

## Modelling Techniques For Limited Data Sets

By Luis Sánchez and Robert Ceske

The relative scarcity of operational risk data means that risk managers often have to adjust either the data that is available to them, or the models that they use (Table A).

Table 1: Sources of data

- Internal operational loss data, collected from within an institution
- Other institutions' operational loss data, used as a proxy for the institution that is being analysed
- Educated opinions, such as management scenarios or self-assessments

In this article we introduce a series of techniques that can be applied to limited data sets, or that estimate/extrapolate data using limited samples.

### Fitting data distributions

The first technique that we will describe assumes that if the distribution were complete, it would take the shape of some known "model" distribution. The risk manager's problem is therefore to find the most appropriate "model" distribution, and to estimate the most appropriate parameters for this distribution.

The frequency distribution of operational risk – the chance that a loss event will occur – tends to be best represented by a Poisson, binomial or negative binomial distribution. Based on observations of public and non-public loss events collected through our RiskOps™ product and in consulting engagements, we have found that the severity distribution – the size of the loss – tends to be best represented by a lognormal, logistic, Weibull, gamma or Pareto distribution. The choice of which distribution to use (when only one model distribution is used – see later examples), is strongly influenced by the region of the severity distribution that is of interest.

After selecting the "model" distribution, the risk manager uses the limited sets of empirical data that are available to him/her to estimate the model parameters, as discussed later. Once the distributions have been established, a Capital-at-Risk ("CaR") model can be applied, and CaR results obtained.

Of course, there are some problems to overcome. Distributions of operational risk data are often characterized by a "fat tail" – a relatively high proportion of "unusual" or "catastrophic" events. So fitting a single curve (e.g. normal, logistic, Weibull, gamma or Pareto distribution) to the severity distribution may result in a "good fit" with the main body of the distribution, but a relatively "poor fit" in the tail. To combat this, the severity distribution can be broken up and

different distributions can be fitted to different portions of the curve. To achieve this, the risk manager must first estimate the shape of the distribution of the underlying data in different regions of the distribution. For example, in the case of a severity distribution, the risk manager might use an empirical distribution for the bulk, lognormal for the middle, and generalized Pareto for the tail.

Using multiple distributions to estimate the distribution of the underlying data means that the CaR results will be more robust. That is, they will be more resilient to the addition of new losses (even large ones) than results obtained using a single distribution.

### Curve-fitting techniques

The key to stable and robust CaR numbers is to find distributions that best fit the data. The parameters used to “fit” the model distributions to empirical operational loss data might be selected on the basis of opinion, or by visually comparing the distributions to plots of the actual data. But often it is better to estimate parameters by applying “goodness of fit” tests to the existing data. For example, Chi-square and Kolmogorov-Smirnov (KS) tests can be used to determine the “closest” fit for a specific distribution: the distribution with the best score is the one that should be used.

Because existing data is used in the choice and calibration of model distributions, “goodness of fit” tests make sense when a moderate amount of data is available. The “best fits” derived from very limited sample sets may not necessarily reflect the distributions that would be expected from the complete distribution (were it available).

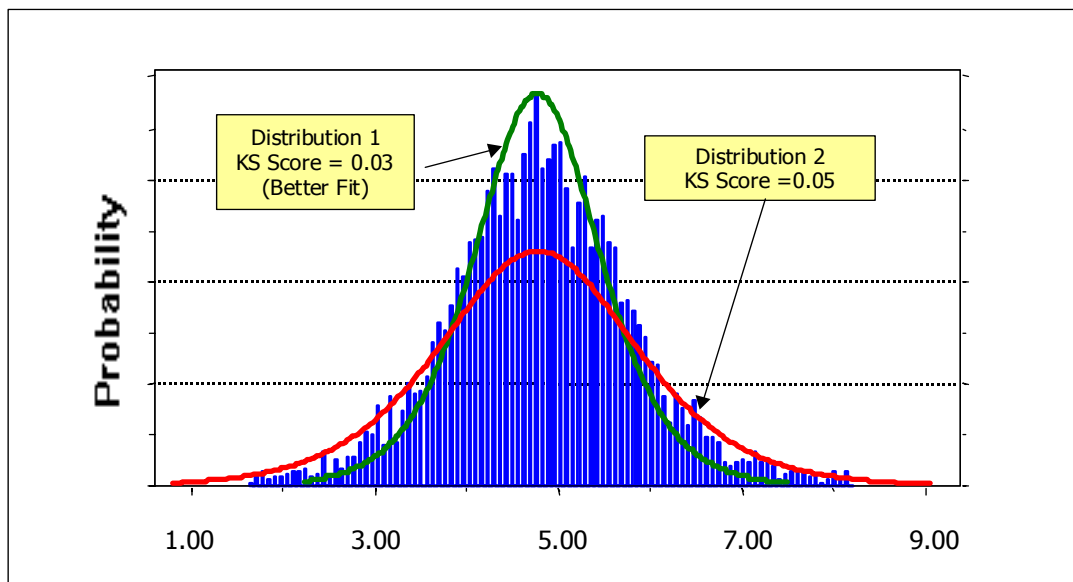


Figure 1: KS scores for 2 types of distributions for variable X

In a further refinement of goodness-of-fit tests, we could use a modified version of the Kolmogorov-Smirnov to assign more weight to the operational risk data in the tail of the distribution (relative to that in the body of the distribution). Such a technique is very useful for operational risk data since the tail of the distribution is often the focus of the analysis.

