained in Appendix A.

Example: Suppose the raw data in Table 3 contains no losses under $100, because of the existence of a $100 deductible. One could fit a distribution to this data under the assumption that the interval $0-100 is empty. (This is equivalent to assuming that the cumulative frequency of loss dollars up to the $250 limit equals the frequency of loss dollars in the interval $100-250.) The error introduced by this assumption (0 cf. .0002), is less than the error intro-

[6] As opposed to some arbitrary, a priori assumption, guess or inaccuracy in the raw data itself.

duced by postulating that there are no claims in the \$0-100 loss interval (0 cf. .1433).

After fitting a distribution to raw loss dollars, one can deduce the parameters of the distribution of claim counts as shown in Appendix C. These latter parameters can be used to calculate the hypothetical proportion of claims under \$100. That proportion can be used to fit the augmented claim count distribution.

RELATIVE FREQUENCIES AND SEVERITIES

Changes in claim frequencies and severities can be determined when deductibles (or retentions) and limits are changed. Assume an insured has a retention of \$1,000 per claim, then what are the relative frequencies when the retention is increased to \$5,000 per claim? From Table 2,

$$\frac{1 - .7109}{1 - .3862} = .471$$

the new frequency is 47.1% of the old frequency.

The relative severities under unlimited coverage are:

$$\frac{\dfrac{1 - .1754}{1 - .7109}}{\dfrac{1 - .0545}{1 - .3862}} = \frac{2.852}{1.540} = 1.852$$

or the new severity is 185.2% of the old severity.

Suppose limits are increased from \$10,000 per claim to \$100,000 per claim. What happens to the relative severities? From Table 2,

$$\frac{.6029}{.2588} = 2.330$$

that is, the new severity will be 233.0% of the old severity.

Suppose a reinsurer has data collected on the basis of a retention of \$10,000 and a limit of \$100,000, and suppose loss costs have increased 100% since the period for which the data was collected. How will the relative frequencies and severities change? From Tables 2 and 3, the relative frequencies are:

$$\frac{1 - .7109}{1 - .8215} = 1.620$$

that is, there will be 62.0% more claims at the same retention because of inflation. The relative severities are:

$$\frac{\dfrac{.4981 - .1754}{1 - .7109}}{\dfrac{.6029 - .2588}{1 - .8215}} = \frac{1.116}{1.928} = .579$$

The new severity is 57.9% of the old severity.

PROGRAM HEWITZ

All computations described in this paper were performed with a computer program called HEWITZ.[7] This program fits five distributions to input data: gamma, log-gamma, log-normal, gamma + log-gamma and gamma + log-normal.

HEWITZ has the following characteristics and capabilities:

1. The user can select different intervals for the input data and the output distribution table.
2. The user can halt the iterative algorithms in one of several ways, but usually by specifying the maximum number of iterations.
3. The user can create a wholly new distribution by presetting any distribution's parameters.
4. The program computes the Chi-square goodness-of-fit statistic.

[7] Program HEWITZ is written in G-Level Fortran IV, and has been implemented on an IBM 370/158 computer. The program occupies about 100k bytes of core. The program took ten seconds to fit all five distributions to the loss data described earlier.

TABLE 1

Automobile Bodily Injury Loss Data

	Number of Cases		
Loss Amount ($)	Actual	Log-Normal	Gamma + Log-Gamma
1– 50	27	18	27
51– 100	4	10	4
101– 150	1	8	2
151– 200	2	6	2
201– 250	3	5	3
251– 300	4	4	3
301– 400	5	7	6
401– 500	6	6	5
501– 750	13	12	12
751–1,000	8	8	10
1,001–1,500	16	12	15
1,501–2,000	8	9	11
2,001–2,500	11	7	9
2,501–3,000	6	5	7
3,001–4,000	12	8	11
4,001–5,000	9	6	8
5,001–7,500	14	10	13
Over 7,500	40	48	41
TOTAL	189	189	189

