Toward a unified approach to fitting loss models[*]

Jacques Rioux and Stuart Klugman

February 6, 2004

Abstract

There are four components to fitting models — selecting a set of candidate distributions, estimating parameters, evaluating the appropriateness of a model, and determining which member fits best. It is important to have the candidate set be small to avoid overfitting. Finite mixture models using a small number of base distributions provide an ideal set. Because actuaries fit models for a variety of situations, particularly with regard to data modifications, it is useful to have a single approach. Though not optimal or exact for a particular model or data structure, the method should be reasonable for most all settings. Such a method is proposed in this article. To aid the user, a computer program implementing these models and techniques is provided.

1 Introduction

Actuaries have been fitting models to data for most of the profession's existence (and maybe even before). Through the years, a progression of techniques has taken place, from graphical smoothing, to methods of moments, to maximum likelihood. At the same time the number of models available has increased dramatically. In addition, the number of diagnostic tools has increased. The combination of possibilities can be overwhelming. This forces

[*] This is a draft paper of a paper being considered for publication. No part may be copied or quoted without the authors' permission.

the actuary to make choices. The purpose of this paper is to encourage actuaries to make a particular set of choices when fitting parametric distributions to data.

One approach could be distribution-by-distribution. When a particular model has been identified, there may be considerable literature available to guide the actuary toward a method that is best for that model. For example, Kleiber and Kotz [5] offer an adjustment to the maximum likelihood estimator for the single parameter Pareto distribution that makes it unbiased and slightly reduces the variance. However, finding unique methods for each model may be difficult because not all models have been as well researched as others.

Our approach is to offer a single estimation method that can be applied in most all circumstances. This is accompanied by a limited set of graphical and statistical tools. As a result, the process may not be optimal for any one situation. The main advantage is that if an actuary becomes adept at our approach, it can be quickly applied in a variety of situations with regard to the nature of the data and the model selected.

There are four components of the model fitting and selection process that will be discussed in turn in the following sections. Throughout, two examples will be used to illustrate the process. The components are:

1. A set of probability models. A small, yet flexible, set of probability models will both lessen the workload and prevent overfitting.[1]

2. A parameter estimation technique. Maximum likelihood estimation will be used throughout. Its benefits and implementation have been thoroughly discussed elsewhere and so will not be covered here.

3. A method for evaluating the quality of a given model. Several graphical techniques and hypothesis tests will be offered. All compare the model to the data. A challenge is describing the data when it has been grouped, truncated, or censored.

4. A method for selecting a model from the list in Item 1.

[1] In this paper, overfitting refers to any process where a large number of possibilities are considered. This could mean an excessive number of parameters or an excessive number of models considered. In both cases there is a tendency for the selected model to be more faithful to the data than to the population.

It should be noted that the above list still provides some flexibility for the model builder with regard to which graphs and tests to emphasize. This allows the experienced actuary to make use of personal knowledge and preferences to aid in the determination of the best model.

The authors have also made software available that implements the ideas in this paper.

2 A collection of models

The collection proposed here may not satisfy everyone. It has been selected with the following goals in mind.

- It should contain a small number of models. Many are available. For example, Appendix A of Klugman, Panjer, and Willmot [7] lists 22 different distributions while Appendix C of [5] lists 14, with the union containing 26. All of these have been used for loss modeling. The comprehensive volumes by Johnson, Kotz, and Balakrishnan ([2] and [3]) contain dozens of distributions, though many are clearly not appropriate. There is a great danger of overfitting when too many models (especially when many are inappropriate for the phenomenon being modeled) are considered. That is, it becomes more likely that the model is matching the data than that it is matching the population that produced the data.
- It should include the possibility of a non-zero mode.
- It should include the possibility of both light and heavy tails.

A collection that meets these requirements begins with the following distribution.

Definition 1 *The* ***mixture of exponentials*** *distribution (to be denoted by M in this paper) has the following distribution function:*

$$F_M(x; \boldsymbol{\alpha}, \boldsymbol{\theta}, k) = 1 - \alpha_1 \exp(-x/\theta_1) - \cdots - \alpha_k \exp(-x/\theta_k)$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)'$ is a vector of positive weights that sum to 1, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)'$ is a vector of exponential means, and k is a positive integer.

This distribution was promoted by Clive Keatinge [4]. For finite values of k this model is not unlike an exponential distribution in that all moments always exist. However, it cannot have an increasing failure rate, making its tail no lighter than a single exponential distribution. As k becomes infinite, with an appropriate choice of weights, the distribution approaches a Pareto distribution. Therefore, even with finite k a fairly heavy tail can be obtained. A second drawback is that the mode of this distribution is always at zero. To add the required flexibility, the following extension is proposed.

Definition 2 *The **augmented mixture of exponentials** distribution (denoted A) has the following distribution function:*

$$F_A(x) = mF_M(x; \boldsymbol{\alpha}, \boldsymbol{\theta}, k) + gF_G(x) + lF_L(x) + pF_P(x)$$

where m, g, l, and p are non-negative numbers that sum to 1 with either $g = 0$ or $l = 0$. In addition, $F_G(x)$ is the cdf of the gamma distribution, $F_L(x)$ is the cdf of the lognormal distribution, and $F_P(x)$ is the cdf of the Pareto distribution.

The addition of the lognormal or gamma distribution (two commonly used models) allows for an interior mode. The Pareto distribution is included for two reasons (given that the mixture of exponentials can approximate the Pareto, it may appear to be redundant). First, the Pareto distribution is commonly used (as are the lognormal and gamma distributions) and this simple model should be part of the modelers toolkit. Second, the principal of parsimony would favor a simple two-parameter Pareto model over a similar many-parameter mixture of exponentials model. Similarly, the suggestion that at most one of the lognormal and gamma distributions be used is again for purposes of parsimony, though if a compelling argument could be made, both might be used. Having both in the mixture also leads to the possibility of a model with two modes. One of the motivations for keeping the collection of models small is to avoid conducting a large number of hypothesis tests. Because the possibility of error is inherent in any hypothesis test, conducting too many tests may nearly guarantee that an error will be made. To further reduce the number of tests, the lognormal, gamma, or Pareto distributions should be added only if there is solid *a priori* reason to do so.

While this distribution could be viewed as a collection of distributions, it still contains a modest number of possibilities. We prefer to view it is a single distribution with a variable number of parameters.

Mixture models are easy to work with. The density function is the same mixture of the individual density functions. Raw moments are the same mixture of individual raw moments. That is,

$$E(A^n) = m\sum_{j=1}^{k}\alpha_j\theta_j^n n! + gE(G^n) + lE(L^n) + pE(P^n).$$

3 Measuring the quality of a proposed model

Given a particular model, there are two steps. The first is to estimate the parameters. This is done by maximum likelihood. The second is to evaluate the quality of the proposed model. This is done by comparing the data to the model. After indicating how the data can be organized, we discuss both graphical comparisons and formal statistical tests. This Section also introduces two data sets that are used to demonstrate the process.

3.1 Organizing the data

The goal is to compare the proposed model to the data. The proposed model is represented by either its density or distribution function, or perhaps some functional of these quantities such as the limited expected value function or the mean residual life function. The data can be represented by the empirical distribution function or a histogram. The graphs and functions are easy to construct when there is individual, complete, data. When there is grouping, or observations have been truncated or censored, difficulties arise. In the spirit of a unified approach, a single method of representing the distribution and density functions of the data will be proposed.

