British Actuarial Journal, Vol. 20, part 2, pp. 366–386. © Institute and Faculty of Actuaries 2015
doi:10.1017/S1357321715000094

Quantitative approaches to model uncertainty: update on recent research

Abstract of the London Discussion

[Institute and Faculty of Actuaries, Sessional Research Event, 29 October 2014]

The Chairman (Mr M. H. D. Kemp, F.I.A.): I now ask Dr Tsanakas to start the presentation and provide some background to the presentation and why we are having this meeting today.

Dr A. Tsanakas (introducing the paper): This is an update on recent research in the area of quantifying model uncertainty. Part of the work we present was sponsored by a grant from the Institute and Faculty of Actuaries.

There is not one specific sessional paper that relates to this work: this is a review of related work. Both Mr Smith and I have worked in this area, mostly independently, but in quite frequent dialogue for a number of years.

I will cover what we know and what we do not know in the context of statistical modelling, and talk about statistical estimation bias. I will then handover to Mr Smith, who will cover a variety of aspects of the same problem of quantifying model uncertainty.

The plot in Figure 1 shows different estimates of distributions of yearly equity returns for the FTSE 100 based on 30 years of experience.

When you do this, you can start at any date in the year and start calculating from that date. There is huge sensitivity to the starting date that you choose. On the left, you have a distribution derived if you start in August. On the right, you have a distribution when you start in October. On the left, a 1-in-200-year event corresponds to a 35% fall in equities. On the right, it is 50%. The same statistical scenario leads to very different implications.

The only difference between the two is a completely arbitrary choice of starting date. There is no judgement separating the two.

That goes some way to demonstrate the level of uncertainty to which we are exposed, especially when dealing with small or even moderate data sets.

Of course, regulators have something to say about model error and model uncertainty. Under Solvency II, two things have been said. One is that the output of the model should not include a material error or estimation error. This is a nice thing to require but one we cannot satisfy. Even in principle, we cannot provide a guarantee that models are error-free. We can never say that our estimates are correct.

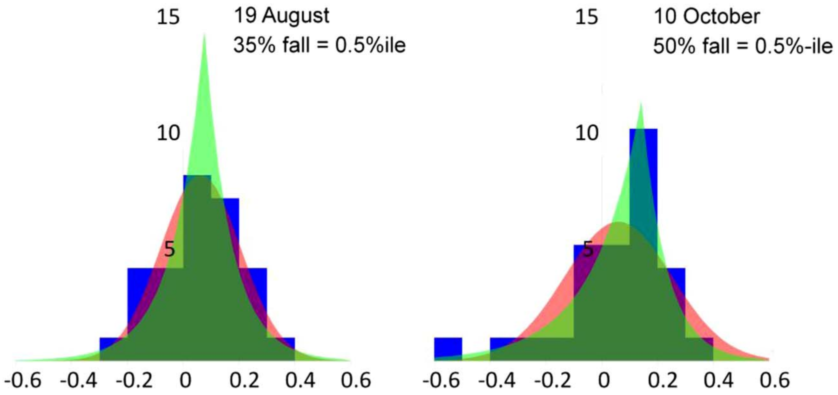


Figure 1. Equity return distributions (FTSE 100). Annual fit at two different rates

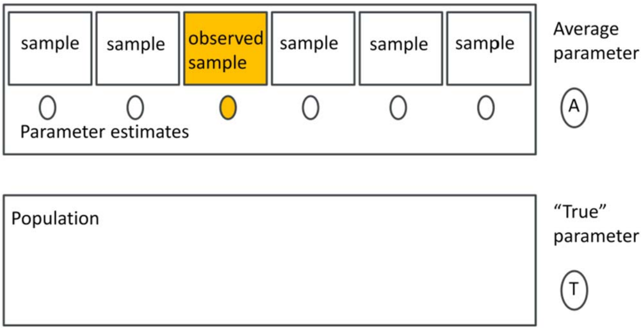


Figure 2. Statistical definition of bias = A-T

The other requirement that comes through Solvency II is that, wherever possible, the probability distribution should be adjusted to account for model error. We can do something about this, at least in principle. It may not be easy but at least it is a problem that we can start to address.

The way to think about this sort of problem is using statistical reasoning. You are required to estimate extreme percentiles and extreme events more generally. You can never say that your estimates are right.

What you can say something about is whether the method that you use in calculating those estimates was good according to a particular statistical quality criterion. This will lead you towards finding an appropriate way to adjust probability distributions for model uncertainty.

The question is: how are we going to provide this adjustment? What do we mean by a "good method"?

We are beginning to discuss biases in statistical estimation. For this, I will start with a very basic definition. The situation here is that you have an observed sample which corresponds to the square yellow area in Figure 2. What you want is to work out a particular parameter from the sample. So, you apply some sort of algorithm and the result will be something like the yellow oval.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

But now, what you have to consider, is that not only is the future random but also the past. The observations that you have are themselves the result of a random process. You may have this particular sample which may lead to this particular parameter estimate. However, you could have had a different sample. Each different sample that you could have observed would have given you a different parameter estimate.

This observation is the foundation of statistical inference. It is rarely discussed in depth because it is fairly counterintuitive. We think in terms of probability modelling, which makes us think about different futures. Here in statistical inference, we have to think about different ways that the past could have been.

When we talk about bias, we think about the different parameter estimates that we could have worked out and different samples that could have been observed, not only the ones we have seen but all the others we have not had the chance of experiencing.

If we take the average of the parameter estimates over those alternative histories and find that this average is the same as a known parameter, then we say that we have an unbiased estimate. This is a very well-established statistical concept.

The problem with bias is that you have to decide which quantity you want to be unbiased. You can start collecting some data (pre-calibration losses), and, as we see in Figure 3, you can work out moments like means or standard deviations, parameters and your required capital. At each of those points, you can require that you have unbiased estimates. But if, for example, you have an unbiased estimate of model parameters, you end up with a biased estimate of capital. If you have an unbiased estimate of capital, you have a biased estimate of standard deviations, because all of these are non-linear transformations of each other. So, we have to focus specifically on the quantity we think is important or interesting.

As part of this, what we think is interesting, especially for the purposes of solvency-type calculations, is the difference between two quantities. One is the random quantity, which is the loss you will experience in the next period (a random quantity because it appears in the future). Then, there is the difference between that loss and your capital. The capital itself is a function of your data that has been estimated in some way. These data, following the previous arguments, can themselves be considered as random.