TABLE 2

Distribution Table

Upper Limit of Loss Amount ($)	Cumulative Frequency of Cases	Cumulative Frequency of Dollars	Deductible Credit
100	.1623	.0003	.0065
250	.2008	.0008	.0156
500	.2724	.0028	.0298
750	.3344	.0057	.0427
1,000	.3862	.0090	.0545
2,000	.5271	.0242	.0943
2,500	.5739	.0320	.1109
5,000	.7109	.0683	.1754
7,500	.7795	.0995	.2221
10,000	.8215	.1264	.2588
20,000	.8992	.2075	.3570
25,000	.9176	.2380	.3908
50,000	.9582	.3432	.4981
100,000	.9803	.4570	.6029
250,000	.9934	.6047	.7263
500,000	.9973	.7040	.8028
1,000,000	.9990	.7873	.8633
Unlimited	1.0000	1.0000	1.0000

TABLE 3

Distribution Table with 100% Inflation

Upper Limit of Loss Amount ($)	Cumulative Frequency of Cases	Cumulative Frequency of Dollars	Deductible Credit
100	.1433	.0002	.0034
250	.1676	.0004	.0081
500	.2008	.0008	.0156
750	.2373	.0017	.0229
1,000	.2724	.0028	.0298
2,000	.3862	.0090	.0545
2,500	.4298	.0126	.0655
5,000	.5739	.0320	.1109
7,500	.6564	.0508	.1463
10,000	.7109	.0683	.1754
20,000	.8215	.1264	.2588
25,000	.8501	.1501	.2891
50,000	.9176	.2380	.3908
100,000	.9582	.3432	.4981
250,000	.9848	.4939	.6350
500,000	.9934	.6047	.7263
1,000,000	.9973	.7040	.8028
Unlimited	1.0000	1.0000	1.0000

APPENDIX A

The following symbols are used:

Symbol	Meaning
$a+1$	Scale parameter of the log-gamma distribution
A	Scale parameter of the gamma distribution
c_i	Actual number of cases in the i-th loss (claim) interval
C_i	Computed number of cases in the i-th loss (claim) interval
DC_j	Cumulative proportion of cases in the first j intervals of the distribution table
DD_j	Cumulative proportion of loss dollars in the first j intervals of the distribution table
DE_j	Deductible credit for the first j intervals of the distribution table
$E_G(X)$	Mean of X, gamma
$E_G(X^2)$	Mean of X^2, gamma
$E_L(x)$	Mean of x, log-gamma
$E_L(x^2)$	Mean of x^2, log-gamma
$\tilde{E}_L(x)$	Estimate of $E_L(x)$ used in the first iteration of the gamma + log-gamma algorithm
$E(X)$	Mean of the compound distribution
$E_L(X)$	Mean of X, log-gamma. Equals $\left(\frac{a+1}{a}\right)^{p+1}$
$E_N(x)$	Mean of x, log-normal
$E_N(x^2)$	Mean of x^2, log-normal
$E_N(X)$	Mean of X, log-normal. Equals $exp\,(E_N(x) + \sigma^2_N/2)$
f_i	Relative frequency of cases in the i-th loss (claim) interval
F_i	Cumulative of f_i
FG_i	Cumulative frequency of gamma distribution in the i-th interval
$H(j)$	Proportion of claims in j-th loss interval after allowing for inflation
$I(y,w)$	Value of the incomplete gamma function ratio for the variable y and the parameter w. This is the cumulative density of the ratio up to and including y
N	Index of last loss (claim) interval
P	Proportion of total claims in log-gamma or log-normal distribution
$p+1$	Shape parameter of the log-gamma distribution
Q	Proportion of total claims in gamma distribution. Equals $1-P$
R	Shape parameter of the gamma distribution

x_i	$Log_e\ X_i$
xh_i	$Log_e\ XH_i$
X	Value of loss
X_i	Midpoint of the i-th loss (claim) interval
XH_i	Upper boundary of the i-th loss (claim) interval
XL_i	Lower boundary of the i-th loss (claim) interval
y_j	$Log_e\ Y_j$
Y_j	Upper boundary of the j-th distribution table interval
λ	Inflation factor. Equals one plus the rate of inflation expressed as a fraction
$\Phi(z_i)$	Normal curve cumulative density from $-\infty$ to z_i
σ^2G	Variance of X in the gamma
σ^2L	Variance of X in log-gamma
$\bar{\sigma}_L$	Estimate of σ_L used in the first iteration of the gamma + log-gamma algorithm
σ^2_N	Variance of x in log-normal

