MODEL RISKS AND CAPITAL REQUIREMENTS

by Parit Jakhria, Stuart Jarvis and Andrew Smith

(Members of the Extreme Events working party)

1. Introduction

Insurance accounting and solvency principles frequently make reference to probability distributions. For example,

- The IASB's draft insurance contract standard refers to a "probability-weighted estimate of the future cash outflows"; a wording which is repeated in Solvency II technical provisions.
- The definition of the solvency capital requirement makes reference to probabilities of default, value-at-risk and 99.5% confidence intervals

Each entity reporting under these standards must create their own probability model, applying judgement to the statistical analysis of past data. Several probability models may be capable of explaining the data, yet with different implications for reported assets, liabilities or capital requirements. There are two distinct kinds of risks: firstly those captured within the judgements made for a single model and secondly risks associated with having chosen that model rather than an alternative.

This paper explores both of these risks in some detail:

Section 2 considers the overall aspect of choice or judgement with respect to modelling, explaining the basic need for judgement, and highlighting several broad areas where judgement / choice is manifested in the context of actuarial modelling. The chapter then considers each area, and introduces some methods of mitigating the need for judgement.

Section 3 looks into model and parameter risk, and describes classical statistical approaches to parameter error, showing constructions of confidence and prediction intervals, incorporating the T-effect, for a range of distributions and data sample sizes. It also tackles model error, with a number of examples, including a discussion of model and parameter error in the prediction of percentiles given a small random sample from an unknown distribution. Some techniques to assess model errors are also discussed.

Acknowledgements

Acknowledgements to Dr Andreas Tsanakas for his expertise on model risk, participants at various seminars including the 2012 Life Convention for valuable discussion and 'difficult questions', two anonymous scrutineers, and most of all, other working party members for stimulating discussion and ideas.

2. Judgement

Firstly, it is important to acknowledge that judgement is a necessary and inescapable part of Actuarial modelling, and it is very difficult (in fact, one could argue impossible with the exception of pathological examples) to simply avoid the need to make any choices. Moreover, as we shall explain, judgement permeates almost every aspect of modelling, which means that one may encounter the need to make choices at various levels.

Another aspect is that any judgement, by definition, depends to a certain extent on who is making the judgement, and the framework within which they need to make the judgement. In the UK, the board is ultimately responsible for all assumptions and judgements, and statutory audits in the UK assert that a set of accounts provide a 'true and fair' view, in accordance with the accounting principles.

A specific element of the audit assesses whether judgements made by management during the production of a set of 'true and fair' accounts, are reasonable. A reasonable judgement may be interpreted as within the range of conclusions an expert could draw from the available data, and will have regard to common practice in the market. The range for reasonable judgements may be narrower than parameter standard errors in a statistical sense, especially when data is limited.

An audit process also looks for errors which are unintentional human mistakes, such as the use of an incorrect tax formula or omitting an expense provision. Any such mistakes are assessed using the concept of a 'materiality' limit. This is an amount of error, expressed in currency terms, which is deemed not likely to change the decisions of the users. Material errors should be corrected but a clean audit can still be granted provided the aggregate effect of mistakes is within the materiality limit.

The exercise of judgement is not a mistake; it is possible for two reasonable judgements to differ by more than the materiality limit. Therefore, decision makers relying on financial statements should be aware that alternative judgements in the account preparations could have led to different decisions. True and fair accounts must be free of material mistakes but cannot be free of material judgement.

Thus, there is a considerable element of judgement embedded in the production of the basic 'realistic' or 'best estimate' balance sheet, even before extrapolating into the calculation of the capital required to withstand a hypothetical 1-in-200 event.

We try to broadly categorise the different manifestation of choice / judgement inherent within modelling, and point the reader to tools that potentially deal with each aspect.

2.1. Manifestations of Judgement

Judgements are an integral part of any modelling exercise, because any model is necessarily a simplified representation of the real world, and as such needs to be stripped down to its most relevant components. This ensures that the model is a useful analogy of the real world for the specific purpose that a model is required for. The process of stripping down to the bare useful components and 'calibrating' the resultant model inevitably has a large amount of judgement associated with it. In the limited context of Actuarial modelling, this judgement can broadly be thought to manifest itself in the following ways:

- Choosing which risk factors to model
- Choice of overall model framework
- Choosing individual parts of the model
- Choice of calibration methodology
- Judgements inherent within the data itself
- Choice of parameters

One can imagine that these have (broadly) decreasing levels of significance to the end results. However, the industry appears to have the greatest focus on the final (and perhaps second to last) elements of parameter choice and calibration methodology, often to the detriment of the overall choice of framework and model fitting. For example, companies may focus most of the documentation and rationale of expert judgement in the final two categories, potentially at the expense of reduced oversight and attention paid to the substantial implied judgements involved in the first two categories.

2.1.1. What risk factors to model?

Perhaps the single biggest modelling choice to be made by a company is simply what risks to model. As briefly explained above, a model can only hope to be a simplified representation of the reality it intends to model. A philosophical way of thinking about it is that any model of the universe needs to be at least as large as the universe itself, and in fact in order for us to project it into the future faster than real time, it needs to be even larger.

A key constraint on the number of risks factors to include in a model is availability of resources. Human resources, computer resources and time are required in most modelling exercise and these have to be employed in the most efficient manner. To that end, we need to choose the most appropriate aspects to model. Suppose we could simplify the modelling problem considerably, by converting the problem into a simple choice regarding the number of 'risk factors' to model. It can be shown that the number of modelling choices faced by a company is simply enormous!

Needless to say, choosing the risk factors to model is an extremely important exercise, and one that should not be taken lightly. Also important, once the modelling choices have been made is how to allow for the risk factors that are not modelled. This would need to be addressed via 'grossing up techniques'.