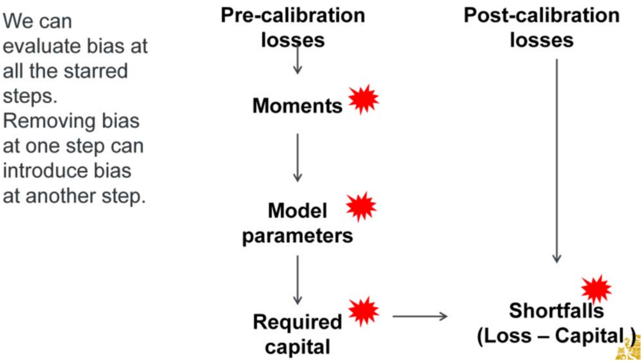

We can evaluate bias at all the starred steps. Removing bias at one step can introduce bias at another step.
Pre-calibration losses
Moments
Model parameters
Required capital
Post-calibration losses
Shortfalls (Loss – Capital)

Figure 3. Method of moments risk calibration

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

We call this difference the “shortfall”. The shortfall is the difference between the losses we will experience and the capital we have estimated. Both of these are uncertain in their own ways. One corresponds to an uncertain future and the other to an uncertain past.

We also need to consider “events not in the data”. These are extreme and rare events which may not be present in your observed sample. There may be some big losses that could have taken place but we happened not to experience them. Within those possible imagined samples, where these losses take place, they would give you a very different parameter estimate.

The problem is that if in our observed sample, we do not have such events, then we probably understate the capital. If we had such extreme events in a sample, we would be overstating the capital. The only way to think about these constructively is to think about the distribution of estimates across all different possible samples, considering both scenarios where the extreme events happen and where they do not happen.

Let us introduce a formal framework for discussing these ideas. What we are interested in is, say, a 99% value-at-risk (VaR) for a particular random variable, call it Y. That is our future loss. What we do is collect a sample, X, and, keeping things simple, assume that the data have the same distribution as the future loss.

Now you have something called an estimation procedure. The estimation procedure is nothing but some function, “g”, which can be extremely simple, or extremely complicated. What this function does is take your data and work out a VaR estimate from it.

Now we have to pose a requirement for the quality of our VaR estimate. What does VaR do? We have defined it as the value of capital such that the probability of the loss being smaller than the capital is 99%. So, essentially, what we are requiring is that only once in 100 time periods should we observe a loss that is bigger than our capital.

In our world, “shortfall unbiasedness” is the statistical criterion for estimators to satisfy. We the same condition in place, but now what we are comparing is the future loss with the estimated capital. The probability of the future loss being smaller than the estimated capital has to be 99%. But now both sides are random. This is the big difference.

There is a specific precedent in the literature on this subject. We thought we had invented it independently, but we found out that estimators satisfying this condition exist in statistics under the name of predictive limits, so there is a noble tradition to this sort of analysis.

One way of thinking about this is to consider backtesting. When we do backtesting of capital requirements, we record historical VaR estimates and historical losses, and then count the number of times the historical losses have been above VaR estimates. We call these “violations”. If the violations are frequent, then it means that the VaR estimate does not work.

What we are doing is something like that. We call this “Monte Carlo backtest”. Think about how you would evaluate the probability of a future loss being smaller than estimated capital. What you would do is to simulate a lot of futures from your loss and also simulate a lot of past data histories (in this, simulating many capital estimates). Then you would compare the two and work out how often a simulated future loss is bigger than the simulated capital estimate. This is the comparison that we make.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

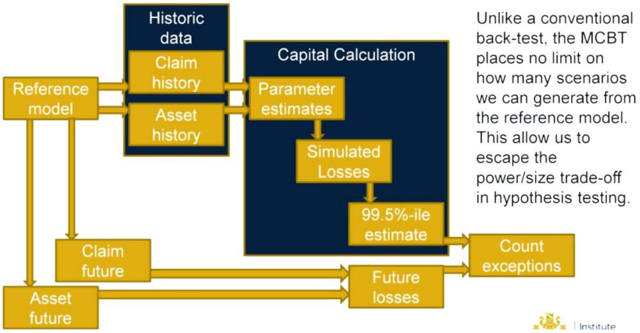
If you view this schematically, as in Figure 4, what you have is the following situation. You start with a reference model that you have estimated. You say, “let us now pretend that this model is the ‘true’ one”. We simulate from it in two ways.

On the one hand, we can simulate future claims and assets, which gives us what the losses would be in the next time period in each scenario. At the same time, we use a parallel path and also simulate claim and asset histories, which lead to parameter estimates, themselves leading to capital estimates.

We compare the future losses with capital estimates and count the number of exceptions (violations).

What sort of results can be seen from such an approach? The first example in Figure 5 is a pedagogical one, involving the normal distribution.

What is plotted here is the expected frequency of violations: the expected frequency of future losses being larger than estimated capital. We assume that the underlying reference model is normal and the standard maximum likelihood estimation (MLE) procedure is followed. The two lines correspond to a 99th and 99.5th percentile. On the vertical axis we have the frequency of exceptions and on the horizontal axis we have the sample size.



Unlike a conventional back-test, the MCBT places no limit on how many scenarios we can generate from the reference model. This allow us to escape the power/size trade-off in hypothesis testing.

Figure 4. The Monte Carlo backtest (MCBT)

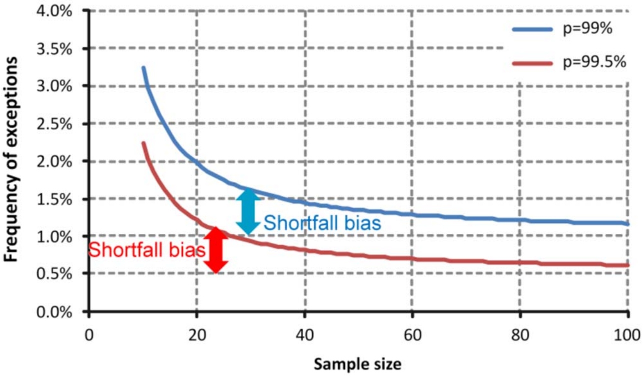


Figure 5. Normal distribution without bias correction

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

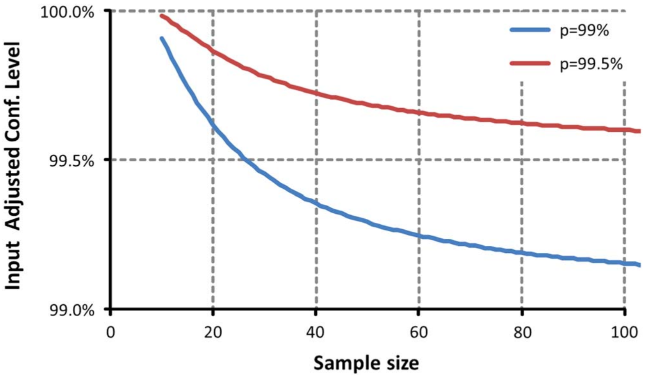


Figure 6. Adjusting input confidence levels to achieve desired exception rate

From the blue curve, you see that if you only have 20 observations, your frequency of exceptions is around 2% as opposed to 1%, which is what you would like to have. If you then compare 2% with the 1% of your nominal exception rate, the difference between those two is what we call shortfall bias in this context: the additional frequency of violations that you would observe because of parameter uncertainty. As your sample size increases, this curve slopes downwards, so things become better but they do not ever become quite right.