1) *Gamma Distribution*

The gamma distribution, actually the incomplete gamma function ratio, is the cumulative density function:

$$I\,(v_i,\ R-1)\ =\ \begin{cases} 0\ , & i=0 \\ \dfrac{1}{\Gamma(R)} \displaystyle\int_0^{v_i\sqrt{R}} y^{(R-1)}\,e^{-y}\,dy, & 0<i<N \\ 1\ , & i=N \end{cases}$$

where

$$v_i\ =\ A\ \bullet\ XH_{i_i}\ /\ \sqrt{R}$$

In the k-th iteration, the distribution parameters

A - the scale parameter

R - the shape parameter

are estimated as follows:

$$A\ =\ E_G(X)\ /\ \sigma^2_G,\quad R\ =\ A\ \bullet\ E_G(X)$$

The k-th iteration values of $E_G(X)$ and σ_G^2 are:

$$E_G(X) = \Sigma f_i X_i \,, \quad \sigma_G^2 = \Sigma f_i X_i^2 - \overline{E_G(X)}^2$$

Initially, $X_i = \tfrac{1}{2}(XH_i + XL_i)$. After the k-th iteration, the repaired interval midpoints are:

$$X_i = \frac{g_i' E_G(X)}{g_i^*}$$

where $E_G(X)$ is the mean of the gamma computed in the $(k\text{-}1)$st iteration, and g_i^* is the proportion of cases in the i-th interval computed using the gamma fitted in the k-th iteration.

$$g_i^* = I(v_i, R-1) - I(v_{i-1}, R-1)$$

The quantity g_i' is the proportion of dollar loss in the i-th interval computed using the gamma fitted in the k-th iteration

$$g_i' = I(v_i', R) - I(v_{i-1}', R)$$
$$v_i' = A \bullet XH_i / \sqrt{R+1}$$

In the k-th iteration, the repaired values of X_i^2

$$X_i^2 = \frac{g_i'' \bullet E_G(X^2)}{g_i^*}$$

where $E_G(X^2)$ is the average of the squared midpoints computed in the $(k\text{-}1)$ st iteration, and g_i'' is the proportion of the X^2-value in the i-th interval computed using the gamma fitted in the k-th iteration

$$g_i'' = I(v_i'', R-1) - I(v_{i-1}'', R+1)$$
$$v_i'' = A \bullet XH_i / \sqrt{R+2}$$

The number of claims in the i-th interval computed using the fitted gamma distribution is:

$$C_i = [C^* \{ I(v_i, R-1) - I(v_{i-1}, R-1)\} + .5]$$
$$C^* = \Sigma c_i$$

The square brackets represent the greatest integer function.

2) *Log-gamma Distribution*

The log-gamma distribution, actually the incomplete gamma function ratio applied to the logarithms of loss data, is the cumulative density function:

$$I(u_i, p) = \begin{cases} 0 \ , \ i=0 \\ \dfrac{1}{\Gamma(p+1)} \displaystyle\int_0^{u_i\sqrt{p+1}} y^p \ e^{-y}dy, \quad 0<i<N \\ 1 \ , \ i=N \end{cases}$$

where

$$u_i = (a+1) \ \bullet \ xh_i / \sqrt{p+1}$$

In the k-th iteration, the distribution parameters

$a+1$ — *the scale parameter*

$p+1$ — *the shape parameter*

are estimated as follows:

$$a+1 = E_L(x) / \sigma_L^2, \quad p+1 = max\{0, (a+1) \ \bullet \ E_L(x)\}$$

The k-th iteration values of $E_L(x)$ and σ_L^2 are:

$$E_L(x) = \Sigma f_i x_i \ , \quad \sigma_L^2 = \Sigma f_i x_i^2 - \overline{E_L(x)}^2$$

Initially $x_i = log_e\{½ (XH_i + XL_i)\}$. After the k-th iteration, the repaired interval midpoints are:

$$x_i = E_L(x) \ \bullet \ f_i' / f_i^* \ ,$$

where $E_L(x)$ is the mean of X in the log-gamma computed in the $(k\text{-}1)$st iteration, and f_i^* is the proportion of cases in the i-th interval computed using the log-gamma fitted in the k-th iteration.

$$f_i^* = I(u_i, p) - I(u_{i-1}, p).$$

The quantity f_i' is the proportion of x in the i-th interval

$$f_i' = I(u'_i, p+1) - I(u'_{i-1}, p+1)$$

$$u''_i = (a+1) \ \bullet \ xh_i / \sqrt{p+2}$$

In the k-th iteration, the repaired values of x_i^2 are:

$$x_i^2 = \frac{f_i'' \bullet E_L(x^2)}{f_i^*}$$

where $E_L(x^2)$ is the average squared log of the midpoints computed in the (k-1)st iteration, and f_i'' is the proportion of the x-value in the i-th interval computed using the log-gamma fitted in the k-th iteration.