2.1.2. Choice of overall framework

This is quite possibly the second most significant judgement to be made within a (capital) modelling context, although it is not always appreciated as such. Of course, the materiality of the different choices depends on the particular problem at hand. We try to illustrate this in the context of aggregation methodology using a very simple case study with two risks (described as two products of equal size, A and B):

- Each risk akin to a simple product with a guaranteed £100m liability, in which not all the risks are hedge-able. There is a residual 1 in 200 risk that the assets (and capital) would lose half their value. The extra capital required at time 0 such that the product has a 99.5% probability of meeting its guarantees at time 1 is (an extra) £200m (£100m for each product).
- The two products are assumed to be uncorrelated.

Let us now consider two commonly used methods of calculating the aggregate capital requirement:

1. Using an 'external correlation matrix' approach, the answer is relatively simple. The aggregate capital requirement can be calculated as $\sqrt{Capital_A^2 + Capital_B^2}$, which is £141.6m in total and £70.8m per product.

2. Using an alternative approach of undertaking a Monte Carlo simulation of the losses of the two products assuming the distribution of the losses are lognormal (again, a popular choice amongst practitioners) with the same 1 in 200 individual capital requirements. This produces a different aggregate capital requirement of £121m and £60.5m post diversification capital requirement for each product

This simple example of the impact of the choice of the aggregation framework shows that the aggregate capital requirement can differ by 20% of the liabilities. This example is not special in any sense, in that, the two products could represent pretty much any two risk factors.

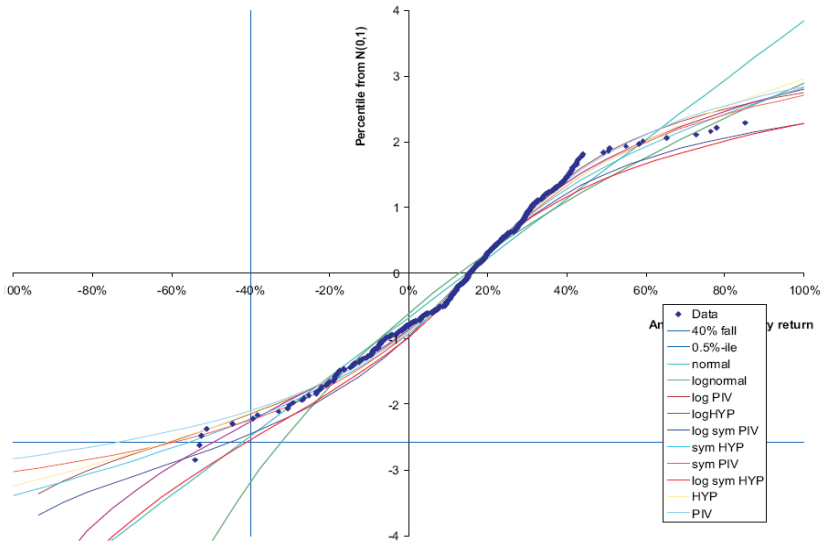
One can appreciate that there are multiple choices to be made in simply aggregating different risk factors. This further compounded by choices that need to be made in relation to exotic copula structures, non-linearities, etc.

Finally, this example only touches on a very specific aspect of framework choice; there are many other implicit judgements necessitated by trying to create a simplified representation of the real world. There is a long list of other judgements that need to be made, examples of which include:

- Using heavy models vs. lite (proxy) models
- If proxy model, choice of proxy model
- What measure is used to estimate capital e.g. VaR , tail VaR , etc
- Granularity of assets, model points
- Use of instantaneous stress approximation (time zero , time 1 or other)
- Holistic model of the business vs. detailed product specific models aggregated
- Treatment of new business
- Fungibility of capital
- Measure of correlation used

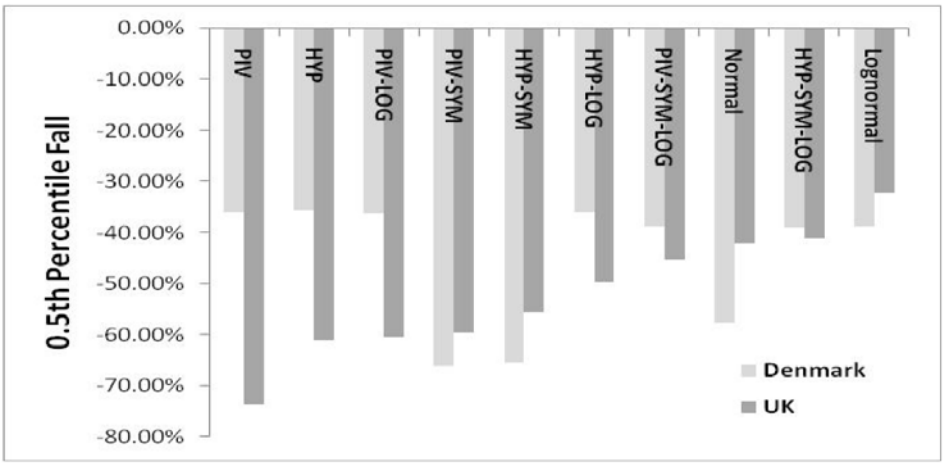
2.1.3. Choice of model components

The choice of model for each risk is an important aspect of the modelling. A previous paper, an extract of which is showed below, from this working party (Frankland et al, 2008) showed that fitting different models to the same historical data can have a range of different results, even when using relatively large amounts of data.



"Even after settling on a single data set, the fitted curves for U.K. produce a wide range of values for the 1-in-200 fall. The most extreme results are from a Pearson Type IV, applied to simple returns, which implies a fall of 75% at the 1-in-200 probability level. At the other extreme is the lognormal distribution, with a fit implying that even a 35% fall would be more extreme than the 1-in-200 event. Other distributions produce intermediate results."

The next graph shows the 0.5th percentile estimated using different models for the UK and Denmark. Interestingly, the same extreme potential model error is also true for Denmark (which also has relatively large amounts of historical data), but in this case the models that results in the most and least extreme values are completely different.



Another example of model risk is described in section 3.2 and the reader is also referred to a recent paper by Currie, Richards and Ritchie, and the subsequent discussion.