What can you do about it? Within a simple example like this, you can do something pretty straightforward, as in Figure 6. You can find an algorithm for adjusting your capital estimates such that your expected frequency of exceptions is the correct one, the nominal one of 1% or 0.5%, for example.

One of the ways of doing this is to adjust the confidence level of your capital calculation. You can say that you start at 99% as the nominal level but if you only have, let us say, 40 data points, you will have to use a confidence level close to 99.4%. As the sample size increases, the penalty becomes smaller. This in itself, albeit in a simple example, is a very explicit capital add-on. It essentially addresses the regulatory concern that capital estimates have to be adjusted.

It is important to note that this does not mean that if you follow the above process, your capital estimates will be accurate. If you have 20 items of data, you have 20 items of data and that is all the information you have. This problem will not be solved with maths. What satisfying the quality criterion of shortfall unbiasedness means is that on average you will derive the right exception frequency.

When can such a procedure be followed? When and in which sort of circumstances can you produce this sort of correction easily? If the distribution is nicely behaved, something like a lognormal, log-*T* or a Pareto, you can do this easily.

If a distribution is not so well behaved, you can still do something which derives a close enough approximation. You can use an approximate method to achieve the same results, more or less.

However, the problem is that the distribution has to be known. That is the elephant in the room. If you do not know the distribution family, you cannot do any of those adjustments.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

If we are to consider this problem of needing to know the distribution, we have to consider the process by which the distribution is chosen, and consider this as part of the capital estimation procedure.

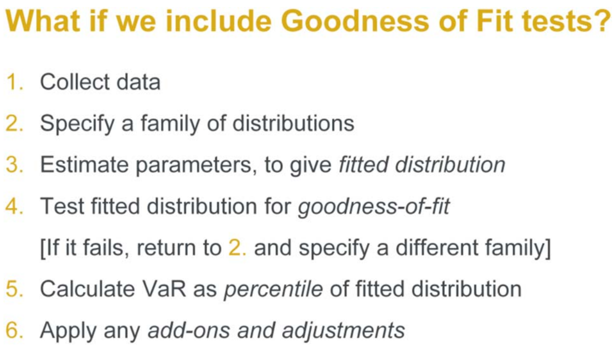
Figure 7 shows one way of working out the distribution. You collect the data and decide what sort of family of distributions you are using – let us say normal. You estimate the parameters and obtain the fitted distribution and then perform a standard goodness-of-fit test. If your distribution is not rejected you stop. If it is rejected, you look for another one. You go all the way until you find a distribution that fits and is not rejected.

Figure 8 shows the output from such a simulation experiment. We are looking at VaR estimation at 99%. We have two different reference models from which futures and data are simulated. One is a normal distribution and the other is a *t* distribution with 4 degrees of freedom, corresponding to the two columns. In the rows we have three different estimation procedures.

One is a normal MLE, so we assume a normal distribution, apply an MLE and make an adjustment to correct for shortfall bias. The second one is a normal MLE, without making such an adjustment.

What we see is if the reference model is normal and we apply the normal MLE with adjustment, then frequency of exception is 1%, which is as it should be.

If we do not apply the correction, it is 1.6%. What if the model was wrong? If the reference model was a *t* distribution, things get even worse because then the exception rate moves from 1.6% to 2.2%. So, the model error increases the frequency of exceptions even further.

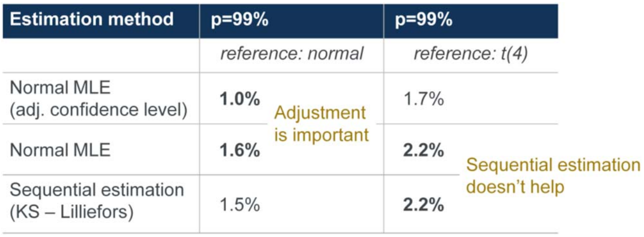


What if we include Goodness of Fit tests?

1. Collect data
2. Specify a family of distributions
3. Estimate parameters, to give *fitted distribution*
4. Test fitted distribution for *goodness-of-fit*

   [If it fails, return to 2. and specify a different family]
5. Calculate VaR as *percentile* of fitted distribution
6. Apply any *add-ons and adjustments*

Figure 7. Distribution fitting process



Estimation method	p=99%	p=99%
	reference: normal	*reference: t(4)*
Normal MLE (adj. confidence level)	**1.0%**	1.7%
Normal MLE	**1.6%**	**2.2%**
Sequential estimation (KS – Lilliefors)	1.5%	**2.2%**

Adjustment is important

Sequential estimation doesn't help

Figure 8. From parameter to model uncertainty. Frequency of exceptions, sample size = 30. MLE, maximum likelihood estimation; KS, Kolmogorov–Smirnov

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

Can we bring this down by applying a model fitting procedure, as described before? In fact, it does not help at all. The reason is that with this estimation procedure, we start by first fitting a normal model and only if it is rejected do we consider other models leading up to the *t* distribution. But with small data sizes, you usually do not end up rejecting any model, so you are stuck with the first one you tested, which happens in this case to be the wrong one.

All the discussion so far has been on VaR. But what about other risk measures? Can the concept of shortfall bias as a quality criterion for capital estimates be transferred to other risk measures?

The shortfall unbiasedness criterion that we described was that the probability of the future loss being smaller than the estimated capital is, say, 99%. That is mathematically equivalent to saying that if we take the VaR of the shortfall, that is, of the difference between the future loss and the estimated capital, this should be 0.

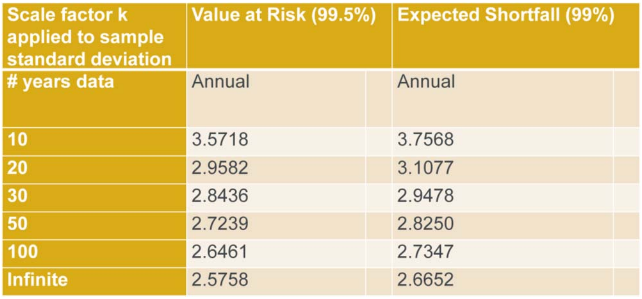
Mathematically, this is equivalent to what we said before. If you take the difference between future loss and estimated capital, both random, the VaR of the shortfall must be 0.

To move to another risk measure like tail-value-at-risk (TVaR), the only thing you do is change the risk measure. Then, you require that the TVaR of the shortfall is 0.

What sort of adjustments do you need to perform to achieve this end? In Figure 9, we show results based again on a normal model. What the figure shows is the number by which you have to multiply the estimated standard deviation in order to satisfy the shortfall unbiasedness criterion for VaR and TVaR.