$$f_i'' = I(u_i'', p+2) - I(u_{i-1}'', p+2)$$
$$u_i'' = (a+1) \bullet xh_i / \sqrt{p+3}$$

The number of claims in the i-th interval computed using the fitted log-gamma distribution is

$$C_i = [C^* \{I(u_i,p) - I(u_{i-1},p)\} + .5]$$
$$C^* = \Sigma\, c_i$$

The square brackets represent the greatest integer function.

3) Log-normal Distribution

The cumulative frequency of the log-normal distribution is:

$$\Phi\ (z_i) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_i} e^{-\frac{1}{2} z_i^2}\ dz_i$$

where

$$z_i = \{ xh_i - E_N(x) \} / \sigma_N$$

In the k-th iteration the distribution parameters

$E_N(x)$ – the mean of the log-normal
σ_N^2 – the variance of the log-normal

are estimated as follows:

$$E_N(x) = \Sigma\, f_i\, x_i$$
$$\sigma_N^2 = \Sigma\, f_i\, x_i^2 - \overline{E_N(x)}^2$$

Initially $x_i = log_e \{ \frac{1}{2} (XH_i + XL_i) \}$. After the k-th iteration, the repaired interval midpoints are:

$$x_i = f_i' / f_i^{i*}$$

where f_i^* is the proportion of cases in the i-th interval.

$$f_i^* = \Phi(z_i) - \Phi(z_{i-1}),\ \Phi(z_0) = 0,\ \Phi(z_N) = 1$$

$$f_i' = \{E_N(x)\,\Phi(z_i) - \frac{\sigma_N}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2}\} - \{E_N(x)\,\Phi(z_{i-1}) - \frac{\sigma_N}{\sqrt{2\pi}} e^{-\frac{1}{2} z_{i-1}^2}\}$$

In the k-th iteration, the repaired values of x_i^2 are:

$$x_i^2 = f_i''/f_i^*$$

where

$$f_i'' = \{(\overline{E_N(x)}^2 + \sigma_N^2)\,\Phi(z_i) - (E_N(x) + x_i) \bullet \frac{\sigma_N}{\sqrt{2\pi}} e^{-\frac{1}{2} z_i^2}\}$$
$$- \{(\overline{E_N(x)}^2 + \sigma_N^2)\,\Phi(z_{i-1}) - (E_N(x) + x_{i-1}) \bullet \frac{\sigma_N}{\sqrt{2\pi}} e^{-\frac{1}{2} z^2_{i-1}}\}$$

The number of claims in the i-th interval computed using the fitted log-normal distribution is:

$$C_i = [C^*\{\Phi(z_i) - \Phi(z_{i-1})\} + .5]$$
$$C^* = \Sigma\, c_i$$

The square brackets represent the greatest integer function.

4) Gamma + Log-gamma

Fitting a compound distribution is a trial-and-error process. Initially, the log-gamma distribution is fitted to the data. Generally this will result in fewer computed claimants in the lower loss intervals than are actually there. The gamma distribution is fitted to the excess claimants. These calculations "split" the data between the gamma and log-gamma distributions. The compound distribution is the weighted sum of the two distributions, the weights being:

P — the proportion of total claims in the log-gamma

$Q = (1\text{-}P)$ — the proportion of total claims in the gamma.

As before, the interval midpoints must be repaired to recognize intra-interval skewness.

The gamma plus log-gamma (*GPLG*) distribution is the cumulative density function:

$$GPLG_i = \begin{cases} 0 \ , \ i=0 \\ Q \bullet I(v_i, R-1) + P \bullet I(u_i, p) \ , \ 0 < i < N \\ 1 \ , \ i=N \end{cases}$$

where

$$v_i = A \bullet XH_i / \sqrt{R}$$
$$u_i = (a+1) \bullet xh_i / \sqrt{p+1}$$

Two methods are used to estimate the log-gamma distribution parameters. One applies to the first iteration only, the other to the remaining iterations. On the first iteration:

$$a = \frac{\bar{E}_L(x)}{\bar{\sigma}_L{}^2} \qquad p+1 = a \bullet \bar{E}_L(x)$$

where

$$\bar{E}_L(x) = \frac{\Sigma f_i X_i \bullet x_i}{\Sigma f_i X_i}$$

$$\bar{\sigma}_L{}^2 = \frac{\Sigma f_i X_i \bullet x_i^2}{\Sigma f_i X_i} - \{\bar{E}_L(x)\}^2$$

