[SQUEAKING]

[RUSTLING]

[CLICKING]

PROFESSOR: All right. Well, today's guest lecturer is James Shepherd with Quantile--

JAMES SHEPHERD: Quantile, yeah.

PROFESSOR: --which is a wholly owned subsidiary of LSEG. And your background in quant finance is quite long. And so I'll let you share some of the highlights.

JAMES SHEPHERD: OK.

PROFESSOR: Thank you very much, James.

JAMES SHEPHERD: Thank you very much. Do I need that.

PROFESSOR: No.

JAMES SHEPHERD: OK. Thank you very much. Thank you for having me. This supposed to finish about-- so a quarter to 4:00. Yeah. I'd like to leave some time for questions.

PROFESSOR: Until-- yeah. 10 to 4:00.

JAMES SHEPHERD: OK. Thank you very much for having me. Yes, I'm James. Yes, I work for Quantile Technologies. That's a fintech company, founded in about 2016, to do counterparty risk optimization, which is the subject of today's talk.

A couple of years ago, we got bought out by the London Stock Exchange Group, which is why my email now says LSEG. Prior to that, I was at Morgan Stanley for quite a long time, which is where I know Vasiliy and Jake from, which is how I've ended up here.

So we're going to talk today, we're going to work our way up to talking about optimizing a thing called initial margin, which is a measure of counterparty risk. So we're going to start off talking about different types of risk. Sorry.

PROFESSOR: [INAUDIBLE]

JAMES SHEPHERD: I've just emailed them to Peter. Sorry about that.

PROFESSOR: [INAUDIBLE]

JAMES SHEPHERD: The first view is just kind of like warmup stuff, so you're not missing too much. So the first part of the talk, we're going to talk about are different types of risk that you have with derivative trading. And we're going to talk about expected shortfall and value at risk.

Just out of interest? Do people know what these terms are, or is it worth my explaining them a little bit?

AUDIENCE: Explain.

JAMES SHEPHERD: OK, I will do that. Then we'll talk about how that relates to initial margin and counterparty risk. And then we'll work our way up to how we optimize this initial margin within a network of financial institutions. And we'll talk a lot about the challenges that you get in the real world.

Because this is a maths talk, I feel morally obliged to chuck some equations in the thing from time to time. But largely, I'll skip over that and talk a little bit about some of the intuition, some of the toy models. And I've put a bunch of references. When you get the thing, there's a bunch of references at the end, which has got some of the proper maths and proofs and more detailed stuff behind it, if you're interested in that.

So in the shape of derivatives, when we talk about the different kinds of risks that can arise when you're trading derivatives, market risk is probably the most common one. So that's the value of the derivative goes up and down as the market changes.

Credit risk, again, is reasonably well understood. It's the rise when you've got exposure to a debtor defaulting. Operational risk is largely when somebody screws up. It's probably actually the hardest one to do any mathematical modeling on operational risk because it's quite hard to model people just messing up, which is really what that is.

Liquidity risk is when you've got a transaction, which is well, as the name suggests, illiquid, which can happen when you've got a very large position. So if you're trying to sell a massive position in some particular stock, you start selling that thing. And then because you start selling it, that moves the market. And so you can't sell the whole thing at the position that the market was when you started.

So liquidity risk happens it's a kind of obscure instrument. And we'll come back to it in this. It happens again if you've got a very, very large position. And what we're mostly going to talk about today is counterparty risk, which is the exposure that you've got when you trade derivatives to a counterparty defaulting.

I think maybe a question you might have here is, what's the difference between counterparty risk and credit risk? So the example you can have to differentiate this is, if I do a trade with Goldman Sachs, let's say, a credit derivative swap that has some JPMorgan corporate bond as the underlying, then I have credit risk with respect to JPMorgan. Because if JPMorgan defaults on its corporate bond, then the value of my derivative will change.

And I have counterparty risk to Goldman Sachs because, depending on what happens to the value of that derivative, Goldman Sachs may or may not pay me that amount of money. That's the difference between those two. Make sense?

So once you've got a type of risk, once you've got some risk, generally speaking, you want to mitigate it, i.e., reduce either the probability of it happening, or reduce the impact of that risk happening, not just for counterparty risk, for all kinds of risk. Hopefully, reasonably obvious that people want to try and make the impact of once you've identified the risk, make the impact of that smaller.

So there's a bunch of different ways that you can do that. We're largely going day-- to talk about the hedging, which means book some trades to offset-- book some new trades, which will offset the risk that you've just discovered that you've got. And we'll talk about collateralization, which is really paying margin payments from the person you've got-- the person you're exposed to, you would collect a margin payment from them, which is kind of like an insurance payment against them defaulting. And that's mostly what we're going to talk about today.

There are other things you can do. And there's a thing called central clearing counterparties, which I thought nobody would have heard of. But apparently, someone's-- Andrew Gunstensen has already mentioned this a little bit. So I hope I might skip over this slide because they are important in the trading of derivatives.

So I have a quick slide on CCPs. So normally, you trade derivatives, and they're bilateral trades. So JPMorgan trades with Goldman Sachs and so on. So we've got here, you've got four parties. Say, A trades with B. A trades with C.

It's got a position of 100 versus B. 125 in the opposite direction versus C. And what a central clearing counterparty does, as the name kind of suggests, is that it steps into the middle of all these trades. And so instead of A facing B and A facing C, A just faces the CCP. And so does B, and so does C, and so does D.