2.1.4. Choices inherent within the calibration

There are other, more commonly acknowledged calibration choices that should be noted in the same context. For example, calibration of a model is required to fit historical data to a 1 in 200 event, and one of the decisions is the choice of data period to use. Very simply, if we are using x years of data, then the model would accept the worst event within this data window as being a 1 in x event. Thus the choice of data window makes an important contribution to the results, and it is important to note that even picking any available dataset comes with a default assumption regarding the data period.

This context results in an obvious place where one may want to exercise judgement i.e. one may want to impose some views on the extremity of actual events observed. An obvious example can be constructed by using very short data series that included the recent credit crisis, which would naively overestimate the resulting extreme percentile calculated by assuming such an extreme event would occur regularly within such a short time period. Of course, it goes without saying that the reverse would also have been true when looking at short term data prior to the recent financial crisis.

Another example of judgement related to calibration is the choice of method to estimate the parameters of a model. The two approaches that are commonly used are maximum likelihood estimation and the method of moments (described further in section 3.1.4). In calibration exercises where data is limited, these different approaches can lead to very different results.

2.1.5. Judgement inherent within underlying data

Before the application of judgement to the data it is useful to consider the nature of the raw data used to calibrate the model. The data might not itself be accurate, being based on estimates, or might contain a systemic effect that would influence our interpretation of the data and perhaps the calibration.

A recent example of data that is based on estimates, which was perhaps not generally appreciated at the time, is the ONS mortality data for older ages in England and Wales. The results of the 2011 Census revealed that there were 30,000 fewer lives aged 90 years and over than expected. In absolute terms, a difference of 30,000 lives is not a large change. However, in relative terms it represents a reduction of around 15%. Between census years the exposed to risk is an estimated figure. The funnel of doubt around the estimate increases as the time since the last census increases. The ONS data, being based on a large credible dataset, is the source of most actuarial work on longevity improvements including the projection model developed by the Continuous Mortality Investigation. One implication of the recent census data is that estimates of the rate of improvement of mortality at high ages might have been significantly overstated.

An example of a systematic effect that may be present in data, and should be considered before using any data to calibrate an internal model, would be the derivation process behind a complex market index. As an illustration, Markit publishes a report on its iBoxx EUR Benchmark Index[1] that documents multiple changes to the basis of preparation. This index may be a considered a suitable starting point to construct a model of future EUR bond spread behaviour but such an analysis should consider the effect of the various changes. Of course, the report from Markit contains extensive details of the past index changes but an equivalent level of detail might not be available for other potential data sources.

[1] http://www.markit.com/assets/en/docs/products/data/indices/bond-indices/Markit_iBoxx%20EUR_Benchmark_Guide.pdf

2.1.6. Choice of parameters, and expert overlay

Whatever model we have chosen to use, we would need to supply it with suitable parameters. This can be done by using some statistical methods of choosing the best parameters, acknowledging the parameter uncertainty inherent from fitting to limited data. Section 3.1.1 looks at possible ways of quantifying the risk capital where parameters are uncertain.

In addition to genuine parameter uncertainty (assuming that past data is truly reflective of the future), we may also have uncertainty as the future conditions may be different to historic conditions (taking interest rates, for example). In this case, management still needs to make some judgements on the choice of parameters.

This judgement on model parameters can be explicit or implicit. For example, at the most explicit level, one may override the 1 in 200 stress itself, by super-imposing the views of investment experts. In many cases, the paucity of data and the changing economic landscape makes this a regular part of the capital calculation exercise. Alternatively, judgement can be more implicit in the structure of the model. For example, conditional on the form of the model, one may have prior views on certain parameters. (e.g. we may have a prior view on the volatility parameter of a lognormal distribution).

However, it should be recognised that any judgement (even though necessitated due to poor data and changing environment) is ultimately subjective. One advantage of explicit judgement (over implicit judgement) is that it is extremely transparent, and openly recognises that models and data can only go so far in terms of predicting future distributions.

For either explicit or implicit judgement, we need to recognise that although one may have a better base assumption, parameter uncertainty (section 3.1.1) still needs to be taken into account. Also, one should aim to follow a good process when coming up with the parameters, and some observations on current practice within the industry are discussed in the next section.

2.2. Key points

The salient point of this chapter is simply that although the calculation of capital requirements is extremely sensitive to judgements made, judgements are also a necessary and inescapable part of Actuarial modelling. To that end, it is important for the companies to recognise where judgement occurs, and we try and broadly categorise different areas where judgements can occur, together with some detailed examples. In particular, the importance of the choice of risk factors to model, choice of framework as well as choice of model should not be underestimated.

The next section of this paper focuses on model and parameter error, although the working party has also carried out research in some of the other areas.

3. Model and Parameter Error

3.1. Allowing for model and parameter risk

With the 'appropriate' model and, 'accurate' assumptions, we can justify statements such as "with €100m of available capital, there is a 99.5% probability of sufficiency one year from now".

However, given that a model is only intended to be a representation of reality, it is unlikely any model will be ever fully correct, or the assumptions fully accurate. This section considers how such a statement may be modified if a firm has concerns about the correctness of models or accuracy of parameter estimates. There might be several models that adequately explain the data. Even when one model is a better fit than another, we may not be able to reject the worse fitting model as a possible explanation of the data. There is a difference, however, between picking one model as the most credible explanation and picking one model as the only credible explanation. We should not discount the possibility that some initially less plausible model could subsequently turn out to be most appropriate.

Model error is one of many risks to which financial firms are exposed. We might hope to quantify model error in much the same way as other risks, by examining the potential losses arising from model mis-specification. These might then be incorporated into a risk aggregation process, making appropriate assumptions about the correlation between model risk and other risks.

In this section, we argue that model risk is of essentially different nature to other modelled risks. To describe interest rate risk, we take as given a model of how interest rates might move, test the model against past data and use this model to explore the likelihood of possible adverse shocks. A probability approach is ideal for such analysis. The probability framework is less equipped to cope with ambiguity in models. For example, we may struggle to find an empirical basis to express the likelihood of alternative models being correct.