To implement the approach, $\mathbf{x}_j$, the jth "data point" consists of the following items.

t_j, the left truncation point associated with the observation.

c_j, the lowest possible value that produced the data point.

d_j, the highest possible value that produced the data point.

w_j, the weight associated with the data point.

Then the data point is $\mathbf{x}_j' = (t_j, c_j, d_j, w_j)$. The weight is the number of times that particular observation appeared. These represent the fewest pieces of information need for writing the likelihood function. If a specific value was observed, then $c_j = d_j$ and the contribution to the likelihood function is

$$\left[\frac{f(c_j)}{1-F(t_j)}\right]^{w_j}$$

while if value was known to be in the interval from c_j to d_j the contribution is

$$\left[\frac{F(d_j)-F(c_j)}{1-F(t_j)}\right]^{w_j}.$$

In addition, these four values provide enough information to obtain the empirical distribution.

A few examples may clarify this notation. A policy with a deductible of 50 produced a payment of 200. Then, the actual loss was 250 and the data point is $(50, 250, 250, 1)$. Repeating the value of 250 indicates that the exact value was observed. Next, consider a mortality study following people from birth. If 547 people were observed to die between the ages of 40 and 50, the data point is $(0, 40, 50, 547)$. Finally, if the policy in the first example had a maximum payment of 500 and 27 claims were observed to be paid at the limit, the data point is $(50, 550, \infty, 27)$. This notation allows for left truncation and right censoring.

3.2 Representing the data

The data will be represented by the Kaplan-Meier estimate of the survival function. Because interval data (as in the second example above) is not allowed when constructing this estimator, an approximation must be introduced. Suppose there were w observations in the interval from c to d. One way to turn them into individual observations is to uniformly allocate them through the interval. Do this by placing single observations at the points $c + b/w$, $c + 2b/w$, $\ldots$, $c + b$ where $b = d - c^2$. If the data were truncated, that t value is carried over to the individual points. For example, the data point $(0, 40, 50, 547)$ is replaced by 547 data points beginning with $(0, 40{+}10/547, 40{+}10/547, 1)$, through $(0, 50, 50, 1)$. When the Kaplan-Meier estimates are connected by straight lines the ogive results. The algorithm for the Kaplan-Meier estimate is given in the Appendix. For the rest of this

[2] If there is a large weight, it is not necessary to use the large number of resulting points. The choice made here is for programming convenience, not statistical accuracy.

article, it is assumed that, for the purpose of performing hypothesis tests and creating most graphs, all grouped data points have been converted to individual data points. An exception is the calculation of maximum likelihood estimates where the data will be used as collected. Finally, note that groups running from c to ∞ must remain as is. These right censored observations will be treated as such by the Kaplan-Meier estimate.

It should be noted that after this conversion, there are only two types of data points. One is uncensored data points of the form (t, x, x, w) and the other is right censored data points of the form (t, x, ∞, w). The formulas presented here assume that all points with the same first three elements are combined with their weights added. The uncensored points are then ordered as $y_1 < y_2 < \cdots < y_k$ where k always counts the number of unique uncensored values.

The Kaplan-Meier estimate provides the empirical distribution function. If a histogram is desired, it can be constructed through differencing. Let $\hat{F}(x)$ be the empirical distribution function and let $c_0 < c_1 < \cdots < c_h$ be the boundaries for the histogram. The function to plot is

$$\hat{f}(x) = \frac{\hat{F}(c_j) - \hat{F}(c_{j-1})}{c_j - c_{j-1}}, \quad c_{j-1} \leq x < c_j.$$

If the data were originally grouped and the same boundaries used, this approach will reproduce the customary histogram. If the user can choose the groups, one suggestion for the number of groups is Doane's rule [10, page 126]. It suggests computing $\log_2 n + 1 + \log_2(1 + \hat{\gamma}\sqrt{n/6})$ and then rounding up to the next integer to obtain the number of groups. $\hat{\gamma}$ is the sample kurtosis. This is a modification of the more commonly used Sturges' rule with the extra term allowing for non-normal data. For a given number of groups it is reasonable to then set the intervals to be of equal width or to be of equal probability. The default boundaries in the accompanying software use this rule with intervals of equal width.

3.3 Representing the model

In order to compare the model to truncated data, begin by noting that the empirical distribution begins at the lowest truncation point and represents conditional values (that is, it is the distribution and density function given that the observation exceeds the lowest truncation point). In order to make a comparison to the empirical values, the model must also be truncated. Let the lowest truncation point in the data set be T. That is, $T = \min_j\{t_j\}$. Then the model distribution and density functions to use are

$$F_T(x) = \begin{cases} 0, & x < T \\ \frac{F(x)-F(T)}{1-F(T)}, & x \geq T \end{cases}$$

and

$$f_T(x) = \begin{cases} 0, & x < T \\ \frac{f(x)}{1-F(T)}, & x \geq T. \end{cases}$$

Besides T, there is another point of interest. When the largest observed value in the data set is a right censored observation, the empirical distribution function is not defined past that point. When the largest observed value in the data set is not censored, the Kaplan-Meier estimate of the distribution function will be 1 at that point and the empirical distribution function is defined for all values. Let U be the largest right censored observation, provided it is greater than or equal to all uncensored observations, otherwise, set $U = \infty$. When providing formulas, the points $y_0 = T$ and $y_{k+1} = U$ will be added to the list of uncensored observations.

3.4 Two data sets

For illustrative purposes, consider the following two data sets. The f irst data set is real and the second is artificially created to illustrate the points being made in this paper.

Example 3 *(Data Set A) 392 dental claims were recorded. The data were grouped with the results given in Table 1. All observations have a left truncation point of zero, so it is not given in the Table. As well, no observations were right censored.*

c_j	d_j	w_j	c_j	d_j	w_j
0	25	6	600	700	15
25	50	24	700	800	13
50	75	30	800	900	8
75	100	31	900	1,000	2
100	150	57	1,000	1,250	5
150	200	42	1,250	1,500	5
200	250	38	1,500	2,000	5
250	300	27	2,000	2,500	7
300	400	30	2,500	3,000	2
400	500	28	3,000	4,000	1
500	600	16	4,000	∞	0