Just by doing that you haven't really achieved much. You've just changed your counterparty risk from being versus B and C to versus the CCP. But once the thing is in the CCP, you've got both the 125 and the 100. And you can net those two positions together. And so you can net that down and say, well, now, instead of having 100 plus 125, I've only got 25 as my risk. And the same story for all the others.

So in the case of B, when we had exactly equal and offsetting positions, B's ended up with no counterparty risk at all. So B's in a very, very happy situation. And everyone else is slightly happier than they were before.

There are other advantages of CCPs, which I'm not really going to talk about too much. But the thing to know is that some derivatives are mandated to clear at CCP. So interest rate swaps-- pretty much if you have a standard vanilla interest rate swap, you have to do this. You don't get a choice.

But more exotic trades like, say, an interest rate swap or something-- an interest rate swaption, they are not eligible to be cleared at a CCP. So some are. Some are not. So what you end up with is a portfolio with some bilateral trades versus a bunch of different counterparties and some trades at the CCP. So almost everybody will end up with a mixture of stuff.

So now we're going to talk about quantitative risk measures. So what you want from a risk measure, or any general risk measure, it should quantify somewhat the size of the total risk or the size of the total impact. How much money could you possibly lose? And it should also quantify the probability of losing that amount. So there's two aspects to it-- the size and the probability.

So in 1999, there's a paper by Artzner-- there's a reference at the back-- came up with these properties of a risk function, which is nice properties, which are known as a coherent risk function if it satisfies those five properties. Mostly, they say things like the function behaves kind of normally.

We'll come back-- so the important one is subadditivity. Anyone heard of subadditivity as a thing? So basically, what that says is if I have two portfolios X1 and X2, and I've got some risk on X1 and some risk on X2, and then I bash those two portfolios together, and now I've got a combined portfolio, the risk on my combined portfolio, according to my risk measure, ought to be less than or equal to the sum of the risk on the two portfolios I had in the first place, which seems kind of sensible.

If I put some stuff here and some stuff here and put it together, it seems unlikely that I've manufactured some risk out of doing this. But now conversely, it seems highly plausible that there might be some diversification of putting these two things together. And so my overall risk might well be smaller. It seems a very plausible and reasonable property to ask for from a risk measure.

So we're going to talk about two risk measures in particular, which are probably the two most common ones for looking at a whole portfolio. One is Value at Risk, which I probably will keep referring to as VaR, and the other is expected shortfall.

So value at risk is this amount here. It says, if I've got-- it says, what is the maximum-- for a given confidence level, what is the maximum amount that I can say that I will lose more than that amount. So for example. can I be 99% certain that I won't lose more than $10 million over, say, the next 10 days?

So I have a confidence level is called beta. There's a time horizon, which we'll talk about in a sec. And the VaR is how much money can I be 99% or beta percent confident that I won't lose more than this amount?

The expected shortfall says, conditional that I've exceeded that amount of loss, so given I'm in the 1% tail if beta is 99%, or the 5% tail if beta is 95%, if I'm in that tail, what is my expected loss, given that I've exceeded VaR? So that's what expected shortfall is.

So hopefully, it's indicated on here that this is the percentage. This dark area here is 1 minus beta. So this is, say, 1% or minus 5%. The VaR is the biggest amount before I get into the tail. And the expected shortfall is the average value of that tail. Clearly, the expected shortfall should be bigger than the VaR. Hopefully, that's obvious for the same confidence level.

So VaR was kind of popularized by JPMorgan in about the 1990s, something like that. Expected shortfall came around about 2000. And both of them are heavily used by regulators and lots and lots of counterparty risk measures today. VaR is probably the most well-known one and most common. It's got a couple of problems.

The first problem is if you have these kind of weird fat tail probabilities. So you've got two probability distributions, red and blue. The blue has this weird fat tail bit over here. And we'll just say, well, the area under the blue bit and the red bit is the same. So both these distributions have the same VaR because the area up to here-- assume there's a little bit of blue here. The area up to here is the same as the area under that. So both these distributions have the same VaR.

But clearly, the blue one has a much bigger probability of doing some fairly horrific loss. And also, hopefully clearly, the blue one would have a much bigger expected shortfall than the red one because the average of the blue one is about here. The average of the red one is about here. So VaR doesn't distinguish between those two different cases.

And you might say, well, this is ridiculous because nobody's going to have a-- in real life that kind of distribution can't possibly exist. But that is not completely ridiculous distribution. It's not totally uncommon trading policy to sell a whole bunch out of the money options.

And if you sell a bunch out of the money options, what most of the time going to happen is you're going to take a small profit from the sale of those options. And most of the time, they will expire still out of the money. They won't get exercised, and you're going to take a profit.

But some of the time the market will move enough. And if it moves enough, they'll get exercised, and you're going to lose a lot of money on those options. So that-- even though I did contrive that for the purposes of this lecture, that kind of distribution is not as ridiculous as you might think. So that's the first problem.

The second related problem is that value at risk is not subadditive. So this, again, a little bit of a contrived example to show that if you take, say, the 95% VaR. I take two portfolios A and B, with those kind of probability of happening. A and B by themselves, the VaR, is zero because there's a 96% chance I'm going to lose nothing. So 95% the VaR is going to be zero.