Several different models might account for the historic data, but they might have different implied capital requirements. Percentile-based capital definitions no longer produce a unique number, but rather a scatter of numbers depending on which model is deemed to be correct. If we want the answer to be a single number, then we have to change the question.

We then give concrete examples of model and parameter risk. We consider possible ways of clarifying the 99.5%-ile question in the context of model ambiguity, and explore the impact on model output.

It is helpful to consider model ambiguity in three stages:

- Models and parameters are known to be correct. This is the (hypothetical) base case to which we compare other cases.
- Location-scale uncertainty, section 3.1.1: Past and future observations are samples from a given distribution family, but the model parameters are uncertain. Specifically we consider situations where candidate distributions are related to each other by shifting or scaling.
- Model and Location-scale uncertainty, section 3.1.2: Both the applicable model and the parameters are uncertain. For example, there may be some dependence between observations and the observations may be drawn from one of a family of fatter tailed distributions. In each case, limited data is available to test the model or fit the parameters.

3.1.1. Parameter Uncertainty in Location Scale Families

We consider an example where the underlying model is of a known shape, but where the location and scale of the distribution are subject to uncertainty. This is usually the case in practice.

We then consider three possible definitions of a percentile where parameters are uncertain.

Method	Construction	Allowance for parameter uncertainty
Substitution method.	Model parameters are estimated and substituted into the formula for the percentile given the parameters.	The objective is to get as close as possible to the answer that would be obtained were the parameters certain; there is no extra margin for the parameter uncertainty.
Confidence interval.	A confidence interval is a function of data that has (at least) a given probability of containing the true parameters. If we want to estimate the 99.5%-ile of a distribution, we could construct a confidence interval that has a 99.5% probability of containing the true 99.5%-ile.	This implies a large impact for parameter uncertainty, as we construct a parameter at a 99.5% confidence level and then look at a 99.5%-ile event given those extreme parameters.
Prediction intervals	Given a series of historic observations and unseen observations from the same distribution, a prediction interval is a function of the historic data with a given probability of containing the unseen observations.	Extreme parameter errors do not necessarily coincide with extreme percentile outcomes given the parameters, so a prediction interval captures some diversification between the two types of risk.

It is not always clear in practice, even for statutory purposes such as computing the solvency capital requirement, which, if any, of these definitions applies.

3.1.2. Comparison of Statistical Definitions to Actuarial Best Estimates

We contrast probability definitions with actuarial concepts of best estimates, as discussed, for example by Jones et al (2006, GRIT report):

> *"best estimates ... contain no allowance or margin for prudence or optimism."*
> *... " they [best estimates] are not deliberately biased upwards or downwards. "*
> *" The estimates given in the report are central estimates in the sense that they represent our best estimate of the liability for outstanding claims, with no deliberate bias towards either over or under-statement. "*

The actuarial best estimate definitions refer to a lack of deliberate bias. This suggests that provided the actuary did not intend to introduce bias, a 'best estimate' results. If there turns out to be a bias in a statistical sense, this does not disqualify an actuarial best estimate provided the bias is unintentional. This highlights the difference between actuarial best estimates defined in terms of intention, and statistically unbiased estimates defined in a mechanical way.

Our statistical definitions refer to probabilities which can in principle be tested in controlled simulation experiments, and indeed we will perform such experiments in the example that follows.

As defined above, any best estimate is a point estimate. However, there is an interval of plausible outcomes around this estimate. There are different statistical definitions for such an interval. The experiment that follows considers two of these intervals (described in section 4.1.1):

- Substitution: β-quantile
- Confidence interval with probability γ of containing the β-quantile.
- Prediction interval containing the unseen observation with probability β.

What the experiment shows is that the properties of a probability distribution do not necessarily determine a unique construction for an interval. There might be (and indeed, there are) several ways to construct intervals satisfying the definitions.

3.1.3. Five Example Distributions

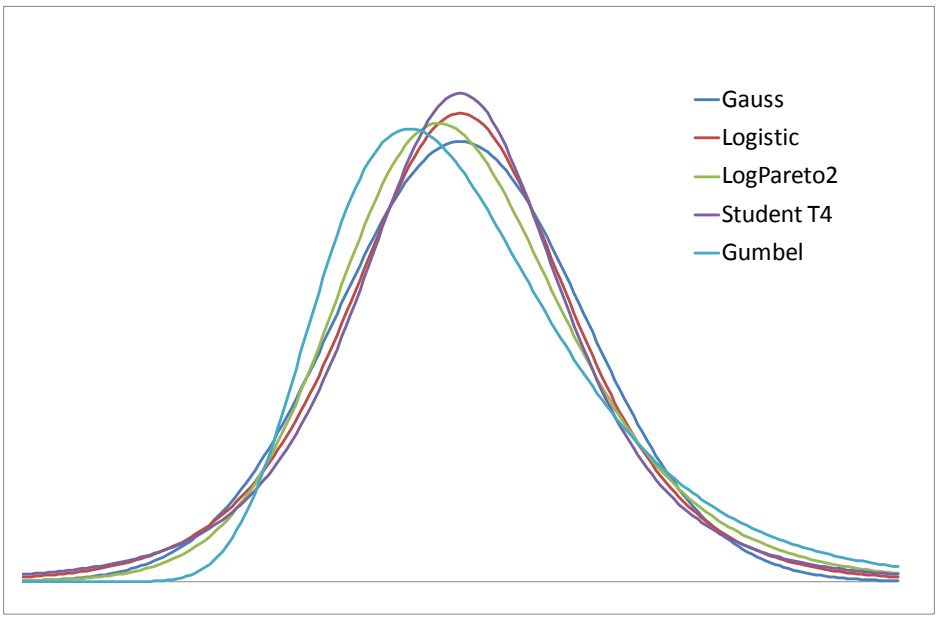
We now show some example calculations of these intervals based on five distribution families. In each case we define a standard version of a random variable *X*; the other distributions in the family are the distributions of *sX+m* where *m* can take any value and *s*>0.