Table 1 - Dental claims

Example 4 *(Data Set B) 100 observations were made on liability claims. The policies had deductibles of 100, 250, or 500 (all values are in thousands) and maximum payments of 1,000, 3,000, or 5,000. The data are given in Table 2. For the column headed* d/w_j*, if the entry is less than 10, interpret it as the value of* w_j *with* $d_j = \infty$ *known, otherwise, interpret it as the value of* d_j *with* $w_j = 1$ *known. Although there were policies with deductibles of 100 and 250 that had maximum payments of 5,000, none of those policies had a payment at the maximum.*

t_j	c_j	d/w_j	t_j	c_j	d/w_j	t_j	c_j	d/w_j
100	182	182	250	931	931	250	1441	1441
100	184	184	100	960	960	100	1495	1495
250	296	296	250	974	974	500	1500	7
250	331	331	250	1016	1016	100	1556	1556
250	381	381	100	1044	1044	250	1564	1564
100	401	401	250	1060	1060	250	1614	1614
250	491	491	500	1064	1064	250	1647	1647
250	495	495	100	1100	7	250	1737	1737
250	505	505	250	1105	1105	100	1744	1744
500	514	514	500	1122	1122	100	1751	1751
100	547	547	500	1131	1131	500	1768	1768
250	553	553	100	1141	1141	500	1807	1708
500	601	601	100	1148	1148	250	1811	1811
250	616	616	500	1156	1156	250	2031	2031
500	653	653	250	1178	1178	500	2080	2080
250	674	674	500	1200	1200	500	2263	2263
250	685	685	500	1213	1213	250	2275	2275
250	693	693	250	1215	1215	500	2671	2671
500	708	708	500	1240	1240	500	2752	2752
100	771	771	250	1250	2	250	2880	2880
100	793	793	250	1259	1259	100	3100	4
250	825	825	250	1294	1294	250	3250	2
500	840	840	100	1301	1301	500	3469	3469
100	872	872	500	1372	1372	500	3500	2
250	885	885	100	1383	1383	250	4254	4254
250	913	913	100	1409	1409	100	4510	4510
250	927	927	250	1434	1434	500	5500	1
250	929	929						

Table 2 - Liability claims

Example 5 *Construct the Kaplan-Meier estimate for each of the two examples.*

For Data Set A, the Kaplan-Meier estimate is the ogive as depicted in Figure 1.[3] For Data Set B, the estimate is depicted in Figure 2. The calculations

[3] When grouped data are made discrete, the empirical distribution function is a step function. Here, when data are grouped with non-overlapping boundaries, the more tradi-

are in the Appendix.

3.5 Graphical comparison of the density and distribution functions

The plots in this Section provide various ways to visualize the difference between the empirical distribution and a proposed model distribution.

3.5.1 Distribution function plot

Example 6 *For Data Sets A and B, plot a lognormal model against the estimated distribution function.*

The functions to be plotted are $\hat{F}(x)$ and $F_T(x)$ with the range being $T \leq x \leq U$. The maximum likelihood estimates[4] for the lognormal model are $\hat{\mu} = 5.35376$ and $\hat{\sigma} = 1.02432$ for Data Set A and $\hat{\mu} = 7.16304$ and $\hat{\sigma} = 0.858883$ for Data Set B. The plots appear below in Figures 3 and 4. While commonly used, this plot is often not effective for skewed distributions because all the distribution functions look somewhat the same. However, if confidence bands are desired, they must be calculated for this plot before they can be modified for the others.

For Data Set A the lognormal model appears to be a good fit. For Data Set B the fit deteriorates after about 1,500.

It is possible to construct a confidence band around the empirical distribution function. A 95% band implies that were a large number of samples to be taken and bands constructed, 95% of those bands would completely enclose the true distribution function. The following version is taken from [6].

tional ogive will be presented, rather than the step function.

[4] When estimating parameters, the data as actually collected (not as made discrete) is used.

Let r_j and let s_j be the quantities needed to obtain the Kaplan-Meier estimate (see Appendix). Define

$$\begin{aligned}
v_j &= \sum_{i=1}^{j} \frac{s_i}{r_i(r_i - s_i)} \\
\delta_j &= \exp\left(\frac{c\sqrt{v_j}}{\ln[1-\hat{F}(y_j)]}\right) \\
c &= 2.6161 - 4.2316a + 2.0946b + 4.3501a^2 + 3.6047ab \\
&\quad -3.3038b^2 - 3.7714a^3 - 0.6092a^2b - 0.7852ab^2 + 1.8838b^3 \\
a &= \frac{nv_1}{1+nv_1}, \quad b = \frac{nv_k}{1+nv_k}.
\end{aligned}$$

Klein and Moeschberger [6] provide the c-values in a table. This function gives a good approximation (Differences from the table are at most 0.045, when a and b are either both very small or both very large, and more typically about 0.015). For computing b, k is set at the highest value for which $\hat{F}(y_j) < 1$. At each uncensored observation, the bounds for the 95% confidence band are

$$1 - [1-\hat{F}(y_j)]^{\delta_j} \text{ to } 1 - [1-\hat{F}(y_j)]^{1/\delta_j}.$$

Two issues remain. For all the plots, there is a question as to what should be plotted between uncensored observations. Also, what should n, the sample size, be? For plotting, when individual observations are available, the convention is to plot the Kaplan-Meier estimate as a step function. For grouped data, connecting the points via linear interpolation produces the ogive. No preference is expressed here.

With regard to the sample size, the formula given above was derived under an assumption that if there was truncation, all observations were truncated at the same value. For Data Set A, there is no problem because none of the 392 losses were truncated or censored and thus the sample size is clearly 392. For Data Set B, many of the observations were truncated at values above 100 and many were censored. It is not reasonable to give the same confidence to its empirical distribution function as if there were 100 observations with no truncation or censoring. When calculating the interval for a given y_j value, the sample size should be the number of observations with truncation points less than y_j. For Data Set B, the sample size is 30 for $100 < y_j \leq 250$, 70 for $250 < y_j \leq 500$, and 100 for $y_j > 500$.

For the exponential model the parameter estimates are $\hat{\theta} = 358.687$ for Data Set A and $\hat{\theta} = 1,597.80$ for Data Set B. The plots are below in Figures 5 and 6. For Data Set A the exponential model is outside the confidence band, making it a poor choice. For Data Set B the model is inside the band, but clearly does not come very close to the empirical cdf itself.

3.5.2 Other cdf based plots

There are a number of interesting plots that are based on the distribution function. When looking at the distribution function graphs produced above, it should be clear that when the model's distribution function is close to the empirical distribution function, it is difficult to make small distinctions. Among the many ways to amplify those distinctions, two will be presented here. The first is to simply plot the difference of the two functions. That is, if $\hat{F}(x)$ is the empirical distribution function and $F_T(x)$ is the model distribution function, plot $D(x) = \hat{F}(x) - F_T(x)$. The relevant range to plot is again $T \leq x \leq U$. Confidence bands are again available. For this plot, calculate the bands for the CDF plot and then subtract $F_T(y_j)$ from each value. Similar adjustments will produce bands for the other plots to be presented.

Example 7 *Plot $D(x)$ for the previous example, using the lognormal model.*

The plots appear, with confidence bands, in Figures 7 and 8. The lognormal model is inside the bands, but the differences for Data Set B have been magnified.

Perhaps the most effective way to highlight any differences is the $p{-}p$ plot which is also called a probability plot. The plot is created by first selecting a set of k values $0 < x_1 < \cdots < x_k$. A point is then plotted corresponding to each value. The coordinates to plot are $(F(x_j), \hat{F}_T(x_j))$. If the model fits well, the plotted points will be near the forty-five degree line running from 0 to 1. The easiest way to construct the plot is to use the y-values at which the empirical distribution was calculated. Once again, confidence bands can be established by plotting the upper and lower confidence bands in place of the empirical cdf.

Example 8 *Create a $p{-}p$ plot for the previous example, using the lognormal model.*

The plots appear in Figures 9 and 10. Once again, both models are inside the confidence limits, but the lognormal model does not appear to be adequate for Data Set B.