The expected shortfall of those things is given. I'm in that last 5% tail. 1% of the 5, I'm going to lose nothing and 4% of the 5, I'm going to lose 100. So I get to 80. And if you do the maths and just add the two things up, and you do the A plus B, you can discover that the VaR of the combined portfolio is 100, and the expected shortfall is 103.

So that is a kind of example that shows that VaR is not subadditive. Expected short-- obviously, this is not proving that expected shortfall is subadditive. It is, in this case. It's quite hard-ish to prove that expected shortfall is subadditive. It is. And I put one of the things at the end is seven different proofs that it is subadditive. So you can pick your favorite one from the thing at the end.

I also put an example at the end-- so even though this one is a very contrived example. So in real life, most things are kind of normal or normal-like. And in that case, VaR is subadditive. So in real-life situations, this is not as much of a problem as you might think.

And there's a paper at the end. It talks also a lot about that and says, actually, lots of people criticize VaR because they come up with these kind of crazy examples, but it's not too bad. For our purposes, given we want to go on to optimizing stuff, it is critically bad. Because subadditivity-- we'll come back to this again in a sec-- is basically the same thing as convexity.

So not being subadditive means VaR is not really a convex function, in general. And convex functions are way, way, way easier to optimize than nonconvex functions. So even though it's not a problem if you just want to know VaR. It is a massive problem if you want to optimize one of these things. And we'll come back to this again in a little bit.

I just wanted to quickly point out that there's lots and lots of other risk measures you can possibly choose here. VaR and expected shortfall just happened to be the two common ones. You can characterize a risk measure by how much weight it puts on a particular percentile.

So VaR is putting all of its weight on the beta percentile. Expected shortfall puts zero on everything less than beta and averages out everything above beta. But you could say, well, both of those-- you can make some criticisms of them.

You could say, well, I should put a little bit of weight on something that's a bit less than the beta percentage, and I should weight the-- the really, really extreme tail losses, I should weight them more than the just beyond the tail losses.

So you can imagine having these exponential spectral risk functions that have these kind of functions, which do exactly those things, that they put an awful lot of weight on the really bad losses at the 99% and 99.9% things. And it scales off going down. It's just to point out there's a few different choices you can make here.

So both VaR and expected shortfall, they've basically got two parameters. They've got the confidence level, and they've got the time horizon. So we'll talk a bit about the time horizon.

So that means I want to make sure that I'm going to lose less than this amount of money over one day, five days, 10 days. That's the time horizon. How long have I got to avoid losing this amount of money?

If you were to make the assumption that the change in the portfolio-- so this delta P is a normal distribution, then VaR and expected shortfall would have these two formats here. So they would also be normal. And if you assume that people make these kind of assumptions in the thing, that the mean change on a daily basis is zero, which is not a crazy kind of assumption, then both VaR and expected shortfall would be proportional to the standard deviation with these things. They would be normal.

And if you did that, and you assumed that all the daily changes were independently-- they were all kind of independent identical normal distributions so you could add them all up. Then you would discover that the T-day VaR and the T-day expected shortfall is proportional to the square root of the one-day VaR, which is the normal approximation that people make with those things.

In reality, it seems that the losses-- the changes that you'd have on a typical portfolio or a typical market data are not independent. If you lose a lot of money one day, it's a reasonable chance you might lose some money on the-- or the market might go down on the second day. So there's a thing called autocorrelation, which is basically the correlation of a time series with itself shifted by one day.

And so you can-- which is what this thing here. And if you assume that that autocorrelation, rho, is some number, then it would modify the T-day VaR for two days by this formula. So I put in this table.

So the top row is assuming that there's no autocorrelation. So that's 1 root 2 root 5, root 10 and so on is the top row. And then as you increase the autocorrelation, the T-day VaR, as a function of the one-day VaR, gets bigger and bigger.

So I did some-- I'll show you some market data in a sec. So I took a set of observations from the EURO-STOXX stock index. And that suggests that the rho, the autocorrelation, is about 0.1. I did some other indices, and they were all between 0.05 and 1.15. So 0.1 is kind of reasonable.

And generally, people take a 10-day VaR thing as a normal thing. So you'd be looking at about this number here, this 3.46. It's about-- that means to say that the standard approximation of using the square root of-- saying the square root of the one-day VAR-- sorry-- the square root of T times the one-day VAR is probably going to underestimate the actual T-day VaR by about 10% because it's 3.46 divided by 3.16. But nevertheless, almost everybody just says that the T-day VaR is the one-day VaR times root T.

So this is where I got the data from. So I picked four stock indexes-- Ibex, CAC, DAX, and EURO-STOXX. That's, if you don't know, Spanish, French, German, and a kind of pan-European stock index.

So from 2006 to 2024, you can see there's a couple of periods of stress in here. One in 2008 time frame, around about here, which was the global financial crash, and one in about 2020, which was COVID. And some degree of calmness in between those two periods. Clearly, there's some correlation between these things that's there.

On a practical note, there's a couple of things worth pointing out. I picked these stock indices because they're all European, which means I don't need to worry about FX rates when I'm playing with these things. But even though they're European, they're from different countries. And the different countries have different holidays. And so you have to decide what I'm going to do-- what you're going to do when one of these indexes is on holiday and not trading.