If you had an infinite number of data, you would have exactly the right standard deviation. Then, the last row contains the that gives you a capital for VaR and TVaR. But because you have a limited sample, as the sample size decreases, the multiple of the standard deviation that we have to use becomes bigger and bigger. So, you need to increase capital by up to 40%. That is both for VaR and TVaR. Change of risk measure does not save you. This demonstrates quite clearly the costs of parameter uncertainty in the given context.

Mr A. D. Smith: I am going to talk about the work we have done in relation to ambiguity and robustness. One of the problems that we face when we are fitting models is how to choose what universe of models to fit. You could decide to fit a small class of models with a small number of



Scale factor k applied to sample standard deviation	Value at Risk (99.5%)	Expected Shortfall (99%)
# years data	Annual	Annual
10	3.5718	3.7568
20	2.9582	3.1077
30	2.8436	2.9478
50	2.7239	2.8250
100	2.6461	2.7347
Infinite	2.5758	2.6652

Figure 9. Example calculations of unbiased shortfalls. Figures based on Monte Carlo backtest, assuming a normal random walk

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

parameters. What you know then is that you will probably have some stability in those parameter estimates. But then there is also very little hope that your small class of models that you have picked contains one that is really like the data.

An alternative would be a much more complicated model with a larger number of parameters. Then you have a bit more reason to be optimistic that the true model, if there is one, looks something like one of the models in your class. You are then caught with another problem: you have a large number of parameters relative to the number of the data points. So, you add one to your data points and the number completely changes and the whole thing is unstable. That is something with which many of us grapple on a daily basis.

The body of research called robust statistics gives us a way out of this bind. You need to distinguish between two sets of models. There is a small set which is what is called the fitting set, and that is the set of models from which you are going to fit one. There is a larger set, called the ambiguity set, and the ambiguity set are all the models that you consider might have generated the data.

You are going to have to keep asking the question: what happens if the real model is something from this big ambiguity set and the model that I am fitting is from a much smaller set? Clearly, in those circumstances, you are not going to end up fitting the right model. You cannot fit the right model because the right model is not in the set that you are fitting.

You might still find that the answer you obtain from fitting the wrong model is good enough. So, your model might not be right but it might be fit for purpose. The idea of robust statistics is to investigate this idea of fitness for purpose; to be able to say, even though I have misspecified the model, the procedure that I am applying is robust.

Let us consider some of what we do at the moment in this context. Quite a lot of what we do is to pick a model, to put this model through a very laborious process of challenge and validation, and then use that model all the way through our business and that model becomes the truth. You have immediately neglected the possibility that any of the other plausible models could have generated the data.

It is a special case where the ambiguity set is very small. We would be making the ambiguity set the same as the fitting set, so you always assume that the set of models you are fitting contains the two models.

What I am proposing is to have an ambiguity set that is a lot bigger than the fitting set. Then you are going to find yourself having to do a lot of rather laborious tests to demonstrate that your fairly simple fitting procedure works across a large number of a big ambiguity set of models.

One of the aspects that comes out there is that you have to be able to define exactly what your fitting procedure is in order to be able to apply it to all these large university models. The *Black Swan* author (Nassim Taleb) calls it model graveyards: all the models that you may have considered but ended up not using. We tend not to discuss model graveyards. When you have fitted the model that you want, all the other models that you did not want have been forgotten.

In order to apply the robustness technique, you have to be able to reapply your method to lots of alternative histories that could have happened, so you need to describe exactly how the graveyard has worked so that you can replicate it.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

How can you say that a fitted model is working even when it is wrong? It comes back to this test, which Dr Tsanakas discussed earlier with a slight difference. We now have greater than or equal to. What is happening is that you cannot expect to hit 99% for every model in the big ambiguity set. What you will find is that if you programme your adjustment to work for one model in the ambiguity set it will not work so well for another. The best that you can hope for is to be at least 99% confident.

You might say that for some models in the ambiguity set, there may be some more favourable models with thin tails, which will be >99% confident. You are perhaps thinking that is ultra prudent? Would you prefer to be 99% on average rather than 99% of the worst case? In general, you cannot formulate 99% on average because you do not have any notion of the relative likelihood of all of these models in the ambiguity set. It is a class of model with no probability relating t to hem.

What you can do, in a very special case of taking a Bayesian approach, is as follows. This is one area where Mr Valeria Bignozzi and Dr Tsanakas have carried out some research. If we do this test for counting the number of exceptions, how well does it work using a Bayesian approach? The answer they came up with, which is comforting, is that if your Bayesian universe contains everything in the fitting set, in the ambiguity set, it works quite well. However, if your Bayesian prior does not contain a big enough set of models, then it goes horribly wrong. So, the Bayesian gives you what you might expect, and if your Bayesian universe contains the right model, you find it. If a Bayesian model does not contain the right model, it is not very robust at all.

So, what are we going to do? Here is an alternative. We are going to play with the model quite carefully so that it does work over a larger ambiguity set. We do not have a theoretical basis for constructing it and so we are going to do it by trial and error. This has not been done very often. Here is one example where it has been done.

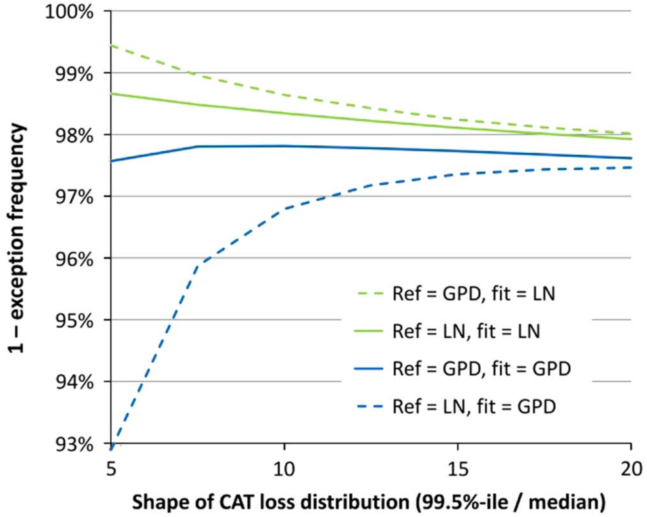


Figure 10. Robust example (Nicholson & Smith, 2013)
Source: Figure 10 is a reproduction of figure 3, from the Actuary magazine http://www.theactuary.com/features/2013/10/gi-prepare-for-the-worst/

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

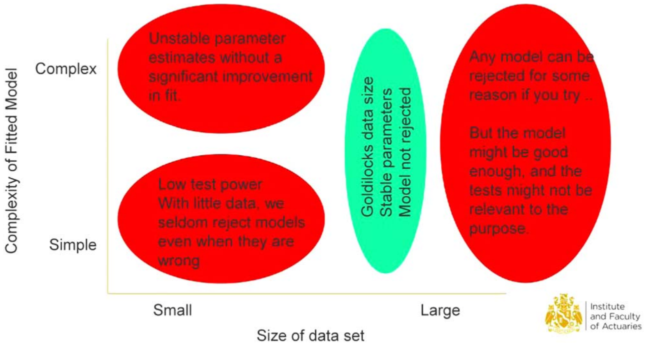
Figure 10 refers to estimating losses for catastrophes. The vertical axis is one minus the exception frequency. How often do you observe losses that are bigger than what you said was your capital requirement?