A third plot uses the limited expected value. It is the expected payment when a limit of u is imposed. The empirical limited expected value function is

$$\hat{E}(X_T \wedge u) = \int_0^u 1 - \hat{F}(x)dx.$$

where X_T is the ground-up loss variable, X, conditioned on $X > T$. At an uncensored claim value of y_i the limited expected value is

$$\hat{E}(X_T \wedge y_i) = T + \sum_{j=1}^{i}(y_j - y_{j-1})[1 - \hat{F}(y_{j-1})].$$

Intermediate values are obtained by linear interpolation. For the model, the limited expected value function is (for $u \geq T$)

$$\begin{aligned} E(X_T \wedge u) &= \int_0^u 1 - F_T(x)dx = \int_T^u xf_T(x)dx + u[1 - F_T(u)] \\ &= T + \frac{E(X \wedge u) - E(X \wedge T)}{1 - F(T)}. \end{aligned}$$

Once again, confidence bands can be obtained by determining the limited expected value function of the upper and lower cdf bands. A plot of each function for all u is possible or an analog of the $p-p$ plot can be constructed by plotting the pairs $(E(X_T \wedge y_i), \hat{E}(X \wedge y_i))$.

Of the various cdf-related plots considered here, this one may be the most informative. Many actuarial applications of loss models are related to the expected cost of providing insurance for losses between two values. Having a close correspondence between model and data at those points where limited expected values are to be calculated may be an important objective.

Example 9 *Construct limited expected value plots for the lognormal distribution for each data set.*

The direct plots appear in Figures 11 and 12 while the $p - p$ versions appear in Figures 13 and 14.

The conclusions match those from the earlier plots.

3.5.3 Histogram plot

Sometimes the quality of the fit is best seen by plotting the model density function against the histogram.

Example 10 *Plot the lognormal model against the histogram for each data set.*

For Data Set A the data were already grouped and the plot appears in Figure 15.

For Data Set B the data must be grouped. The default number of groups is 8, but 16 seemed to give a better picture. The picture appears in Figure 16.

These plots add further evidence that the lognormal model is reasonable for Data Set A, but not for Data Set B.

3.6 Hypothesis tests

A picture may be worth many words, but sometimes it is best to replace the impressions conveyed by pictures with mathematical demonstrations. One such demonstration is a test of the hypotheses:

$$\begin{aligned} H_0 &: \text{ The data came from a population with the stated model.} \\ H_1 &: \text{ The data did not come from such a population.} \end{aligned}$$

The test statistic is usually a measure of how close the model distribution function is to the empirical distribution function. When the null hypothesis completely specifies the model (for example, an exponential distribution with mean 100), critical values are well known. However, it is more often the case that the null hypothesis states the name of the model, but not its parameters. When the parameters are estimated from the data, the test statistic tends to be smaller than it would have been had the parameter values been pre-specified. That is because the estimation method itself tries to choose parameters that produce a distribution that is close to the data. In that case, the tests become approximate. Because rejection of the null hypothesis occurs for large values of the test statistic, the approximation tends to increase the probability of a Type II error while lowering the probability of a Type I error. For actuarial modeling this is likely to be an acceptable trade-off. Our goal is to find a useful model, and we are more likely to come to that conclusion when parameters are estimated from data.

3.6.1 Kolmogorov-Smirnov test

The first test considered is the Kolmogorov-Smirnov test. The test statistic is

$$D = \max_{T \leq x \leq U} \left| \hat{F}(x) - F_T(x) \right|$$

When all the observations are individual data points (rather than intervals), the maximum must occur at one of the data points. Strictly speaking, this test should not be used for interval data because the empirical distribution is poorly defined within intervals. In the spirit of being more interested in getting answers than in being precise, interval data will be discretized as indicated earlier and a step function used for the empirical cdf.

To complete the test, two things must be done. The first is to standardize the test statistic so that it can be compared to tabled values. D'Agostino and Stephens [1, pages 112-113], recommend using $D^* = n^{1/2}D + 0.19n^{-1/2}$ where n is the sample size. If there are multiple truncation points, the standardization should be done prior to obtaining the maximum. For each x, the sample size to use is the number of observations with truncation points less than x. If the largest observation is censored (as in Data Set B) the critical value should be reduced. That is because the maximum is taken only over part of the range of possibilities and thus there are fewer opportunities to have a large value for D. In [1] a table is presented. The following cubic functions provide a good approximation.

$$\begin{aligned} 10\% &: 0.9289p^3 - 2.6822p^2 + 2.5761p + 0.4011 \\ 5\% &: 1.1803p^3 - 3.2402p^2 + 2.9628p + 0.4555 \\ 1\% &: 1.6886p^3 - 4.3535p^2 + 3.7262p + 0.5764 \end{aligned}$$

where $p = F_T(U)$. These functions match the tabled values within 0.002 (10%), 0.004 (5%), and 0.01 (1%). The tabled values are generally valid for $n \geq 25$ and $0.2 \leq p \leq 1$.

Example 11 *Conduct the Kolmogorov-Smirnov test at a 5% significance level for both the lognormal and exponential models for both data sets.*

If v is the critical value from the above table, an alternative is to compare D to $v/\sqrt{n} - .19/n$. For Data Set A the test statistics are 1.669 and 0.342 for the exponential and lognormal models respectively. The 5% critical value is 1.358 indicating the lognormal model is acceptable while the exponential model is not. For Data Set B the test statistics are 0.974 and 0.937 with a critical value of 1.358. Both models are acceptable with the lognormal model providing a slightly better fit.

3.6.2 Anderson-Darling test

This test is similar to the Kolmogorov-Smirnov test, but uses a different measure of the difference between the two distribution functions. The test statistic is

$$A^2 = n\int_T^U \frac{[\hat{F}(x) - F_T(x)]^2}{F_T(x)[1 - F_T(x)]} f_T(x)dx.$$

That is, it is a weighted average of the squared differences between the empirical and model distribution functions. The weight is the reciprocal of the variance of $\hat{F}(x)$, $F_T(x)[1 - F_T(x)]/n$. Note that when x is close to T or to U, the weights might be very large due to the small value of one of the factors in the denominator. This test statistic tends to place more emphasis on good fit in the tails than in the middle of the distribution. Calculating with this formula appears to be challenging. However, because the empirical distribution function is constant between uncensored data points, the integral simplifies to

$$\begin{aligned} A^2 &= -nF_T(U) + n\sum_{j=0}^{k}[1 - \hat{F}(y_j)]^2\{\ln[1 - F_T(y_j)] - \ln[1 - F_T(y_{j+1})]\} \\ &\quad + n\sum_{j=1}^{k}\hat{F}(y_j)^2[\ln F_T(y_{j+1}) - \ln F_T(y_j)]. \end{aligned}$$

Note for the first sum that when $j = k$ and $U = \infty$ the first factor will be zero. Then the term in braces (which involves the logarithm of zero) will not need to be evaluated. When there are multiple truncation points, the formula needs to be modified to reflect the non-constant sample size. A reasonable modification is to use

$$\begin{aligned} A^2 &= \sum_{j=0}^{k} n_j[1 - \hat{F}(y_j)]^2\{\ln[1 - F_T(y_j)] - \ln[1 - F_T(y_{j+1})]\} \\ &\quad + \sum_{j=1}^{k} n_j\hat{F}(y_j)^2[\ln F_T(y_{j+1}) - \ln F_T(y_j)] - \sum_{j=0}^{k} n_j[F_T(y_{j+1}) - F_T(y_j)]. \end{aligned}$$