What I did here was that if all of them were on holiday, I'm ignoring the day completely. And if one of them was not on holiday, but the rest were, I just fill forward. And so you'd see a daily change of zero on those.

That might not necessarily be the most sensible thing to do. It doesn't really make that much difference. But the reason I mention it is because if you work in this kind of industry for that, you end up spending a surprising and disproportionate amount of time dealing with both FX and holiday calendars. You probably don't expect to be doing this. But holiday calendars-- you spend a lot of your time trying to work out what are you going to do with these kind of things?

So there's a couple of different ways you can estimate VaR. One is historical simulation, and one is doing model building. We'll have a quick look at both of them.

So the historical simulation, basically is we're going to go back and look at the time series. We'll look at the value of each of these indices, the CAC, DAX, IBEX, and EURO-STOXX on each day. And we'll come up with a portfolio. So I just made up some portfolios. Say I'm going to have 3,000 units of CAC, 4,000 of DAX, and so on and so on. And we'll call that the portfolio weight.

And so now, I can figure out that on the day-- the change in value on day I is going to be the relative change in the underlying market index. For example, the second row minus the first row for the first day and so on, multiplied by the weight for that index. Then just sum it up over all the indices. And that will give me the change on the I-th day. So then I can take the distribution of those things, and I will get something that looks like that. Hopefully, that's big enough.

So one thing to add here is that when you're doing that, when you're looking at these historical simulations, you have to decide over what period, over what historical period, am I going to look at? And the main thing to notice here is that depending on what historical period you look at, it makes a massive difference to both the VaR and the expected shortfall.

So what normally happens, if you normally talk about VaR, you talk about the most recent, say, two years. These are all two-year periods, by the way. So the top left, 2022 to the end of 2023. Given my series went up to the end of 2023, this top left box would be what you would normally put in as VaR, the most recent 500 days, basically, assuming there's 250 days in a year, which is more or less true.

The ones on the right, they're from the two stress periods-- the global financial crisis and COVID. And as you might imagine, if you compute VaR based off those two periods, you get a much, much bigger number than you do of either of the calm periods, the most recent one, 2022, or a period in between the global financial crisis and COVID, 2015 to 2016 with that.

So that makes, by far and away, the biggest difference, which period you choose. And because of that, regulators have recently introduced-- well, not that recently-- introduced the concept of a thing called stress VaR.

So in addition to taking the most recent two years, five years, 10 years, whatever you decide to choose, you also pick a fixed period known to be stressful. So people will pick either COVID, or they will pick the global financial crisis. And then that's generally known as S-VaR, or stressed VaR because there's such a big variation in these terms that are there.

And then just to break it down in terms of what you actually did here, what I actually did was I took all those changes-- so there was 500 changes in each period, more or less. I ordered them from the biggest change to the-- from the biggest loss to the biggest gain, these things here. And then the 5th worst one-- that's the VaR, for the 99% VaR, because that's 1%. And the 12th worst one, that's the 97.5.

It's not that complicated. You just order the thing. You take the 5th worst one, 12th worst one, that's what VaR is. And expected shortfall, for the 99%, you take the average of the worst four. And for the 97.5, you take the average of the worst 11.

There is some debate about that. Some people claim you should take the average of the worst five, so including the 320, and the average of the worst 12. I think standard market convention is to take the four. Doesn't matter that much. But that's all that's actually happening here to do these. All OK so far?

To do the model building approach, it's kind of similar. We just make the assumptions that everything is normally distributed. The daily changes are normally distributed with mean zero and a sigma I. We assume that there's some kind of correlation between all the different indices. And

So you say, well, the sigma for the portfolio is going to be this term here. So there's a variance-covariance matrix between the-- right-- that you get. And using the assumption that VaR and expected shortfall are proportional to the sigma again, you get these two expressions.

So all you need to do for the model-building approach is figure out what the variance-covariance matrix is between these different elements, which I computed here. And I come up with a VaR, and I get these two numbers-- 225 and 257.

The main thing to note about that is that it's significantly, significantly less than the historical VaR. So the 99th-- this is computed for the 2022-2023 period. So the historical VaR-- those equivalent numbers were 320 and 377, basically showing that these normal distributions, it's clearly not a normal distribution. There's a fatter tail in there than you would expect.

But you can say, well-- you can work out, is this an acceptable VaR model? Yes or no. So what you can say is given I've got something where I'm going to take 500 trials, and I expect to see-- and I expect the VaR-- the actual loss to have exceeded the VaR 1% of the time, that's the definition of VaR.

And what is the probability that I'm going to see 12 or more-- so there's 12 violations, by the way. So if I take this 225 and say, how many times did it actually-- according to that period, how many times was it actually-- the loss was bigger than 225, or there were 12 scenarios bigger than 225? So what is the probability that I would see 12 or more days out of a sample of 500 that exceeded the VaR?

So that comes from the binomial distribution, which is that formula, and you get 5.2%. And the general rule of thumb is that if the probability is 5% or more of getting that degree of-- that number of violations of the VaR, you say that the model is OK. And if the probability is less than 5%, you'd probably reject the model and say the VaR is not a good model. So even though the VAR here is significantly lower than the historical simulation, you would probably accept that as an OK kind of model.

And the only other point I wanted to make on that is that what we've basically done there is a back-testing of VaR. So we basically came up with a model, and we said, how good is that model? How many times, based on some historical simulation, was the VaR exceeded? So it's really, really easy to do back-testing of VAR.