Let us suppose, for the sake of argument, we were trying to set capital at 98% confidence. The green line shows what happens if you fit a lognormal distribution to your path data and you use that fitted lognormal distribution and you calculate the 99.5 percentile of your fitted lognormal distribution. What we are saying is even if you have used the 99.5 percentile of your fitted distribution, you still obtain exceptions more to the 1 in 200. But it does not become much worse than 1 in 50. Because these green lines are mostly above the 98% line, it is saying that this is a robust framework which gives you at least 98% confidence, regardless of, in this case, the shape of the reference distribution for the horizontal access, different reference distributions for different shapes.

In this case, you have shown by carrying out a large number of model runs with a not particularly impressive ambiguity set (the ambiguity set contains lognormal and generalised Pareto distributions (GPD)) that for ambiguity test and for fitting lognormals, even if the model is misspecified, you still have at least a 98% chance of having enough capital. It also shows, incidentally, that if you fitted GPD, the result completely fails to be robust.

That may not be what you would expect. Maybe somebody told you there was an extreme value theory justification for using GPDs and lognormal is completely arbitrary. By this measure, you are better off using the arbitrary distribution than using the one which appears on a theoretical basis. That is an example of a real live robustness calculation. There are not very many of these out in actuarial literature. I hope that there will be some more in the future. There was quite a lot of effort to produce them, but there is the example of one.

One more theme to cover is validation test materiality. There is a feature of what happens with modelling as you increase the amount of data. In Figure 11, the horizontal axis shows the size of data set and on the vertical axis, I have the complexity of the fitted model. It is very difficult to fit a complex model to a small amount of data because the parameters are unstable. If you have a small data set, you almost always end up fitting a simple model even though, of course, complex models are capable of generating small data sets. You just would not be able to know what was the complex model.



Unstable parameter estimates without a significant improvement in fit.

Low test power
With little data, we seldom reject models even when they are wrong

Goldilocks data size
Stable parameters
Model not rejected

Any model can be rejected for some reason if you try ..

But the model might be good enough, and the tests might not be relevant to the purpose.

Complexity of Fitted Model
Complex
Simple
Small
Large
Size of data set
Institute and Faculty of Actuaries

Figure 11. Expect rejection when testing big data

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

On the right-hand side, if you have a huge amount of data, for example, individual claim amounts for a motor insurance portfolio, you have the reverse problem, which is that no model really fits. There is always some pattern in your data which your model does not capture. However complicated you make your model, you never quite capture everything in the data.

There is a sweet spot in the middle which is where the amount of quantitative data is just right and you can pin down the model and you know what is that model.

I am going to look now at the right-hand side and think about that in a bit more detail. The concept I am going to bring here is of test materiality. So, for large data sets, some tests will fail but it might not matter if the way in which that test is failing is not material for your purpose.

Let us take an example. Suppose I had some distributions of claims from household flooding events and I looked at those claims and they seemed to be roughly a lognormal distribution. Somebody else might come along and say, "I realise all of your claims are integer numbers of pence so they cannot be a lognormal distribution because a lognormal distribution is a continuous distribution". They could do a statistical test and they could conclusively demonstrate that it had not come from a lognormal distribution because every one of those thousands of claims was to an integer number of pence.

So, what would I do? I do not say, "your test is wrong". I say "your test is not relevant to how I'm using that distribution". We need to have a relevant filter for how we do the tests.

We have one example here. Suppose you were fitting a model to some data, normal for some losses. In this example, what has happened is I have a fit that looked reasonably good over most of the distribution. I do a statistical test: the Kolmogorov–Smirnov test. It says the model is fine. When I look out to the tail, I see it is fitting rather badly and my fitted distribution fails to pick out some of those observations.

In that case, our test is leading us not to reject the model but the model that we fitted is not very useful because the errors are located precisely where we are trying to use the model.

You could have the converse happening. You could have a situation where the fit of the model is not very good in the middle but it works fine in the tail. That is quite a common bodge in stochastic modelling. You might fit a normal distribution, obviously not right in the tails, so you bodge up the standard deviation a bit, fit the tails better and now it does not fit in the middle. Does that matter? In a fundamental sense the model is wrong, but then all models are wrong. The model could still be useful.

When you are doing your battery of tests, you should be thinking: is the test that I am doing relevant to the purpose to which I am applying this model, or is the test that I am doing a nice-to-have slightly irrelevant thing?

This starts a challenge about some of the ideas that we have been experiencing in the insurance industry about use tests and the idea that you have a model which you use for lots of different purposes. The whole concept of test relevance is that for the purpose you are doing it the model might be fine but the model might not be fine for some other purpose.

Some of the ideas of ambiguity sets, the quality criteria, robustness and the idea of statistical testing materiality are concepts which might be new to many of us but are also areas which are continuing

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

to evolve. I think that they are areas on which we will all need to construct a more solid foundation so that we have a better basis for the way we perform our day-to-day jobs.

The last section is about social model validity. I think we know that being technically valid is not sufficient for a model to be used. There might be several criteria.

On the right-hand side of Figure 12, you can see textbook criteria of unbiased asymptotically efficiency. On the bottom left, you have commercial criteria of being stable or not requiring excessive capital. On the top of it, you have the idea of social validity.

You might say what is this doing in a scientific discourse? My co-author and I have discussed this at great length and have come to the conclusion that social validity deserves much more of a place at the table.

Let me give some examples of social validity. Some of you might be building models of mass-lapse events. Maybe you have come to the conclusion that a mass-lapse event, the 1 in 200, 99.5 percentile, involves 40% of the population lapsing.

Perhaps the reason that you have come to that conclusion is because that is the number that sits in the Solvency II standard formula, and it is also the number that you know your peers are using.

What we are arguing is you need to accept the legitimacy of social and commercial constraints on modelling decisions. There is quite a lot of work being done on the social acceptability of that data.

If you have social decisions that are masquerading as science, then that inhibits careful thinking, and applying your mind to think what could go wrong with the model requires a degree of intellectual honesty and independence. That becomes compromised the moment that you start defending the numbers you are using because it is what the benchmarking told you, and you start relating it to some irrelevant body of data.