The modified sample size n_j is calculated as follows: Let $t_1 < t_2 < \cdots < t_m$ be the truncation points and let d_i be the number of observations with a truncation point of t_i. For $t_i < y_j \leq t_{i+1}$,

$$n_j = \sum_{h=1}^{i} d_h[1 - \hat{F}(t_h)].$$

The critical values are given below (based on a cubic approximation to the values in [1, Table 4.4] where again $p = F_T(U)$).

$$\begin{aligned} 10\% &: -0.4579p^3 + 0.3589p^2 + 2.0106p + 0.0243 \\ 5\% &: -0.9301p^3 + 0.8149p^2 + 2.5519p + 0.0548 \\ 1\% &: -1.8586p^3 + 1.3585p^2 + 4.3242p + 0.0545 \end{aligned}$$

for $0.2 \leq p \leq 1$. The approximation is accurate within 0.007, 0.010, and 0.029 for the 10%, 5%, and 1% tests respectively. For this test, a sample size of 5 is sufficient.

Example 12 *Repeat the previous example using the Anderson-Darling test.*

For Data Set A the test statistics are 5.5881 and 0.1962 for the exponential and lognormal models respectively. The critical value is 2.4915. This test magnifies the difference in the quality of fit of the two models. For Data Set B the test statistics are 1.1224 and 0.668 with a critical values of 2.442 and 2.424. While both models are acceptable, this test again magnifies the difference.

3.6.3 Chi-square goodness-of-fit test

Unlike the previous two tests, this test allows for some discretion. It begins with the selection of $m-1$ arbitrary values, $t = c_0 < c_1 < \cdots < c_m = \infty$. Let $p_j = F_T(c_j) - F_T(c_{j-1})$ be the probability a truncated observation falls in the interval from c_{j-1} to c_j. When determining $F_T(x)$, $U = \infty$ is assumed (that is, having the largest observation be censored does not require an adjustment). Similarly, let $\hat{p}_j = \hat{F}(c_j) - \hat{F}(c_{j-1})$ be the same probability according to the empirical distribution. The test statistic is then

$$\chi^2 = \sum_{j=1}^{m} \frac{n_j(p_j - \hat{p}_j)^2}{p_j}.$$

The sample size, n_j, measures the number of observations that could have been in the interval and is calculated the same way as done previously.

The critical value for this test comes from the chi-square distribution with degrees of freedom equal to the number of terms in the sum minus one minus the number of estimated parameters. There are a number of rules that have been proposed for deciding when the test is reasonably accurate. They center around the values of $n_j p_j$. The most conservative states that each must be at least 5. Some authors claim that values as low as 1 are acceptable. All agree the test works best when the values are about equal from term to term. If the data are grouped, it is best, though not necessary to use the groups as given. For individual data, Moore (in [1, page 70]) recommends using $2n^{0.4}$ equiprobable (using the empirical distribution) groups. When there are multiple censoring and/or truncation points, they should be considered as potential boundaries. The software uses this guide, however, excessive censoring may reduce the number of groups by one or two.

Example 13 *Perform the chi-square goodness-of-fit test for the continuing example.*

For Data Set A the recommendation is for 21 groups (the original groupings for this data set could also have been used). The test statistics are 37.29 for the exponential model (p-value of 0.0073 with 19 degrees of freedom) and 4.15 for the lognormal model (p-value of 0.9997 with 18 degrees of freedom). Once again the lognormal model is both acceptable and superior.

For Data Set B the recommendation is for 11 groups. The test statistics are 14.51 and 10.81 for the two models and the p-values are 0.1054 and 0.2126 indicating that both models are acceptable, but the lognormal model continues to provide a better fit.

Because all three tests are trying to measure the same thing (the difference between the data and the model), the results will usually be similar. When there is no grouping, the K-S and A-D tests make the most sense because no arbitrary decisions need to be made. Between the two, the A-D test is more sensitive to differences in the tails and so may be better for actuarial data. When data are grouped, the chi-square test can be done directly while the other tests require an ungrouping process that is likely to reduce their accuracy.

4 Selecting a model

Often, more than one model will be deemed acceptable by the processes described so far. Selection of a single model should be based on two simple goals.

1. Use a simple model if at all possible.

2. Restrict the universe of potential models.

Both of these goals are important to prevent overfitting. This occurs when the model more closely fits the data than it fits the population that produced the data. With enough parameters and a rich collection of models, it is possible to fit the data perfectly. For example, with discrete data the empirical distribution fits the data perfectly, but will not match a population that is known to be continuous.

For situations where there is a lot of data, a complex model becomes tempting. The area of analysis commonly called “data mining” provides a set of tools to prevent this from happening. They tend to be based on using only a portion of the data for determining the model and then using the remaining data for parameter estimation.

Any model selection process must involve judgment as well as the purpose for constructing the model. For example, the 1941 CSO mortality table follows a Makeham distribution for much of its range of ages. In a time of limited computing power, such a distribution allowed for easier calculation of joint life values. As long as the fit of this model was reasonable, this advantage outweighed the use of a different, but better fitting, model. Similarly, if the Pareto distribution has been used to model a particular line of liability insurance both by the analyst’s company and by others, it may require more than the usual amount of evidence to change to an alternative distribution.

Finally, it should be noted that the approach outlined below does not always lead to a unique model choice. In that case judgment is most definitely required.

We recommend fitting a model using the following steps.

1. Use Kaplan-Meier to construct the empirical distribution.

2. Construct pictures.

3. Conduct hypothesis tests.

4. Calculate the Schwarz Bayesian Criterion for each model.

For the Kolmogorov-Smirnov and Anderson-Darling tests, no adjustments are made when the number of parameters is increased. As a result, more complex models will often fare better on the tests (as outlined here). The chi-square goodness-of-fit test adjusts the degrees of freedom. All three tests are sensitive to the sample size. In particular, as the sample size increases, the test statistic tends to increase (when the null hypothesis is false). Because we know that the null hypothesis *is* false, a large sample size will lead to a rejection of all models. Regardless, the test statistic itself can give some guidance in the selection process.

One method that can lead to an unambiguous choice is the Schwarz Bayesian Criterion [9]. Each model is given a score of $l - (r/2)\ln n$ where l is the logarithm of the likelihood function at its maximum, r is the number of estimated parameters (more accurately, the dimension of the space over which the likelihood function was maximized), and n is the sample size. The model with the highest score is the recommended selection. The penalty for adding parameters is explicitly stated with this method and for large samples it becomes more difficult to add a parameter (relative to other information criterion such as Akaike or using the likelihood ratio test). Unlike the likelihood ratio test, this method does not require that the models under consideration be related in any way. While there are other information criteria available, the authors prefer this one because it adjusts both for sample size and the number of parameters.

Example 14 *Determine models for Data Sets A and B.*

For Data Set A the decision is relatively simple. With a clear interior mode, only models incorporating the lognormal or gamma distributions should be considered. The table below lists some relevant calculations. Because the data were grouped, only the chi-square goodness-of-fit test was done.