The single main reason why people still use VaR over expected shortfall is because doing this for expected shortfall is quite tricky. It's not even clear what it means to do back-testing of expected shortfall. Because, I mean, the expected shortfall is a function of the probability distribution. It's the mean of the tail. You can't really compare that with one sample that I've drawn from the distribution. It's like comparing apples with oranges.

There's a whole bunch of debate around whether back-testing expected shortfall is flat-out impossible, difficult, or not. And there's a bunch of papers at the end on this as well. But it's definitely a lot trickier than doing it for VaR. And that is the main reason, people still use VaR, despite its shortcomings around the subadditivity and the handling of the tails.

So I'm going to skip over some stuff. This is some stuff around regulation. And I'll just quickly mention this slide because-- so one of the problems with expected shortfall is it's kind of defined in terms of VaR, because it's the mean loss that you would get, given you've exceeded VaR in the first place.

What you can do is there's a-- this function here-- this is in the paper by Rockafellar and Uryasev, which is in the back. It says, if you take this function F, which is this expression here, and you try-- and then you minimize that over alpha. You consider this to be a function of X, which is the portfolio, and alpha is just a number, and you minimize this over alpha, then the minimal value of that function is the expected shortfall.

And so you can find the expected shortfall by taking that F and minimizing it. And that is basically if you replace the integral with a sum, that's basically a linear problem. And so it's relatively easy to find the expected shortfall.

Moreover, if you minimize it with respect to alpha and X, now what you've done is you've found the portfolio with the minimal expected shortfall, i.e., you have minimized the expected shortfall, which is basically what we're trying to do in this whole thing. So you just have to consider this function here. You can minimize that over X and alpha, which is basically a linear problem. And that essentially is how you minimize expected shortfall.

No such equivalent thing exists for value at risk. So even though expected shortfall is trickier to back test, it's easier to minimize, which is what I care about more I'll skip over it. So that's expected shortfall and VaR.

So now we're going to start talking about counterparty risk and margin. So if you have in the derivative world-- so if I have A has got two trades, one with B, one with C. Derivatives typically swaps, means I'm going to exchange a series of cash flows between the two parties. And typically, those cash flows will be exchanged, say, over a three-month period or six-month period.

So A, in this case, is receiving some cash flows. It's paying the series Si to B and receiving Ti from B. A is exposed to the fact that before the next cash flow arrives from B, B might default and not pay the cash flow. And the same story for C.

So A has counterparty risk to both B and C because when we get to the next cash flow date in, say, three months time or six months time or a year's time, they might not pay that cash flow that they're due. So that's the fundamental reason why they have counterparty risk for derivatives.

In order to mitigate against the fact that there might not pay that cash flow, what actually happens is that they pay-- is that all these counterparties exchange on a daily basis the exchange variation margin, which means that-- which is equal to the value of these derivatives. So as the value goes up and down of these things, they'll be paying variation margin to one another. And that mitigates against the cost of the cash flow not getting paid.

But if they're paying variation margin from one to another, and then one party or other goes bust and defaults, then you've still got to pay the variation margin to the other person. And as soon as one party goes bust, what A would try and do-- so if C goes bust here, A would try and replace that trade.

It would take it some number of days to replace that trade with something else or unwind that trade, during which time the value of that trade might go up. And because the value of that trade is going up, it's having to pay an increased amount of variation margin to B. And so therefore, it incurs a loss.

And so it incurs a loss because during the period between when C defaults and when it manages to replace that trade, the value of the trade will change or go up, basically. And that is the loss. So in order to mitigate against that risk, A will collect initial margin. That's what initial margin is. That's essentially a covering you for the risk in between when a party defaults and when you manage to unwind or replace that position.

That slide just says what I just said. And then I have another slide, which says the same thing again in a picture. So you have the variation margin is getting paid up until here between all these things. You get some default will happen here. We don't know what's going to happen. The risk can go up or down. And some loss could potentially be incurred at this point.

This should hopefully sound to you a little bit like what we just discussed with VaR and expected shortfall. And so initial margin-- in order-- how much initial margin should I collect? Well, basically it's going to be a VaR based on how long you think it's going to take you to unwind that position, which is called the margin period of risk.

There's two kinds of initial margin, depending on whether your trade is a cleared trade or a bilateral trade. The cleared margin is set up by the clearing house, the CCP. And that's always been the case for years and years and years.

Up until 2016, people were not paying initial margin on bilateral trades. That changed when ISDA came up with their standard initial margin, which is the form-- I've put the formula down here, but it's kind of similar-- it's basically VaR is what they put here. And the only difference compared to what we just discussed is that because it's handling nonlinear portfolios, there's a gamma term in there, which I didn't talk about before.

And when you work it all the way through, you make exactly the same assumptions, that everything is normal. Mean is zero. You end up with the exact same term we saw before. Or you also-- SIMM assumes that it's a 10-day margin period of risk. And it's a 99% confidence interval.

So the 99% confidence interval, that's the inverse normal of 99% That's the 10-day margin period of risk. That's the exact same term we looked at in the VaR before. And then you get two additional terms that come from the convexity, which is kind of similar. The main point is that initial margin is basically some variant of VaR, or potentially expected shortfall.