Our five distributions include fatter and thinner tailed examples, and include asymmetric distributions:

Distribution	Probability Density *f*(*x*)	Cumulative Distribution Function *F*(*x*)	Inverse CDF $F^{-1}(p)$
Gauss	$\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$	$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{\xi^2}{2}\right)d\xi$	Not tractable
Logistic	$\frac{e^x}{(1+e^x)^2}$	$\frac{e^x}{1+e^x}=1-\frac{1}{1+e^x}$	$\ln\left(\frac{p}{1-p}\right)$
Log Pareto(2)	$\frac{2e^{2x}}{\left[1+e^x\right]^3}$	$\left(\frac{e^x}{1+e^x}\right)^2$	$\ln\left(\frac{\sqrt{p}}{1-\sqrt{p}}\right)$
Student T with 4 degrees of freedom	$\frac{12}{\left[4+x^2\right]^{5/2}}$	$\frac{1}{2}+\frac{x^3+6x}{2\left[4+x^2\right]^{3/2}}$	$\frac{4\sin\left\{\frac{1}{3}\sin^{-1}(2p-1)\right\}}{\sqrt{1-4\sin^2\left\{\frac{1}{3}\sin^{-1}(2p-1)\right\}}}$
Gumbel	$\exp\left[-x-e^{-x}\right]$	$\exp\left[-e^{-x}\right]$	$-\ln\{-\ln p\}$

Our families consistent of cumulative distribution functions $F\left(\frac{x-m}{s}\right)$ and density $\frac{1}{s}f\left(\frac{x-m}{s}\right)$ where *F* and *f* are taken from a (known) row of this table.

The chart below shows these probability density functions. We have chosen values of *m* and *s* for each family to make the distributions as close as possible to each other:



3.1.4. Methods of Estimating Parameters

We consider four ways of estimating model parameters given a random sample of historic data. These are as follows:

Method	Description	Formulas based on observations x_1, ... x_n
Method of moments	Choose the distribution by equating the sample mean and standard deviation to the theoretical values.	$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$ $\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2}$
Probability weighted moments	Choose the distribution by equating the sample mean and L-scale to the theoretical values.	$\hat{\lambda} = \sum_{i=1}^{n} \frac{2i-n-1}{n(n-1)} x_i$ In this expression the x_i are sorted into increasing order.
Maximum likelihood	Find the distribution that maximises the density function multiplied over all observations.	Choose *s* and *m* to maximise: $L = \prod_{i=1}^{n} \left\{ \frac{1}{s} f\left(\frac{x_i - m}{s} \right) \right\}$
Bayesian methods	Treat the parameters as random variables with a prior distribution. Calculate intervals based on a posterior distribution of parameters given the data.	Depends on choice of prior.

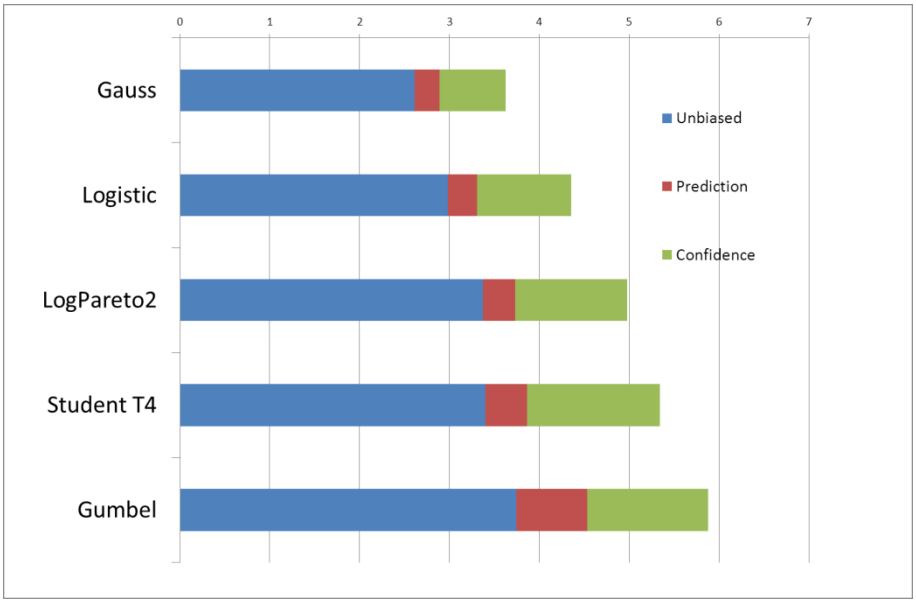
3.1.5. Interval Calculation

We now provide algorithms for calculating the different types of intervals.

Method	Substitution: β-quantile	Confidence interval with probability γ of containing the β-quantile	Prediction interval containing the unseen observation with probability β
Method of moments	$\hat{m} + \hat{s}F^{-1}(\beta)$ where $\hat{m}, \hat{s}$ are MOM estimates.	$(-\infty, \hat{\mu} + k\hat{\sigma})$ where *k* satisfies: $\mathbf{Prob}\left\{m + F^{-1}(\beta)s \le \hat{\mu} + k\hat{\sigma}\right\} = \gamma$ Equivalently: $\mathbf{Prob}\left\{\frac{m - \hat{\mu} + F^{-1}(\beta)s}{\hat{\sigma}} \le k\right\} = \gamma$	$(-\infty, \hat{\mu} + k\hat{\sigma})$ where *k* satisfies: $\mathbf{Prob}\left\{X_{n+1} \le \hat{\mu} + k\hat{\sigma}\right\} = \beta$ Equivalently: $\mathbf{Prob}\left\{\frac{X_{n+1} - \hat{\mu}}{\hat{\sigma}} \le k\right\} = \beta$
Probability weighted moments	$\hat{m} + \hat{s}F^{-1}(\beta)$ where $\hat{m}, \hat{s}$ are PWM estimates.		
Maximum likelihood	$\hat{m} + \hat{s}F^{-1}(\beta)$ where $\hat{m}, \hat{s}$ are maximum likelihood estimates.		
Bayesian methods	$\hat{m} + \hat{s}F^{-1}(\beta)$ where $\hat{m}, \hat{s}$ are means under the posterior distribution.	(0,*k*) where: $\mathbf{Prob}\left\{m + sF^{-1}(\beta) \le k\right\} = \gamma$ Under the posterior distribution for (*m*,s)	(0,*k*) where: $\mathbf{E}F\left(\frac{k-m}{s}\right) = \beta$ Under the posterior distribution for (*m*,s)