The mean of the compound distribution is:

$$E(X) = P \bullet E_L(X) + Q \bullet E_G(X)$$

where

$$E_L(X) = \left(\tfrac{a+1}{a}\right)^{(p+1)} \text{ and } E_G(X) = \frac{R}{A}$$

The formulas for repairing x_i, x_i^2, X_i and X_i^2 are shown earlier.

The number of claims in the i-th interval is:

$$C_i = [C^* \{GPLG_i - GPLG_{i-1}\} + .5]$$

$$C^* = \Sigma c_i$$

The square brackets represent the greatest integer function.

5) Gamma + Log-normal

Fitting the gamma + log-normal (*GPLN*) distribution is analogous to fitting the gamma + log-gamma distribution.

The *GPLN* distribution is the cumulative density function:

$$GPLN_i = \begin{cases} 0 \ , \ i=0 \\ Q \bullet I(v_i, R-1) \ + \ P \bullet \Phi(z_i) \ , \ 0<i<N \\ 1 \ , \ i = N \end{cases}$$

$$v_i = A \bullet XH_i \,/\, \sqrt{R} \ , \ z_i = \{xh_i - E_N(x)\} \,/\, \sigma_N$$

The formulas for estimating the parameters of the gamma and log-normal, A, R, $E_N(x)$ and σ_N^2, are the ones used to fit the gamma and the long-normal separately. Similarly, the procedure used to split the total loss data between the two distributions is the one used in fitting the gamma + log-gamma, but with the log-normal distribution substituted for the log-gamma distribution where appropriate.

In all other iterations:

$$a+1 = \frac{E_L(x)}{\sigma_L^2} \ , \quad p+1 = max\,\{0, (a+1) \bullet E_L(x)\}$$

The log-gamma is fitted to the intervals, yielding theoretical cumulative frequencies:

$$DL_i = I(u_i, p)$$

The next step in the calculation splits the total distribution between the gamma and the log-gamma, and estimates the proportion of total claims in each distribution. The calculation consists of the following steps:

Determine whether the data can be split

1. Set the split proportion estimate $P' = 1$, and the interval index $j=0$
2. Compute proportion of total claims in first interval of gamma $G_1 = F_1 - DL_1$
3. If $G_1 \leq 0$, then the gamma distribution cannot be fitted to the data. This can happen when small valued losses go unreported.

Split the data

4. Increase interval index j by one

5. Compute approximate proportion of total claims in the log-gamma and in the gamma

   $H_j = P' \bullet DL_j$

   $G_j = F_j - H_j$

6. Compute estimate of P:

   $P = 1 - G_{j+1}$ (Initially $G_2 = G_1$; see Step 2)

7. Compute proportion of total claims in the j-th and $(j+1)$st intervals of the log-gamma, and the gamma

   $H_j = P \bullet DL_j$

   $G_j = F_j - H_j$

   $H_{j+1} = P \bullet DL_{j+1}$

   $G_{j+1} = F_{j+1} - H_{j+1}$

8. Compute the difference of successive intervals of the gamma

   $\Delta = G_{j+1} - G_j$

9. If $\Delta < 0$, go to Step 10, otherwise set $P' = P$ and return to Step 4.

Compute Frequencies of the Gamma

10.

$$G_j = \begin{cases} F_i - P \bullet DL_i \ , & i < j \\ 1 \ , & i \geqslant j \end{cases}$$

After the data has been split, the parameters of the gamma distribution are estimated as follows:

$$A = E_G(X) \,/\, \sigma_G^2 \quad , \quad R = A \cdot E_G(X)$$

where

$$E_G(X) = \Sigma\, G_i \cdot X_i \ , \quad \sigma_G^2 = \Sigma\, G_i \cdot X_i^2 - \overline{E_G(X)}^2$$

The formulas for repairing x_i, $x^2{}_i$, X_i^2 *are shown earlier.*

The number of claims in the i-th interval is

$$C_i = [C^* \{GPLN_i - GPLN_{i-1}\} + .5]$$

$$C^* = \Sigma\, c_i$$

The square brackets represent the greatest integer function.