In conclusion, you can never know that a model is right. The best thing you can know is that what you did was a sensible process. If you are thinking about criteria for good models, then you should focus on what comes out at the end of the process, which in the case of solvency regulation is controlling the number of companies that fail (or controlling the magnitude by which they fail).

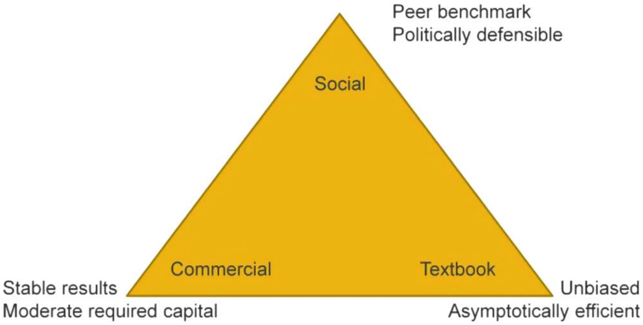


Figure 12. What makes a good capital calculation

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

We think robust statistics are very helpful. That will mean avoiding downward capital bias under a wide range, under an ambiguity test of underlying models.

Your models do not need to be perfect. Some models will fail on large data sets, but a model can still be useful even if it is wrong. You need to think about your statistical testing in the context of its materiality and in the context of the relevance for the purposes that you are using the model.

There will always be social and commercial constraints placed on model choice. So, the purpose of validation should be to understand the impact of those constraints rather than to pretend that they have never been applied.

The Chairman (opening the discussion): That was a fascinating set of points and topics that you raised.

You have spent a lot of time seeking ways of arranging for results to be unbiased. I can see the statistical appeal of such an approach.

However, I note that in modern regulatory thought, there seems to be a tendency to look at the data and, if it does not appear to have enough stresses in it, to instruct firms to replace some of the data in the data set, with new data that is more representative of stressed conditions. In this respect, regulatory approaches seem to be deliberately biasing their output (if the data set is deemed to relate to a time period that doesn't have "enough" stresses within it) and the results towards distributions that show more extreme outcomes.

How would that adjust the prescription that you have proposed?

Dr Tsanakas: This requirement of unbiasedness does not mean to start with that VaR estimates have to be unbiased. This procedure we propose produces a positive bias for VaR estimates.

In relation to the requirement to use biased data, essentially you are saying that you could use biased data as an alternative approach to increasing your VaR requirements. This is quite ad hoc. It is reasonable to say that you should not exclude any data that relate to relevant experience. The idea of considering past histories already will include some adjustment. In some of those past histories the extreme events will be present.

The problem is where, if your original sample does not contain a sufficient number of extremes, then when you go back to generate past data histories, they may also not have a sufficient number of extremes. So, you can only partially address the problem.

But if we are going down the route of saying that we should be deliberately biasing samples, then we have to think carefully about why exactly we want to do this and in which way. We have to be transparent about what sort of purpose we are serving as opposed to just saying "put more extremes in the data so that the estimated capital is bigger".

Mr Smith: I have not talked about the biases that are introduced deliberately by the model when people are cherry picking, perhaps, certain data series which they know will give the answer for which they are looking.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

It may be that this idea to trawl through possible past data sets and find one that is particularly hairy, is a reasonable antidote to what might otherwise be a tendency to trawl through past data sets and pick one which is particularly benign.

I can see it makes logical sense when you have an industry which, by and large, is trying to jig models to minimise capital requirements. It is entirely sensible that regulators might try to dig a bit deeper and discover the model graveyards. If you regard this as a model graveyard discovery exercise, asking people to come up with particularly bad data sets that can be quite a valuable exercise and is certainly not conflicting with what we are recommending.

Dr M. Cocke, F.I.A.: The examples you gave were all around a single risk. I was wondering whether you had given any thought to when you have multiple risks and you are looking at the problem of how to aggregate risks robustly.

Mr Smith: Suppose you have ten different lines of business which all have the same expected degree of profitability. Somebody has taken those lines of business, examined their historical experience (which will be different for each of them) and worked out some optimisation algorithm (which will invariably involve them biasing their portfolio towards the ones that have done historically well). There is something which you can test in this framework.

An example would be the floods that happened in the United States. Everybody was avoiding Florida because their model said that it was very risky. The hurricane then hits a little bit further west and nobody is really prepared.

You certainly can do that model. You have to allow for not only the model calibration process but also the portfolio optimisation process that may have taken place. That means, in your different parallel historical universes, you are always biasing your actual strategy towards those which in the corresponding data set have a very benign history. Unsurprisingly, it is pretty valuable in those situations to have this concept of parallel universes.

To give you another example, somebody might advocate a strategy of picking lottery numbers and picking the same numbers that came up last week. On a backtest, that works absolutely brilliantly. On a Monte Carlo backtest, it is obviously complete rubbish because you have to allow for your optimisation algorithm applying only on the data that you have, and not knowing the whole model.

The framework can cope with that from a theoretical point of view and you can do a few limited experiments. What is clear is modelling the entire decision process of a large financial services business, somehow embedded inside tens of thousands of possible scenarios, is not something that is easy from a computational point of view. This is not something where you can download a bit of software off the internet, press a button and have those results.

At the moment we are at the stage of obtaining some insights as to how it would work. There are certainly some steps further to go to industrialise it to the scale of an insurance company, or a pension fund, with a very large number of different risks.

Mr B. Bergman, F.I.A.: We talk about the social convention of using a 40% mass-lapse stress assumption, but perhaps the biggest social convention of all is the 99.5th percentile capital requirement threshold which is now enshrined in Solvency II legislation.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

Just playing devil's advocate, when the 99.5th percentile capital requirement was first thought up at the outset of Solvency II, taking into account the sophistication of our modelling and the level of knowledge at the time, it was presumably deemed to produce an acceptable capital requirement. The number would most likely have assumed a stable past and not a random past.

So, if we are going to start adjusting our models to allow for a random past and to allow for bias, then arguably the 99.5th has to be reduced, say to 99th in order to hold overall capital requirement levels at levels originally intended based on the modelling techniques back then! Of course, you do get rid of the cross-subsidies between firms if everybody does accurate modelling, rather than using broader brush approximations.

In all this correction for bias, we have been focussing on the capital requirements. We have been also talking about the 1-year risk measure. There is another component of "capital" under Solvency II, namely, the risk margin. I think that most people now would probably acknowledge that this is just "a number". Whether the risk margin will do what it says on the tin under situations of stress, that is, being able to transfer your business to third party at arm's length, when nobody is pricing based on the stipulated 6% cost of capital yardstick, is debateable. Are we not simply polishing the decks on the *Titanic* by trying to correct bias in the capital requirement calculation aspect of Solvency II, rather than asking ourselves, whether the whole Solvency II capital regime (taking into account both the risk margin and the capital requirement, i.e. the SCR) works in the round?