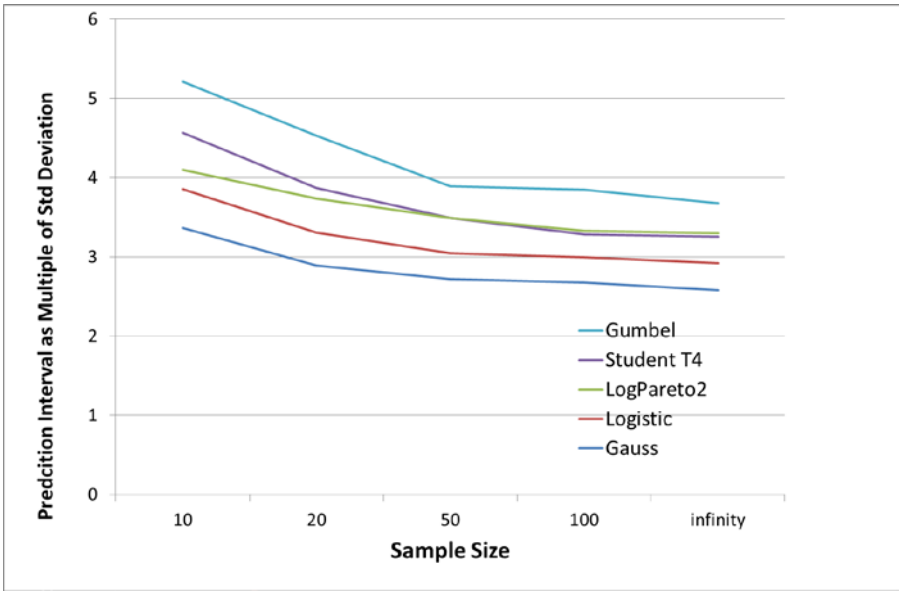
When talking about probabilities, care needs to be taken in respect of the set over which averages are calculated. The frequentist methods (method of moments, probability weighted moments, maximum likelihood) measure probabilities over alternative data sets. The Bayesian statements average over possible alternative parameter sets but not over any data sets besides the one that actually emerged. A 99.5% interval under a frequentist approach is not the same as a 99.5% Bayesian interval.

We use Monte Carlo methods to calculate confidence and prediction intervals. We show these in the case of 20 data points, using the method of moments, based on the 99.5%-ile. These measures differ only because the data is limited; as the data increases these are all consistent estimators of the 'true' percentile. None of these is *the* right answer; the different numbers simply answer different questions.



The interval size (expressed as a number of sample standard deviations) varies from one distribution to another. Note that although the T4 distribution has the fattest tails in an asymptotic sense (it has a power law tail while the others are exponential), the Gumbel produces the largest confidence intervals. This is because the asymptotic point at which the T4 distribution becomes fatter, lies way beyond the 99.5%-ile in which we are interested.

We also consider how the intervals vary by the number of observations. The prediction intervals are shown below. We can see that the prediction interval requires a smaller number of standard deviations as the sample size increases, because a better supply of data reduces the errors in parameter estimates:



3.2. More on model risk

We have described ways of calculating 99.5%-iles in the presence of parameter risk, at least in the case of location-scale probability distribution families.

Uncertainty about shape parameters or about underlying models is more difficult to address. The difficulty is constructing an interval that covers 99.5% of future outcomes, uniformly across a broad family of models. In general, the best we can hope is an inequality, so that at least 99.5% is covered.

We set out two examples of the impact of using different models to estimate variables of interests. These show that the choice of the models has a very material impact of the estimate of the variable of interest. We then set out some techniques that can be used to assess the impact of model errors. The techniques described later are by no means an a exhaustive list, but a list of the techniques commonly used in practice.

3.2.1. Example of Model Risk

A recent paper by Currie, Richards and Ritchie compares seven different models of longevity. They compare the value of a life annuity to a 70 year-old male, limited to 35 years and discounted at 3%. The paper constructs a probability distribution forecast for this annuity in one year's time using different models of mortality improvements. We reproduce table 5 from that paper, below:

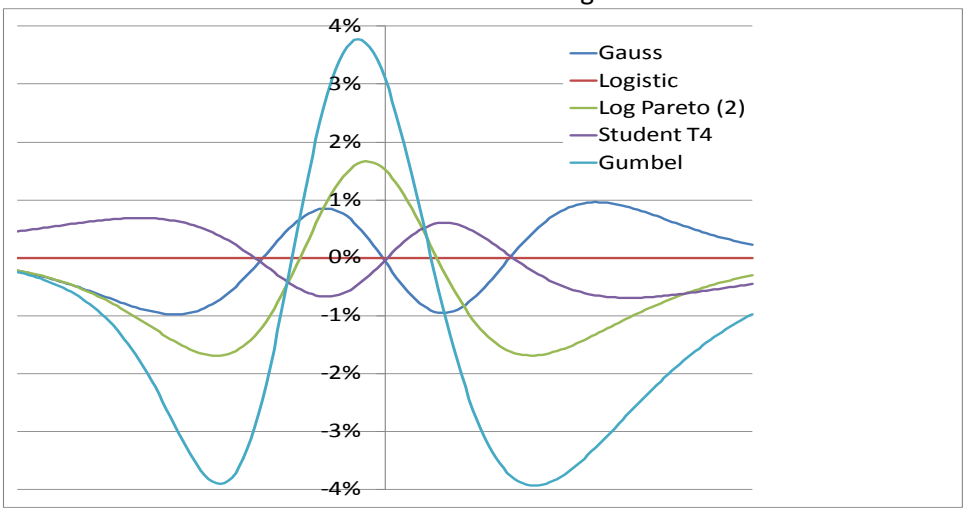
	Value of $\bar{a}^{3\%}_{70:\overline{35}|}$	
Model (Appendix)	(a) average value	(b) 99.5^{th} percentile
LC (A1)	12.14	12.72
DDE (A1)	12.15	12.77
LC(S) (A1)	12.15	12.76
CBD Gompertz (A2)	11.98	12.44
CBD *P*-spline (A2)	11.89	12.36
APC (A3)	12.61	13.04
2DAP (A4)	12.80	13.69

Please refer to the paper [Currie et all, 2012] for more details about the mortality models. We might hope for a degree of consensus between the different modelling approaches, but these figures show the opposite. Indeed, a future annuity value of 12.5 lies below the mean for two of the models, but above the 99.5%-ile for two other models.