Model	l	r	χ^2 p-value	SBC
Lognormal	-1068.79	2	0.4037	-1074.76
Gamma	-1100.14	2	0.0000	-1106.11
Lognormal/exp mix	-1067.63	4	0.4949	-1079.57
Gamma/exp mix	-1071.85	4	0.1007	-1083.80
Lognormal/exp/exp mix	-1067.63	6	0.3508	-1085.54

The SBC clearly favors the lognormal distribution and tests and graphs done earlier confirm that it fits well. Adding an exponential or two to the lognormal bring a slight improvement, but it is not noticeable. The density functions for the two models are indistinguishable. The lognormal distribution is the best choice for this example.

For Data Set B, a few additional models are now considered. We begin with Steps 2 and 3, looking at a number of calculations for each model. For the Pareto model, there was no maximum likelihood estimate and for a mixture of two exponentials the second exponential as a small weight placed on an exponential distribution with a mean at infinity. The loglikelihood for this model is −628.16 making it not viable versus the other models. For the chi-square test the *p*-value is given.

Model	l	r	A-D	K-S	χ^2	SBC
Exponential	−628.23	1	1.2245	0.9739	0.1054	−630.53
Lognormal	−626.26	2	0.6682	0.9375	0.2126	−630.87
Gamma	−627.35	2	0.8369	1.0355	0.2319	−631.96
Lognormal/exp mix	−623.77	4	0.2579	0.5829	0.5608	−632.98
Gamma/exp mix	−623.64	4	0.2804	0.5773	0.5260	−632.85
Lognormal/exp/exp mix	−623.39	6	0.1484	0.4494	0.3472	−637.21
Gamma/exp/exp mix	−623.26	6	0.1353	0.4652	0.3348	−637.08

From these numbers, there are three reasonable choices depending on the analyst's approach.[5] An analyst who favors parsimony will pick either the exponential or lognormal model. They are favored by the SBC and both pass the goodness-of-fit tests. The graphs presented earlier favor the lognormal model. The other extreme is to choose between the two mixtures with two exponential distributions included. If the augmented mixture presented in this paper is viewed a single distribution, then these two clearly maximize the likelihood (adding a third exponential in each case does not increase the likelihood) and parsimony is not relevant. The lognormal mixture has a better K-S statistic while the gamma mixture has the best A-D statistic. This indicates that the latter may do better in the tails. Finally, a compromise position is to choose one of the mixtures with a single exponential. The goodness-of-fit measures are clearly better than the simpler models and

[5] One referee insisted that the SBC should dominate the decision process and thus asked us to conclude in favor of the exponential distribution. Another referee took the "one model" approach and asked us to conlude in favor of the lognormal/exp/exp model because it maximizes the likelihood function.

compare well the more complex models. The graphs (NEED TO REPLACE THE FINAL GRAPHS) indicate....

References

[1] D'Agostino, R. and Stephens. M., eds. 1986, *Goodness-of-fit Techniques*, New York: Marcel Dekker.

[2] Johnson, N., Kotz, S., and Balakrishnan, N. 1994, *Continuous Univariate Distributions, Vol. 1, 2nd ed.,* New York: Wiley.

[3] Johnson, N., Kotz, S., and Balakrishnan, N. 1995, *Continuous Univariate Distributions, Vol. 2, 2nd ed.,* New York: Wiley.

[4] Keatinge, C. 1999, "Modeling Losses with the Mixed Exponential Distribution, *Proceedings of the Casualty Actuarial Society*, Vol. LXXXVI, pp. 654–698.

[5] Kleiber, C. and Kotz, S. 2003, *Statistical Size Distributions in Economics and Actuarial Sciences*, New York: Wiley.

[6] Klein, J. and Moeschberger, M. 1997, *Survival Analysis: Techniques for Censored and Truncated Data*, New York: Springer-Verlag.

[7] Klugman, S., Panjer, H. and Willmot, G. 1998, *Loss Models: From Data to Decisions*, New York: Wiley.

[8] London, D. 1997, *Survival Models and Their Estimation*, Winsted, CT: Actex.

[9] Schwarz, G. 1978, "Estimating the Dimension of a Model," *Annals of Statistics*, Vol. 6, pp. 461–464.

[10] Venables, W. and Ripley, B. (1994), *Modern Applied Statistics with S-Plus*, New York: Springer-Verlag.

A The Kaplan-Meier Product-Limit estimator

The estimator can be found in most biostatistics texts such as [6] and in the actuarial-oriented text [8]. It will be presented here using the notation of this paper. The observations are (t_j, c_j, d_j, w_j) where t_j is the left-truncation point, $c_j = d_j$ indicates that an uncensored value of c_j was observed, $d_j = \infty$ indicates that an observation was censored at c_j, and w_j indicates the number of observations that had the preceding values. It is assumed that the data have been ungrouped so that $c_j < d_j < \infty$ is not possible. Then let $y_1 < \cdots < y_k$ be the ordered collection of the c_j values for those with $c_j = d_j$. Let $y_0 = T = \min\{t_j\}$ and let $y_{k+1} = U$ where $U = \infty$ if the largest value of c_j is such that the corresponding $d_j = c_j$, otherwise, set U as the largest value of c_j. Note that $U = \infty$ occurs when there are no censored observations larger than the largest uncensored observation.

To construct the estimate, two sets of quantities must be calculated. The first is s_i for $i = 1, \ldots, k$ which gives the number of uncensored observations equal to y_i. That is,

$$s_i = \sum_{c_j = d_j = y_i} w_j.$$

The second is r_i for $i = 1, \ldots, k$ which gives the number of policies that could have produced an uncensored observation of y_i given that the observation was known to be at least y_i. This is the sum of the weights for all policies with a deductible less than y_i less the sum of the weights for all policies with observed values c_j (whether or not they were censored) less than y_i. That is,

$$r_i = \sum_{t_j < y_i} w_j - \sum_{c_j < y_i} w_j.$$

The Kaplan-Meier estimate is then

$$\hat{F}(x) = \begin{cases} 0, & x < y_1 \\ 1 - \prod_{j=1}^{i} \frac{r_j - s_j}{r_j} & y_i \leq x < y_{i+1}, \quad i = 1, \ldots, k. \end{cases}$$

If $U = \infty$ it will turn out that $\hat{F}(x) = 1$ for $x \geq y_k$ while if U is finite, $\hat{F}(x)$ is undefined for $x \geq U$.