In real life, there's lots and lots more simplifying assumptions that people make around all these things to try and simplify stuff. And all these kind of correlations between all the different parameters-- this covers the entire world, all different asset classes, by the way. And so they get kind of recalibrated on a yearly basis.

If you Google SIMM COVID now, you'll probably see quite a lot of stuff in there because SIMM 2.7 is about to come out in the beginning of December. And that'll be the first one that's calibrated without any of the COVID or global financial crisis data in there. And what's going to happen, what people expect to happen, is that the initial margin calculated by this SIMM is going to massively reduce by somewhere in the region of 5% to 20%.

And what that will do is release something in the region of $40 billion of margin back out into the wild. Because today, people are paying initial margin, all these different things with these high levels of-- using these high levels of VaR. When they recalibrate in the first week of December, everything will come down. Everybody will release all the initial margin. So there's quite a lot of stuff around that in the news at the moment.

That just talks about how the structure-- you roll everything up into one number across all the different portfolios. And I'll skip over that. So that's what margin is.

So now we come on to trying to optimize the initial margin. I put margin here. I mean initial margin. So the situation that we're thinking about is we've got multiple different financial participants-- A, B, C, D, E. Think of them as banks.

And we've got-- I've put one CCP. There could, in fact, be multiple CCPs in the middle. And all these participants have got a bunch of bilateral trades versus each other, and they've got a bunch of trades versus the CCP. That's the situation that we've got. And they're all posting margin to each other in this situation.

And what we are trying to do by optimizing stuff is come up with a bunch of hedge trades, which is these X, H variables between all the different participants in order to try and minimize the margin, i.e., the VaR, in the entire system. That's the goal of what we're trying to get to.

And you might say, well, why are there no trades versus the CCP? Well, you're ignoring the cleared margin. That's just a practical issue. You can't book trades between the CCP and a bank. All you can do is book bilateral trade between A and B, A and E, all these different things. And depending on which trades you choose, some of them will be cleared or not cleared.

If you pick a trade, where the trade type is like an interest rate swap, then that will be mandated to clear. And so even though it's between A and B, it'll get cleared. And it will affect the cleared margin and not the bilateral margin That's why there's no actual trades booked between the CCP and any of these people, Because that's what we're trying to do.

So that's the problem we're trying to solve, which looks like a lot of maths, but in fact, is not that complicated. So what we're trying to do, is we're trying to minimize summed over all of the parties, so all of the banks-- A, B, C, D, E-- the margin of all those people.

So the margin is made up of the SIMM, the bilateral margin summed over all the counterparties, plus the cleared margin summed over all the CCPs that are there. So that just says the sum of all the margin in the system is what I'm trying to minimize. And we say, well, where SIMM and the initial margin, there's some functions of risk, as we've discussed. We know what the function is. It's Var-like, let's say.

So we know that SIMM and-- oops, sorry-- initial margin are some functions of the risk. And we know that the risk position that's there is equal to the initial risk, which is the zero, plus the sum over all the hedges that we're going to book.

So we're going to propose a bunch of trades. So we sum over all the hedges that we're going to book, the size of the hedge which is the X, which is the thing we're trying to find, multiplied by the risk of that particular hedge. So that's how we know what the new risk is.

And we've got a couple of constraints that we need to put in the system. This is the most important one. This is what we call the symmetry constraint, even though it looks like it looks like an asymmetry constraint. What this says is that if P sells something to Q, Q had better buy that same thing from P. Exactly.

And this was-- when we first set this up, this was, in reality, the thing that people were most worried that we would mess up. Because the biggest problem that could have happened here is that we told-- we set up a thing. We tell a Goldman Sachs, go and buy 100 million of stuff from JPMorgan. And then we tell JPMorgan, go and sell 200 million of stuff to Goldman Sachs.

So Goldman Sachs sees their thing, and says buy 100 million. They think, that looks good to us. We'll agree to this. JPMorgan sees their thing, which says, sell 200 million of stuff to Goldman Sachs. And JPMorgan goes, that looks good. We'll agree to this. Goldman Sachs doesn't see what we told JPMorgan, and JPMorgan doesn't see what we told Goldman Sachs. And they only discover that something's gone horrifically wrong when they try and do that trade.

So the reason that doesn't happen is because we've got this constraint in the system that says that, whatever Goldman Sachs buys from JPMorgan, JPMorgan had better sell that exact same thing to Goldman Sachs. That is the most important constraint in the whole system.

The second constraint, the big one that we've got in there, is this one, which says-- this is cash flow flatness. This says that for any party, let's say, Goldman Sachs, if I sell something to one party, I'd better buy that exact same thing from some combination of other people. That's what that constraint says.

We don't need that to be in the system. It's a kind of safety valve that's in there. The advantage of having this in here is it says, if I'm definitely sure that everything is cash flow flat across the system, nobody's going to make any money or lose any money out of the whole system. Because I'm going to sell something. I'm going to sell something to there. I'm going to buy it from somewhere else. So I'm just-- all I'm doing is shuffling risk around the system.

You could relax that, and you could say, well, some parties could make some money. Some could lose some money. Some could change some risk. But it makes everything a lot trickier. So we basically enforce this around the system.

And then the last constraint is a risk constraint which comes from the parties themselves. And what this says is that Goldman Sachs or some party might say, well, I don't really want to go-- I might not want to go long a particular risk factor versus some particular counterparty for whatever reason.