Link - http://www.actuaries.org.uk/sites/all/files/documents/pdf/value-risk-framework-longevity-risk-printversionupdated20121109.pdf

3.2.2. Another example: Gauss and Gumbel Distributions

Let us return to the set of five probability distributions described in in section 4.1.3. The distribution functions are sufficiently close that the curves are difficult to distinguish by eye. However, we can separate them by subtracting one of the cumulative distribution functions, for example the logistic. The chart shows the CDF for each distribution minus the logistic CDF.



Although these distributions have different shapes, the distribution functions differ nowhere by more than 4%. This suggests that it will be difficult to separate the distributions using statistical tests based on small sample sizes.

Goodness-of-fit tests such as Kolmogorov-Smirnov and Anderson-Darling have low power when data is scarce. The table shows the power of KS and AD tests with 20 data points. The chance of rejecting an incorrect model is often only marginally better than the chance of rejecting the correct model, and in a few cases the correct model is more likely rejected than an incorrect one.

Probability of rejecting a model fitted using the method of moments, tested using a Kolmogorov-Smirnov (or Anderson-Darling) statistic based on 95% confidence and 20 observations.

		Fitted distribution				
		Gauss	Logistic	Log Pareto 2	Student T4	Gumbel
True distribution	Gauss	5.0% (5.0%)	3.7% (3.8%)	6.8% (7.3%)	2.6% (4.0%)	18.0% (28.3%)
	Logistic	8.7% (10.7%)	5.0% (5.0%)	9.0% (10.0 %)	2.6% (2.8%)	21.8% (33.8%)
	Log Pareto(2)	10.5% (13.2%)	6.9% (7.5%)	5.0% (5.0%)	3.9% (4.3%)	10.0% (14.7%)
	Student T4	17.4% (22.3%)	10.5% (12.1%)	15.6% (18.0%)	5.0% (5.0%)	28.7% (40.4%)
	Gumbel	20.3% (27.5%)	15.4% (19.0%)	7.2% (7.9%)	9.8% (11.4%)	5.0% (32.2%)

Even with samples of 200 or more, it is common not to reject any of our five models. This implies that model risk remains relevant for applications including scenario generators, longevity forecasts or estimates of reserve variability in general insurance.

3.3. Techniques to assess model errors

The examples described above shows that there are a range of models that can be used to undertake a task. This motivates a quest to understand any error introduced by model choice.

Model risk remains relevant for virtually all aspects of actuarial modelling, from longevity forecasting to capital aggregation. Are we then at risk of channelling too much energy and expense into the 'holy grail' route of modelling, justifying each parameter and component of a single model? Does our governance process consider model risk, or does board approval for one model implicitly entail rejection for all others?

Given the inevitable uncertainty in which model is correct, how can we make any progress at all? There are several possible ways to proceed.

- Pick a standard distribution, for example the Gaussian distribution, on the basis that it is not rejected. The statistical mistake here is to confuse 'not rejected' with 'accepted'.

- Taking the prudent approach – the highest 99.5%-ile from all the models

- Build a 'hyper-model' which simulates data from a mixture of the available models, although expert judgement is still needed to assess prior weights. Ian Cook (2011) has described this approach in more details on the context of catastrophe models.

Industry collaboration on validation standards may lead to generally accepted practices. For example, over a period of time practice may converge on a requirement to demonstrate at least 99.5% confidence if the data happens to come from a Gaussian or logistic distribution, but not for Student T or Gumbel distributions. This may not always be a good thing, particularly if it leads to a false sense of security.

A commonly used mitigant against model and parameter error already in use within the industry is to carry out scenario and stress testing approaches. These were described in section **Error! Reference source not found.**. A more statistical technique that may prove useful is the concept of robust statistics and ambiguity sets – described below:

3.3.1. Robust Statistics and Ambiguity Sets

Robust Statistics is the study of techniques that can be justified across a range of possible models rather than a single model. It can be used to help derive prediction intervals that are robust to the choice of distribution

Suppose, using the method of moments we wanted to construct a prediction interval of the form $\left(-\infty, \hat{\mu} + y\hat{\sigma}\right)$ valid across a class of distributions (this set is known as the *ambiguity set*). We cannot achieve exact α-coverage for all distributions. We can, however, achieve *at least* α-coverage for a class of distributions simply by taking the largest *y* across that class. For example, we might specify that the methodology should cover at least a proportion α of observations X_{n+1} for an ambiguity class including Gaussian and logistic distributions. In this case, our chart in section 4.1.4 shows that the logistic distribution produces the highest *y* and so the prediction interval is defined based on the logistic distribution, knowing this is conservative for other distributions in the specification of the ambiguity class.

Robustness for a given ambiguity class says nothing about models outside the class. We might construct prediction intervals robust across uniform, normal and logistic distributions but these intervals would not be valid for Gumbel distributions. They would also not be valid if other assumptions are violated – if the observations are not independent or are drawn from more than one distribution.

The size of the ambiguity set is a key determinant of prediction interval size. A bigger ambiguity set means a harsher test of coverage and larger prediction intervals. It also means more wastage, in the sense that if the true distribution is one of the more benign from the ambiguity set then the prediction interval will contain a higher than intended proportion of future observations.