The mean of the compound distribution is:

$$E(X) = P \cdot E_N(X) + Q \cdot E_G(X)$$

where

$$E_N(X) = exp\ (E_N(x) + \frac{\sigma_N^2}{2}\ ) \text{ and } E_G\,(X) = \frac{R}{A}$$

APPENDIX B

EFFECT OF INFLATION

In the formulas given below the subscripts 1, 2 and 3 refer to the gamma, log-gamma and log-normal distributions respectively, and the index j runs over the Distribution Table loss intervals. The parameter λ, is one plus the rate of inflation expressed as a fraction.

$$F_1(j) = I(v_j, R - 1)\ , \quad v_j = \frac{Y_j}{\lambda} \cdot \frac{A}{\sqrt{R}}$$

$$F_2(j) = I(u_j, p) \quad , \quad u_j = \frac{a + 1}{\sqrt{p + 1}}\ log\ \frac{Y_j}{\lambda}$$

$$F_3(j) = \Phi(z_j) \quad , \quad z_j = log\ \frac{Y_j}{\lambda} - E_N(x)\,/\sigma_N$$

$$G_1(j) = I(v_j^*, R) \quad , \quad v_j^* = \frac{Y_j}{\lambda} \cdot \frac{A}{\sqrt{R+1}}$$

$$G_2(j) = I(u^*, p) \quad , \quad u_j^* = \frac{a}{\sqrt{p+1}} \cdot \log \frac{Y_j}{\lambda}$$

$$G_3(j) = \Phi(z_j^*) \quad , \quad z_j^* = z_j - \sigma_N$$

$$H_1(j) = G_1(j) + \frac{Y_j}{\lambda} \{1 - F_1(j)\} / E_G(X)$$

$$H_2(j) = G_2(j) + \log \frac{Y_j}{\lambda} \{1 - F_2(j)\} / E_L(x)$$

$$H_3(j) = G_3(j) + \log \frac{Y_j}{\lambda} \{1 - F_3(j)\} / E_N(x)$$

For the two compound distributions, we get

Gamma + log-gamma

$$F(j) = Q \cdot F_1(j) + P \cdot F_2(j)$$

$$G(j) = (1 - S) \cdot G_1(j) + S \cdot G_2(j), \text{ where}$$

$$S = \frac{P \cdot E_L(X)}{E(X)} \quad \text{and} \quad E(X) = P \cdot E_L(X) + Q \cdot E_G(X)$$

$$H(j) = G(j) + \frac{Y_j}{\lambda} \{1 - F(j)\} / E(X)$$

Gamma + log-normal

$$F(j) = Q \cdot F_1(j) + P \cdot F_3(j)$$

$$G(j) = (1 - S) \cdot G_1(j) + S \cdot G_3(j), \text{ where}$$

$$S = \frac{P \cdot E_L(X)}{E(X)} \quad \text{and} \quad E(X) = P \cdot E_N(X) + Q \cdot E_G(X)$$

$$H(j) = G(j) + \frac{Y_j}{\lambda} \{1 - F(j)\} / E(X)$$

APPENDIX C

Parameter	*Fit on $*	*Fit on #*
	Gamma	
Shape	$R + 1$	R
Scale	A	A
	Log-Gamma	
Shape	$p + 1$	$p + 1$
Scale	a	$a + 1$
	Log-Normal	
Mean	$E_N(X)$	$E_N(X) - \sigma_N^2$
Variance	σ_N^2	σ_N^2

Example of the use of this Appendix: If a gamma distribution with parameters R and A is fitted to numbers of claim counts (#), then the parameters of the distribution of loss amounts ($) are $R + 1$ and A respectively.

REFERENCES

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

[2] Dropkin, L. B. "Size of Loss Distributions in Workmen's Compensation Insurance," *PCAS* LI (1964), p. 198

[3] Hewitt, C. C. "Distribution by Size of Risk—A Model," *PCAS* LIII (1966), p. 106

[4] Hewitt, C.C. "Loss Ratio Distributions—A Model," *PCAS* LIV (1967), p. 70

[5] Mood, A. M. and Graybill, F. A. *Introduction to the Theory of Statistics*, McGraw-Hill Book Company (1963), p. 308

[6] Walters, Michael A. "Homeowners Insurance Ratemaking," *PCAS* LXI (1974), p. 23