For Data Set B, the calculations are given in Table A. Because no uncensored values were repeated, $s_i = 1$ for all i and so these values do not appear in the Table. The $\hat{F}(y)$ values are interpreted as applying to the interval from the current y_i value to the next value. The smallest truncation value is $T = 100$ which is where the Table starts and the largest observation was censored and so $U = 5,500$. Thus, the empirical distribution function is not defined past 5,500.

y_i	r_i	$\hat{F}(y_i)$	y_i	r_i	$\hat{F}(y_i)$	y_i	r_i	$\hat{F}(y_i)$
100	–	0	913	75	0.316	1409	40	0.578
182	30	0.033	927	74	0.325	1434	39	0.589
184	29	0.067	929	73	0.334	1441	38	0.599
296	68	0.080	931	72	0.343	1495	37	0.610
331	67	0.094	960	71	0.353	1556	29	0.624
381	66	0.108	974	70	0.362	1564	28	0.637
401	65	0.122	1016	69	0.371	1614	27	0.651
491	64	0.135	1044	68	0.380	1647	26	0.664
495	63	0.149	1060	67	0.390	1737	25	0.677
505	92	0.158	1064	66	0.399	1744	24	0.691
514	91	0.168	1105	58	0.409	1751	23	0.704
547	90	0.177	1122	57	0.419	1768	22	0.718
553	89	0.186	1131	56	0.430	1807	21	0.731
601	88	0.195	1141	55	0.440	1811	20	0.745
616	87	0.205	1148	54	0.451	2031	19	0.758
653	86	0.214	1156	53	0.461	2080	18	0.772
674	85	0.223	1178	52	0.471	2263	17	0.785
685	84	0.232	1200	51	0.482	2275	16	0.798
693	83	0.242	1213	50	0.492	2671	15	0.812
708	82	0.251	1215	49	0.502	2752	14	0.825
771	81	0.260	1240	48	0.513	2880	13	0.839
793	80	0.269	1259	45	0.524	3469	6	0.866
825	79	0.279	1294	44	0.534	4254	3	0.910
840	78	0.288	1301	43	0.545	4510	2	0.955
872	77	0.297	1372	42	0.556	5500	1	–
885	76	0.306	1383	41	0.567			