Nassim Taleb (2007) has popularised the notion of 'Black Swan' events, that is, unpredictable events that arise with no prior warning (and by definition unpredictable with models existing prior to the event). Models popular in the mid 2000'S, many based on Gaussian distributions, were criticised for producing too few Black Swan events and therefore understating risks. Black Swan events also highlight the classical philosophical 'problem of induction' that inference from past events can only work to the extent that the same model applies in the past as in the future. We could define a 'Black Swan' model as one where one set of formulas applied until yesterday, and tomorrow another unrelated set of formulas takes over. There can be no robust statistical procedure that works for such models as the past data is of no value for future inference. It may potentially be the case that the world has fundamentally and unrecognisably changed, but one also needs to be extremely careful that less likely models are not rejected early on in the decision making process in favour of the most plausible model without allowing for model risk. One conclusion is that any claims of robustness in respect of a statistical procedure should include a careful description of the ambiguity set against which the procedure is robust. This comment may also to the previous section in the context of what risks to model, and the framework within which to model risks.

3.4. Key points

This section focused on model and parameter error and considers in details two cases, where observations are from a given distribution but there is uncertainty about the parameters (parameter error) or where we are uncertain about the actual model itself (model error).

It also highlighted the important case that when we have model or parameter uncertainty, the capital estimation problem itself is not unambiguous, going on to discuss three possible definitions of a percentile where parameters are uncertain. We also provided an extended example, based on different methods of estimating parameters and based on different possible distributions.

Finally, we also discussed different techniques to assess model errors, and introduced the idea of robust statistics and ambiguity sets, which look at techniques that apply over a range of candidate models, rather than a single model that is considered to be most plausible, finishing with a caution to properly allow for model risk before labelling too many events as 'Black Swans'.

4. Summary

4.1. Concluding thoughts

In this paper, we set ourselves the task of tackling some of the challenges involved in estimating extreme percentiles of profits for complex financial institutions given limited data, and using models that are at best an approximate representation of reality.

It might be argued that any attempt to extrapolate from limited data to extreme percentiles is doomed to failure. For example, Rebonato derides the use of 'science fiction' percentiles. At the other extreme, there is a risk of professing more belief in our models than is warranted, a risk exacerbated by tests requiring firms to demonstrate model use to regulators in order to receive model approval.

We have argued that extreme percentiles can be estimated, but clarity is required about the problem to be solved (for example, unbiased percentile estimate, confidence interval for a percentile, or prediction intervals). Clarity is also required about the range (or ambiguity set) of mathematical models for which an approach is required to work.

Available resources usually constrain firms to pick a single internal model for decisions, subject to stress and scenario testing. However, we should always remain cognisant of the fact that other models could just as well have been picked; accepting one model does not imply all others are rejected. It is very rare to have a solid basis for believing the model we have is the only one.

The regulatory process and regulatory environment often discourages deviations from what is seen to be best practice among peers. Given all-pervasive model uncertainty, it is not practical to explore all alternatives, and it is inevitable that some modelling practices are social constructs, reflecting cultural aspects as much as statistical influence. We have to learn to live with this: cultural context is a bad thing only when it masquerades as hard science.

This paper was originally motivated by alarm at the extent to which risk models failed to predict outcomes in the 2008 financial crisis. We have tried to highlight conceptual pitfalls to avoid, and we have highlighted specific remedies for model and parameter risks, judgemental aspects and computational approximations. It is our hope that these techniques will allow actuaries to close the gap between the risks we capture in our models and those revealed in the wake of financial losses.

References

Cook, Ian M (2011).Using Multiple Catastrophe Models. Institute & Faculty of actuaries (slides). http://www.actuaries.org.uk/sites/all/files/documents/pdf/plenary-5-ian-cook.pdf

R.M. Cooke, L. H.J. Goossens, Procedures Guide for Structured Expert Judgement, Delft University of Technology Delft, June 1999

Currie, I. D, Richards, S. J. and Ritchie, G. P. (2012) A value-at-risk framework for longevity trend risk. BAJ, forthcoming.

European Commission – Procedures guide for structured expert judgment (1999). ftp://ftp.cordis.europa.eu/pub/fp5-euratom/docs/eur18820_en.pdf

Frankland, R (chair), Biffis E, Dullaway D, Eshun S, Holtham A, Smith A, Varnell, E and Wilkins T (2008). The Modelling of Extreme Events. British Actuarial Journal. http://www.actuaries.org.uk/sites/all/files/documents/pdf/sm20081103.pdf

G.R. Grimmett and D.R. Stirzaker (1982). Probability and random processes. Clarendon press Oxford

Howie, David (2002). Interpreting probability. Controversies and developments in the early twentieth century. Cambridge University Press.

A. R. Jones, P. J. Copeman, E. R. Gibson, N. J. S. Line, J. A. Lowe, P. Martin, P. N. Matthews and D. S. Powell (2006). A Change Agenda for Reserving. Report of the General Insurance Reserving Issues Taskforce. British Actuarial Journal / Volume 12 / Issue 03 / September 2006, pp 435-599

Michael Ashcroft, Roger Austin, Peter Scolley (on behalf of Solvency & Capital Management Research Group), Expert judgement on expert judgement, Life Conference 2012

Oeppen and Vaupel (2002) Broken limits to Life Expectancy. http://user.demogr.mpg.de/jwv/pdf/scienceMay2002.pdf

Ouchi, F, A Literature Review on the Use of Expert Opinion in Probabilistic Risk Analysis, World Bank Policy Research Working Paper 3201, February 2004

Rebonato, R (2007). Plight of the Fortune Tellers: Why We Need to Manage Financial Risk Differently. Princeton University Press

Rothwell, M & others (2010). Winners' Curse – the Unmodelled Impact of Competition. http://www.actuaries.org.uk/sites/all/files/documents/pdf/winnerscurse-mainreport.pdf

Smith, A D and Thomas, R G (2002) Positive Theory and Actuarial Practice. The Actuary. http://www.guythomas.org.uk/pdf/posth.pdf

W.K. Hastings (1970). Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57(1):97-109.