Looking at Figure 1, we can see that using a standard KS test on this sample data indicates that Distribution #1 offers the best overall fit of the data. However, if we applied modified KS tests to the data we might well be able to demonstrate that Distribution #2 fits better in the tails of the distribution. Figure 1 seems to be an instance in which using different distributions for different parts of the data might result in a better overall fit.

### Empirical vs. Actual Distributions

Up to this point, we have assumed that the “true” distribution of the data is unknown, and that it has to be estimated. Instead, we might assume that the shape of the “true” distribution is represented by the data in our dataset, and is therefore known. But because the dataset is not complete, we might wish to use a modelling technique to expand our dataset using data from the same distribution.

One approach to this problem is to resample with replacement. The procedure is carried out by randomly selecting a finite set of data points from the total collection of events. Each point is randomly selected from the entire data set (i.e., if it is chosen, it is put back in to the “bucket” to be chosen again). With such random selections of events, multiple distributions of possible points can be created from given data, with each point given equal weighting. Because no assumptions about the shape of the underlying data have to be made, this approach – often called “bootstrapping” – can be very powerful (see Box 1).

Another advantage of “resampling with replacement” is that it allows us to generate paths of simulated data that retain the properties of the historical data. Each loss is assumed to be independent. But this is usually a realistic assumption to make, and does not appear to compromise the results significantly.

The technique of resampling with replacement has also been used to determine the potential returns of catastrophe risk bonds, and weather options. Here again, data on the underlying event is not as extensive as the data available in most financial markets.

Nevertheless, the technique has limitations – especially when applied to a small sample. Most fundamentally, the approach assumes that the dataset contains the

complete universe of operational losses – i.e., all the losses that might ever be experienced. Depending on how the model is set up, the analysis can also give undue weight to specific recent events. The technique can also be computer intensive, depending on the number of factors involved.

### Bootstrapping operational risk data – An example

Suppose that we want to calculate the 95<sup>th</sup> and 99<sup>th</sup> percentile Capital-at-Risk (CaR) for an investment bank, using all relevant events in a historical loss database that are in excess of \$50 million and \$100 million. Assume that we have N events, each with a magnitude M. We need to treat the data so that we can better determine the volatility around our CaR estimation.

The procedure is carried out by sampling from the N events with replacement, to give the combinations presented in Table A. (Bootstrapping assigns equal probability to each event.)

**Table A. Combinations obtained by resampling with replacement**

Combination	CaR
55,55,55	CaR1
55,75,75	CaR2
55,87,87	CaR3
87,87,87	CaR4
75,75,75	CaR5
87,75,75	CaR6
75,75,87	CaR7
55,55,87, etc.	CaR8

Each of these sequences, or series of loss events, allows us to calculate a different set of CaR values. By averaging the different CaR calculations, we can gain a better idea of the standard deviation of the distribution around our CaR estimation.

Continuing with our example, we have the following data:

Case	Frequency	# of Events
ABOVE 50	0.14	N1
ABOVE 100	0.09	N2

**Note:** N1>N2

ABOVE 50: Selected loss events, including only losses in excess of \$50 million

ABOVE 100: Selected loss events, including losses in excess of \$100 million

Frequency is displayed as number of events per year

In effect, the “resampling with replacement” has allowed us to set up different combinations of losses from our database. Using an appropriate methodology or software<sup>1</sup> we can then use this data to calculate many CaR estimates (Table C).

**Table C. Calculating the CaR estimates**

Loss Distribution		CaR -95% (000s)	% of Deviation	CaR-99% (000s)	% of Deviation
ABOVE 50	StError	28,908		230,552	
ABOVE 50	Expected	186,838	15.47%	1,140,884	20.21%
ABOVE 100	StError	23,083		232,142	
ABOVE 100	Expected	171,070	13.49%	1,158,157	20.04%

Table C illustrates the cumulative loss distribution for the selected data, as generated by the software.

At the 95% confidence interval, capital-at-risk falls as the database threshold rises from \$50 million to \$100 million. This seems intuitively correct as the lambda (frequency) of such events falls when losses of <US\$100 million are excluded.

## Summary

Small data sets mean that modelling operational risk is a challenge. It is critical that attention be paid to how well the distributions employed by the analyst fit the empirical operational risk data. Evaluating different families of curves for different parts of the operational risk data significantly strengthens any CaR results, as long as the basis for selecting the curves can be justified.

Tests such as the KS test, and weighted KS test, are very useful in justifying the selection of distributions. Resampling with replacement can allow users to create multiple distributions for analysis, all of which are based on empirical data – thus eliminating the need to “assume” any distribution

---

The authors would like to thank Marta Johnson and José V. Hernández for their valuable assistance.

---

<sup>1</sup> In this case, the appropriate NetRisk module, RiskOps™