So the first two are system-wide constraints that we impose on the thing. And the last one is a constraint which basically says, don't change the risk too much, which is what parties say. So all of that maths, that's all it's saying.

So if we look at-- it's kind of-- if we look at a very, very simple example. If I take a three-party system-- A, B, C-- so this might be a bit small-- and you put it in there and say, well, how could I minimize the margin around that system? And the way you'd minimize it, you've got AC's got to trade with 10, AB's got to trade with 2, and BC has got a trade with 9.

Well, what you do is you move the median trade around the whole triangle. If you move one median trade around-- one trade around the whole triangle, I'm guaranteed to satisfy by construction those first two constraints because I'm only moving the exact same thing around the whole system. And if I'm moving, just say in this case, 9. I'm going to-- A is going to buy 9, sell 9. B is going to buy 9, sell 9. C is going to buy 9, sell 9. So I'm going to satisfy the second constraint as well.

I've ignored the third constraint for now. And the reason I picked that is-- those particular examples with 2, 9, and 10, is that actually if you fiddle around with the maths a bit, the thing you should move is the median trade given any triangle. And you can take any network, and you can always break a network into a bunch of triangles. And for every triangle, you just move the median around. That's what you should do.

The problems are this won't satisfy all the different-- if you have additional constraints, it causes problems. And obviously, some edges of this triangle would also be edges of other triangles. And so you end up and you can-- it won't generally solve the problem. But it's a good way of thinking about it. We always think about triangles and how you would move risk around in a particular triangle. So that's that.

In reality, we just use a big numerical solver around this. What we're in the business here is we're solving minimization problems or operational research type problems. They're of the form minimize F subject to some other stuff.

And as I mentioned earlier, solving complex problems is way, way, way easier than solving general, nonlinear problems. And convexity is basically the same thing as subadditivity. That's the definition of convexity. And that should remind you of the definition we saw earlier of subadditivity.

So we saw earlier minimizing expected shortfall. That's basically a piecewise linear problem. All these things we're talking about, SIMM is nearly convex. There are different add-ons that you can have, which are usually convex. And some of the maths that we get into is, how do you come up with approximations of these things which genuinely are convex.

We tried using proper nonlinear solvers on these things. Way easier just to approximate the function you want and use a convex solver on it. Because the state of the technology for solving convex optimization problems is really, really, really good. It would be crazy to write your own version of a convex optimization solver because there are loads of open source and commercial available-- things available. So don't do it. Just plug it into a system, and it works.

The thing to note is that how difficult it is to solve the problem is really, really, really sensitive on the specifics of the problem. And that doesn't really mean the size of the problem. The size is not really too much of an issue. It's all kinds of the structure of the problem, how sparse is the problem. And the thing these are the things that matter in the real world

Also, it's a good idea to use modeling languages. So there are lots of modeling languages available. Gurobi is the one we use. I think you may-- people might have used it. I think you can get it for free if you're an academic institution. In case anybody wants to spend 10 seconds trying to solve that in their head, this is an integer problem, trying to minimize 5 x1 plus 8 x2 subject to those two constraints, you can probably do that in your head. Hang on. So we did that. Sorry.

So we did it on Gurobi. Using the 24 physical cores probably is overkill for that problem, but it comes up in 10 milliseconds with the answer of 31. Hopefully, if you did it, that's what you tried, just to show how easy it is to solve some of these problems.

In real life, a typical problem that we would have might have 25 different parties in the network. Might be, say, for an FX delta 30-odd different currencies that we've got in there. We pick different kinds of hedge instruments, so we try and pick simple ones.

The reason for picking things like nondeliverable FX forwards is that each individual hedge type there has exposure to exactly one underlying, namely the currency. If I pick a dollar euro nondeliverable, that's got exposure to the dollar-euro FX rate and nothing but the dollar-euro FX rate in this case, which turns out to be useful.

And if you do that, you end up with a system which is something like, say, 20,000 variables, 40,000 constraints. Something like that is the real size of the problem. And really, when you're solving convex optimization problems, solving convex optimization problems is easy because the convex problem looks like a bowl. And so they have one minimum.

And so you can convert the optimization problem into a root-finding problem, which is find where the root of the derivative is-- where the derivative is equal to zero. So that's root-finding. People generally use something like Newton-Raphson for root-finding.

When you use Newton-Raphson, you've got to compute a Hessian matrix. And then you need to invert that Hessian matrix. That's the slow part of the entire system is that you're repeatedly inverting Hessian matrices.

So the matrix might be 40,000-ish. Matrix inversion is basically an n cubed kind of operation. So you end up with this unfeasibly large number of operations to solve. In general, most computers will not be able to factorize a 40,000 matrix for a randomly chosen 40,000 by 40,000 matrix.

However, the problems that we have here are very much not randomly chosen matrices. They have an awful lot of structure involved, and they're massively, massively sparse. And most solvers, the reason why they're very mature, is they take advantage of all this matrix structure that's there. And so they can do way, way, way better than you'd expect from this.

So in particular, for FX delta, the matrix looks a little bit like this right. So you've got a bit stuff at the top of the matrix, which is the thing I talked around around the symmetry constraints and the flatness constraints. But the vast, vast majority of the constraints are those ones at the bottom of that picture, which are the risk constraints specified by the individual parties. And they are guaranteed to be block diagonal.