Table A - Kaplan-Meier estimates for Data Set B

B Figures

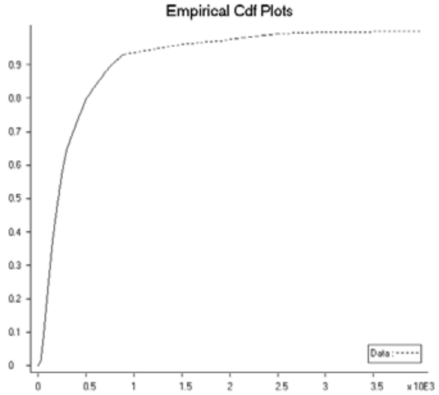


Fig. 1 - Empirical cdf for Data Set A

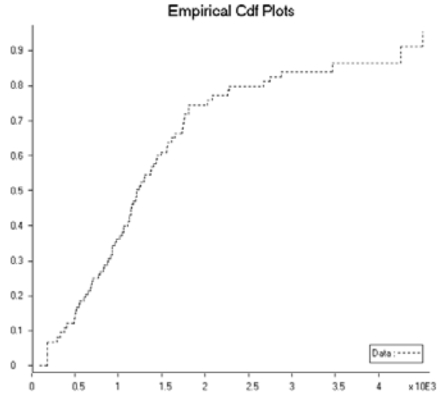


Fig. 2 - Empirical cdf for Data Set B

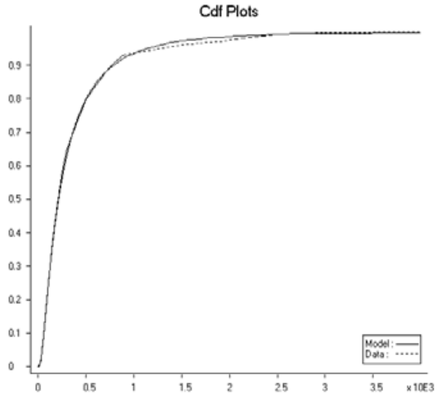


Fig. 3 - lognormal cdf plot for ds A

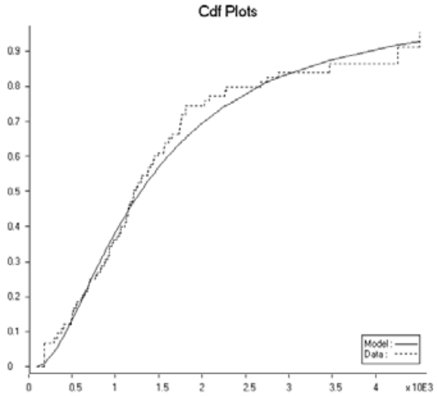


Fig. 4 - lognormal cdf plot for ds B

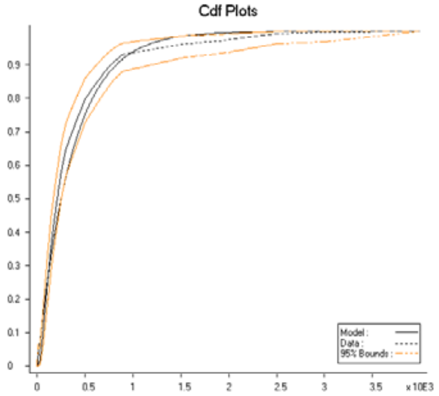


Fig. 5 - exponential cdf plot for ds A

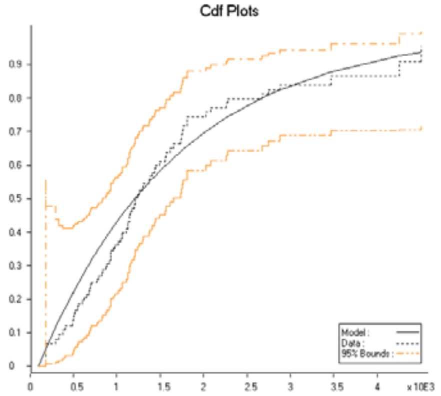


Fig. 6 - exponential cdf plot for ds B

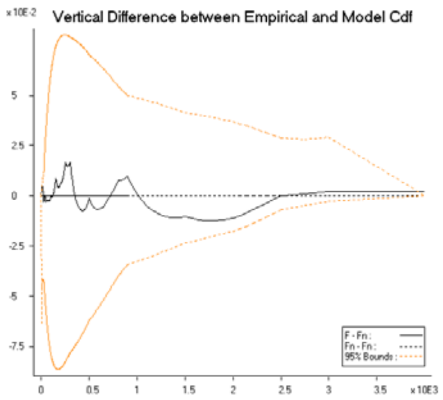


Fig. 7 - lognormal difference plot for ds A

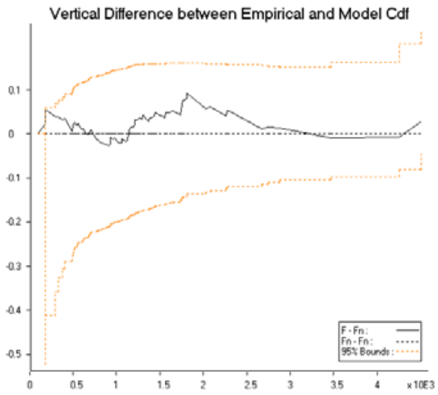


Fig. 8 - lognormal difference plot for ds B

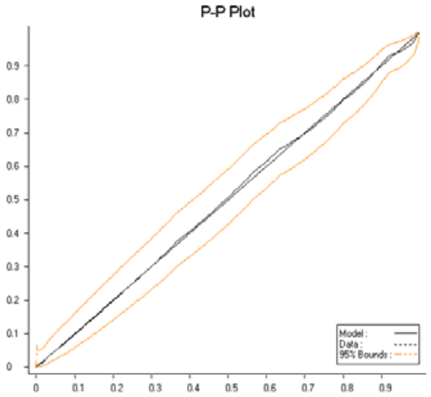


Fig. 9 - lognormal p-p plot for ds A

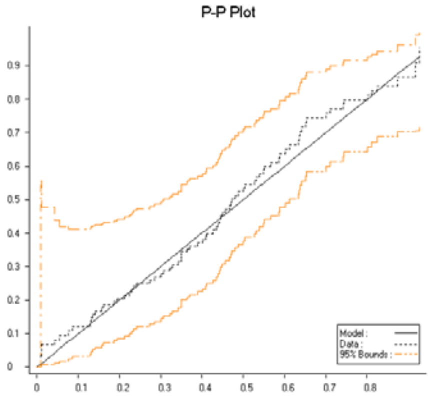


Fig. 10 - lognormal p-p plot for ds B

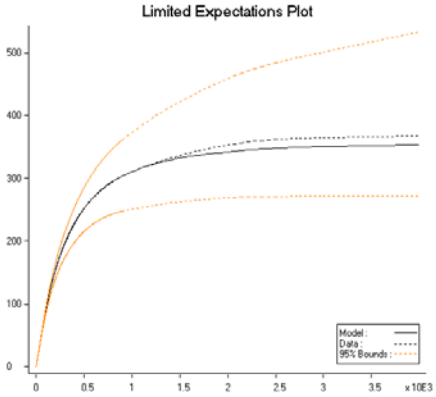


Fig. 11 - lognormal lev plot for ds A

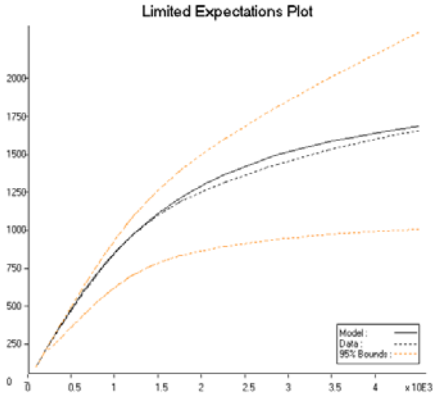
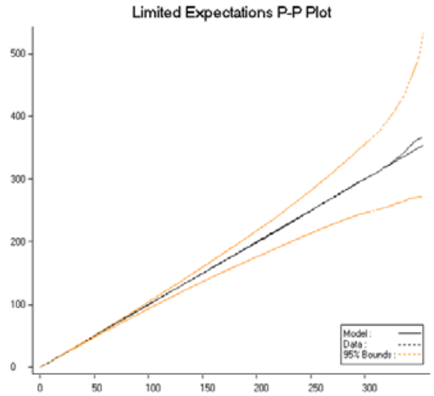


Fig. 12 - lognormal lev plot for ds B



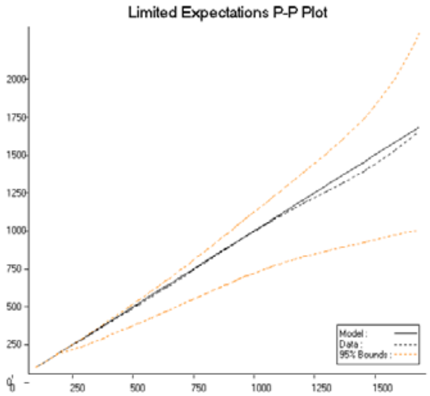
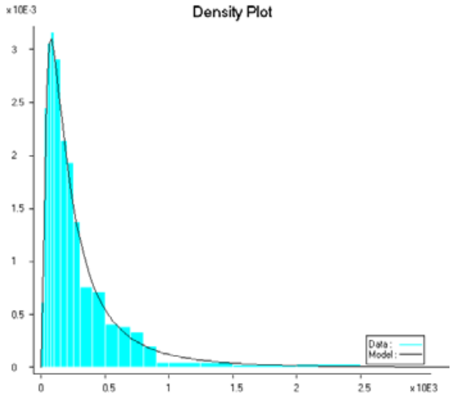


Fig. 13 - lognormal lev p-p plot for ds A Fig. 14 - lognormal lev p-p plot for ds B



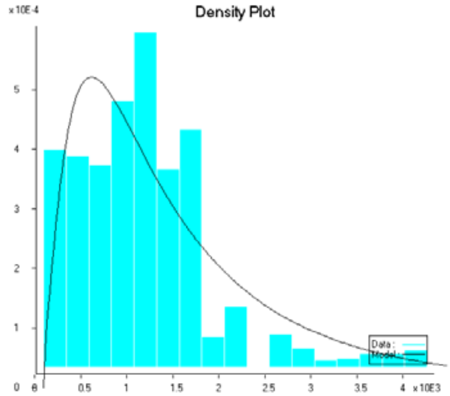
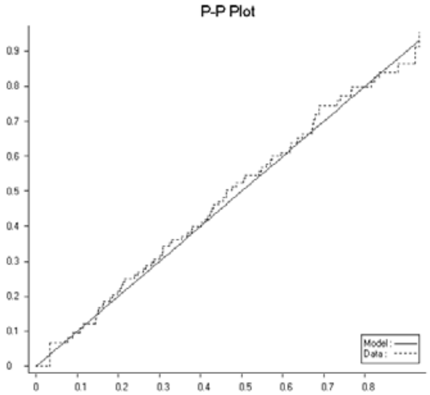


Fig. 15 - lognormal density plot for ds AFigure 16 - lognormal density plot for ds B



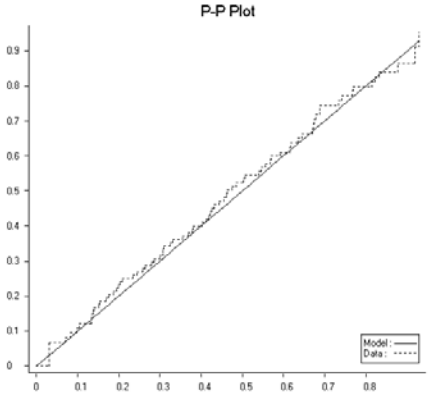


Fig. 17 - lognormal/exp p-p plot Fig. 18 - gamma/exp p-p plot

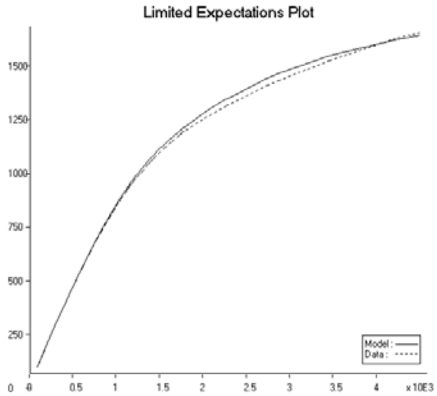


Fig. 19 - lognormal/exp lev plot

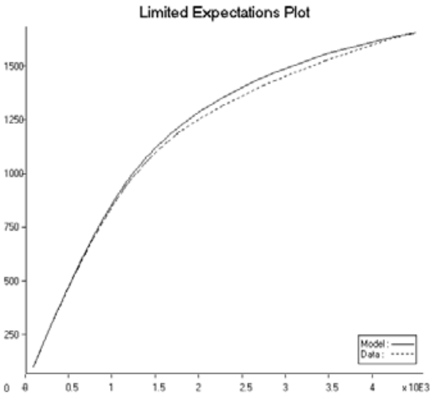


Fig. 20 - gamma/exp lev plot