The Chairman: On the risk margin point, I remember making exactly the same sort of comment. It turned out that the person to whom I was making the comment had thought up the 6% cost of capital figure. He obviously proceeded to explain to me that there were good reasons justifying the selection of this figure. So, sometimes there is greater rationale than initially meets the eye with this sort of computation.

Mr J. G. Spain, F.I.A.: In terms of the social and commercial constraints within model choice, surely there is something that comes afterwards when the board say "That is fine. You want us to have a fund of £100 billion to meet the risks. We do not think that is possible".

It is not clear to me that some of that material should be in the model choice modelling before recommendations are made, as opposed to afterwards.

Mr N. D. Morjaria, F.I.A.: I work in a bank and so I am a little voice as an insurer in a bank. The banking industry has very much moved away from the 1-in-200 VaR-style approach to a stress and scenario testing-based approach where management can really understand what are the emerging risks, running those through their models and seeing what that looks like.

I wonder whether, in the insurance industry, if something really bad does ever happen, then we may be in a similar position where for all of these complicated models that have been developed, the regulators say "They were too complicated. Nobody understood them and we should go back to something simple".

The Chairman: The current regulatory requirements in the insurance sphere do require you to do both a VaR-style computation and stress and scenario testing. Maybe we are talking about the degree of weight that we apply to those two techniques.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

Mr D. Murray, F.I.A.: One of the messages that we might take away is that if we do not have enough data, then we need to go further into the tails than we thought to get a 99.5% result. We might need to go to 99.7%.

I guess that depends on whether we have enough data or not. I am interested whether Dr Tsanakas and Mr Smith have a view as to whether, as an industry, we do have enough data to do what we need to do appropriately. It seems to me that we have loads of mortality data, so we ought to be able to get very good tail estimates in that area. We also have quite a lot of past data around things like equity risk and other market risks. We probably have quite a lot of data as an industry on persistency risk. Perhaps in something like credit risk you could argue that we do not.

I wonder whether you have a view on where the other pinch points are in the areas of Dr Tsanakas's conclusion that we need to go more into the tail because of lack of data.

Mr Smith: In the context of any particular organisation, you also have to look at the materiality of the risk to that organisation. In the work that I have done, for financial market data, there is good equity data for many countries going back to the 1970s. Interest rate data for some countries goes back a lot longer. For the UK, it goes back more or less to the Commonwealth or Oliver Cromwell era.

The question then comes down to relevance rather than the data points. There are some areas where the data is much more limited. Persistency is one of those: the FCA produces a persistency survey. You do not have to go back very far before you start seeing negative lapse rates and all sorts of other things which it appears nobody can explain.

There is definitely some question about the data and the relevance and the quality of that data. If you are looking at mass population behaviours, that is definitely a pinch point.

Another consequence of the way that we have looked at this is there is not an abrupt cliff where you have or have not enough data. There is a smooth graduation where, if you have a small data set, you are going to have to think very hard or make a substantial adjustment for the sampling area in that data. As you move to larger and larger data sets, that adjustment becomes smaller. There may be a stage where you say, the adjustment is so small compared to what I am trying to measure, that it is de minimis and I am not going to bother.

There are quite a lot of the data sets, including some of the ones you have mentioned, where there is apparently lots of data. The model and parameter uncertainty is still a material concern. That is selling the case for longevity projections where you might have a huge amount of data, but, as always, there are things in the future which plainly will not replicate that. For example, the number of people who have given up smoking in the 20th century cannot be repeated in the 21st century because there are not that many smokers.

Even when you have large data sets where there are big social trends that can have an effect. The fact that you have a large data set does not necessarily protect you from model uncertainty.

Dr Tsanakas: Mr Smith said that with large data sets there are structural changes which make them less useful. The effect of model and parameter uncertainty is also related to model complexity. When you have large data sets, you also need complex models to describe them successfully.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

The other aspect that you raised was whether we advise that you should go further into the tail because of parameter uncertainty. I would put it differently. If you want to satisfy a particular criterion of quality which says the expected number of violations should be fixed at a nominal level, then you need to go further in the tail in a given way. But you must first agree on our criterion.

What I would welcome is a discussion about what we think is a relevant criterion for managing parameter and model uncertainty as opposed to what is the right add-on to use.

Dr M. C. Modisett, F.I.A.: I think the criteria you used to go further into the tail is a rather conservative one. We are trying to make a very small chance that your model is understating capital. This is not the criterion that I think the regulators use. They talk about a 1-in-200 estimate of capital. They are not trying to produce a conservative estimate for capital but a best estimate of capital. That means it can have an error on either side. Your criterion is a statistically based one. It is a very natural one if you have a statistical background. It is probably more conservative than the regulators are thinking.

The Chairman: That is a good point that also links to the point made earlier about whether 99.5% is the right number in the first place.

Mr A. K. Howe, F.I.A.: The paper takes, understandably, quite a macro approach and clearly focuses on VaR. I wondered whether you had done much thinking about things close to the mean and how this might vary according to the various parameters.

Mortality might be a good example where mortality can vary by lots of factors, some of which are rating factors, some are risk factors, which we cannot or do not take into account and just linking it to the point made by someone else about the mass of past data. The mix of those factors may be changing quite rapidly, one example being the annuity market where lots of people may take cash out very rapidly. Your techniques work for the mean.

Mr Smith: Mr Bergman mentioned the 99.5th percentile and is it the right standard. A lot of us are focused at the moment on regulatory requirements because they are new and we are trying to keep up with them.

Most of the literature that we have drawn on is not about insurance regulation. It is about estimating future outcomes, both extreme and moderate. The concept of a predictive limit does not have in its description a hard coding that α is 0.995. You could say α equals 0.6 and have a 60% level of confidence if you wanted. The maths and the concepts still carry across. We have hooked a few things on the 99.5 percentile, but that is something a lot of us are forced to confront in our day jobs. I would not want you to think that this set of work was only about complying with regulations.

There are various aspects of the regulation that nobody would necessarily try to justify as being logical: it is the outcome of a process of negotiation, not a process of scientific enquiry. It would be naive to expect it to come out with something that was theoretically perfect, and indeed, it has not.

That is not the problem we are trying to fix. We are trying to do a decent job of answering the question posed.

Dr D. J. P. Hare, F.I.A.: The presentation made me feel that my knowledge of some statistics is quite inadequate for doing some of the roles that modern actuaries have in life offices. I can use these results, but being able to produce them myself is something that would require a certain amount of CPD.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

I take comfort in the fact that it is not a one-pillar-solvency regime within which we operate: it is a three-pillar regime. I go back to the Sharma report from 2002 which highlights that firms get into trouble not because of capital per se but because of the decisions that management make.

Part of what Solvency II is doing, particularly through the ORSA, is highlighting to those making decisions about firms, exactly what risks they are running and the consequences of doing so.

I do not think that disagrees with anything that has been said. What I want to highlight is the challenge upon us as risk professionals to take the statistical theory from other areas, as well as our knowledge of the financial risks, and communicate that to boards in a way that we can support them make decisions. Additionally, to give policyholders security and shareholders the profits that they are looking for from firms in as wide a range of scenarios as is possible and to manage both of those sometimes competing objectives.

I would encourage many people to challenge ourselves whether, as actuaries advising firms and maybe as actuaries taking the decisions in firms, we know enough about statistics to be doing the job. And, as things move on, is our skillset moving with it to continue to give the quality actuarial advice that we want?

The Chairman: Towards the end of the presentation, Mr Smith noted that some of the concepts included in the presentation might be relatively familiar, whilst others might be newer.

Would anybody like to comment on some of the strengths and weaknesses of applying these newer concepts in the areas about which we have been talking? Are we here talking about something really radical? Can we take some of these newer concepts and express them in a form that is easier for the layman, or the lay actuary, at least?

Dr Modisett: In an ambiguity set, you have to decide at what universe of distributions you are going to look.

But formalising it into considering what you did not consider is maybe a little bit of a different angle on the problem. The entire idea of materiality, every assumption we make in Solvency II, we have to write down whether it is material or not, so it is writing down things that we have been doing for at least a few years.

These are things that at least I have done and my colleagues have done in our work. Maybe this is a slightly more formalised take on them, but I do not think that these are new animals.

The Chairman: So, perhaps, we are merely introducing new jargon or new terminology?

Mr Smith: I put it slightly differently. We do not claim to have invented any of these techniques. In that sense they are definitely not new. There are publications going back for decades looking at these questions, although not always applied to actuarial work. It is a point well made.

It is also fair to say a lot of actuaries in their day jobs are doing quite sensible things to the best of their ability. The difficulty is that when you push on the justification for what has been done, the justification does not always make a lot of sense.

For example, there is a lot of concern about measuring correlations with the possibility that in certain economic scenarios, all those correlations that you have measured historically turn out to be irrelevant and suddenly everything becomes correlated with each other. That is a legitimate concern.

It is one that our community has struggled to articulate in statistical terms that would make sense from a purely statistical perspective. So we have a disconnect. We have an idea of something that bugs us and we have a few bodges that we might do to make us feel better.

Then you have statistical principles which we are trying to articulate in terms of principles which do not really connect in the middle. So, you end up with a slight wild goose chase that is tail correlation. That is a term you will not find in any statistical textbook. It is a term that you will find in pronouncements from regulators, and it is a term you will find in consulting marketing literature in which they claim to be able to address, on your behalf, the concerns of the regulators. But it is a completely empty, vacuous concept that does not really exist, for which we are struggling to find the language to articulate.

You can articulate that in the context of ambiguity tests. There is a class of different dependent structures. We know that we do not have enough data to pinpoint which one applies. All we can do is consider each of those as reference models in an ambiguity test, look at the consequences and understand where are our vulnerabilities. That would be an example of where you might see common practice becoming a bit more consistent, a bit better codified, by rooting it in statistical principles. At the moment it is floating in the ether and is not sufficiently grounded to be able to apply really rigorous statistics.

The Chairman: Does anybody else have further insights on these concepts?

Mr P. O. J. Kelliher, F.I.A.: The idea of simulating past histories could prove very useful in things like demonstrating standard formula appropriateness, because you have in the standard formula a reference distribution.

Mr Bergman: Have the authors applied any of their techniques and methodologies to validate the standard formula model to see whether it produces the 99.5 percentile when measured against the data upon which it was calibrated?

Mr Smith: We have done that and it failed. If you take random regenerated past history and look at the methodology, for example, by which the interest rate is calibrated and you generate lots of alternative paths, some of those alternative paths will be very smooth and some of those alternative paths will be very volatile. Your future outcomes are also generated from the same model but not necessarily the one you fitted to the one that generated those paths.

There are a significant number of cases where you underestimate the stress because you have had a luckily smooth past. At that point, far more than your 1-in-200 trickle through in the future.

There are other cases where you overestimate your stress with the consequence in terms of exceptions violations is much smaller. As a rough order of magnitude, you would be doing quite well if you got exceptions of 1 in 50. The methodology applied to a standard formula did not really consider the parameter uncertainty contained therein. If you measure it by a metric, as we are advocating, it does

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press

cut to the parameter uncertainty (and it is not just us who are advocating that but it is also part of the requirements), and it manifestly fails.

The Chairman: It does look as if many of us will need to think further about how to validate approaches versus the standard formula. This perhaps ties in with earlier comments about how the regulators might not have wanted to be quite as conservative as some might have expected. Or do you feel that there are other factors involved?

Mr Smith: What personally bothers me is the apparent deceit of proclaiming models that have a much higher confidence than we know they really have. I do not have a strong view as to whether insurers should be required to hold twice as much capital, half as much capital or something in between that. That is a political call, part of the social contract, if you like, between the insurance industry and policyholders.

All I think I find uncomfortable is being expected to say that something is a 1-in-200 event and you know that you have ignored a whole load of things which even the regulations say you are not allowed to ignore. It is a total muddle as you do have a standard formula and an option to use that, in which case apparently, you do not have to make those assertions.

It ought to worry everybody because it is a threat to the reputation of people who work in those modelling areas.

Mr B. E. Johnson, F.I.A.: You raised the idea of model simplicity as being a very important factor and something that we would look for. You referred to robustness quite a lot and determining whether a model is sufficiently robust.

Where there is more than one model that might satisfy your robustness requirements, how would you then handle the trade-off between robustness and simplicity? I guess in some cases your most simple model might happen to be your most robust model, but where there is a trade-off, what kind of optimisation techniques would you use? Is it just a judgement call or do you have techniques in mind?

Mr Smith: What we have described is some tests which could be applied. As you have noticed, you may be able to construct more than one method which passes those tests.

My take on it is, from the point of view of complying with what you have been asked to do, any of those alternative methods should tick the box. And from the point of view of regulatory approval or somebody validating that you are fitting a 1 in 200, it should be a matter of indifference which of those you pick. The different methods could have several different properties. Some of them might produce capital results which fluctuated more than others over time. Some might produce capital results that explode to very large numbers in certain stress situations and there would be a cyclicality problem with those. Some might produce capital requirements that are higher than others on average. You might want to look at the average capital requirements over the cycle. That will depend on a firm's individual preferences. You are not going to have a one size fits all. Those are trade-offs rather than model simplicity. The trade-offs are if you are going to hit this 99% or 99.5% target, you can measure the volatility of the requirements and measure the average level of the requirements. That is roughly where the trade-offs reside.

The Chairman: I would like to thank the presenters and all the people who have contributed to the discussion.

https://doi.org/10.1017/S1357321715000094 Published online by Cambridge University Press