Because each-- if I have a particular hedge, say, the hedge AB, that's only going to affect the risk between party A and party B at worst case. And then the hedges between A and C, they're only going to affect the risk between parties A and C. So I'm definitely going to get a block diagonal matrix down here. And everything else is going to be zero. So the vast majority is going to be zero.

And if I pick these NDF trades as well, and those SDF trades only have exposure to one risk factor themselves, I'm just going to get a straight diagonal matrix here. And obviously, inverting a straight diagonal matrix is trivially easy.

So I've got a bit at the top, which makes it a bit more complicated. But it's a little bit at the top, plus basically a massive, big diagonal matrix. That's called a Dantzig-Wolfe structure, and there are techniques known to factorize and invert those kind of things.

So then I like this slide. I think this is the one slide I've used all three years. The problems that you get into in real life are not solving the big minimization problem. You just literally write down the math that you want, more or less plug that into the solver.

But in real life, you get situations where you've got nearly parallel risks. You have two trades which are nearly the same, but not quite exactly the same, or can't be represented exactly the same on a computer.

So this example here, if I really wanted to solve this first equation here, obviously everyone-- hopefully, you can see that essentially, there's lots of solutions to that right with this. But you can't represent a third properly on a computer.

So if you actually tried to represent that on a computer, what you might end up with, for example, instead of putting a third, a third X1 plus a third Y1 as your second equation, you might end up with this thing here 0.333 x1-- sorry-- 0.3333 x1 plus 0.333 Y1. And that has this solution. 0, 1.

But you make a tiny, tiny change to that thing-- so all I did was I flipped over the solution here. Instead of having three 3's at the end, having four 3's at the end. And instead of getting a solution 0, 1, I've now got a solution 1, 0.

And if I make another small change to those things, it says instead of-- I'm going to get 0.3. Then that has a single solution of minus 110, 111. That's the solution to that problem.

Now, it's highly plausible I might also say, well, I want all these X's to be positive. That might be a thing in there. And so entirely due to numerical noise, it will say if I want X to be positive, and the only solution to this problem is minus 110, 111, any system is going to go, well, that's infeasible and so it will refuse to solve it.

And so the problems that we get into real life is trying to understand these things and coming up with techniques to remove these kind of situations from the system, rather than trying to actually solve convex optimization problems in their general form. That's there.

This is the same thing as before, except now for the FX delta, the full FX delta problem, we're up to 48 cores now. And you can see here, it really was 114,000-- it was 100,000 by 100,000 matrix is what we ended up trying to solve. It's worth noting this presolve thing here removed close to half of it.

What presolve is, is sometimes you get all these different constraints, and you have them in there. And you end up with a constraint that says X1 is greater than 1, and another one that says X1 is greater than 2. So you might as well chuck out the first constraint and say, well, if X1 is greater than 1 and it's greater than 2, chuck out the first one, and just keep the X1 is greater than 2 constraint. And that happens a lot. And so you can massively reduce the size.

And so we get down to a thing like this. But you can see even this problem, you can solve it in 11 seconds and with very, very few iterations, which is basically a testament to the amazing power of Newton-Raphson quadratic convergence.

So what happens when you try and solve it? This is how you have the initial risk, and this is with the optimized risk on the right-hand side. As you might expect, it tries to form-- it tries to basically get it, but I don't have some positive numbers and some negative numbers. It tries to nestle the risk together. So what you can see here.

You try this on a 20-party system, again just removing all the constraints. Again, it can remove almost all the margin between the different systems. And you end up with either-- for any given row, either pink stuff or green stuff. It's either positive or negative. But you don't have pink and green stuff in the same row or in the same column.

And what that says is it managed to get rid of all the situations where I was up. I was up here and down there. I just want to be up or down. So I've netted all the risk together. That's what the system was basically trying to do. You try it on a real system with all the constraints and stuff like that, it's less convincing what it was trying to do because the constraints hold this stuff in check here with their--

Just out of interest, this kind of diagonal line down the front, that's essentially the margin versus itself. So you obviously cannot pay margin to yourself, which is why you get a kind of blank line down the diagonal. And the square stuff in the middle-- I originally thought that was some bug. But what that is, is trading between interaffiliates of the same entity.

So somebody like JPMorgan or Bank of America will have lots of entities, some of which don't pay margin to each other. What I probably did here was I ordered all the entities alphabetically. So this is probably, I'm guessing, given I did that, that's probably all the interaffiliated of JPMorgan is the one in the middle since that's-- or maybe it's Morgan Stanley. It'll be some company which has got a letter roughly in the middle of the alphabet with many entities not trading margin versus each other. That's that. That's another example doing interest rate stuff.

So what we've discussed so far is solving the optimization problems. It has to be feasible, which means it satisfies all the constraints. And it has to be optimal, which means it finds a good margin saving.

But on top of that, it's interesting to try and make them nice and try and make them fair. So what a nice thing would be is that you don't really want to have horrible looking notionals. You don't want to have to have a trade with some complicated notional in there. You'd like to just trade 100 million of this or 200 million of that.

The biggest amount of headache we got into ever was trading with these NDF FX trades, where we would trade it with a $1 notional with some 200,000 million point 74, and then something else with the Euros, like 0.53 in the Euros. And then we compute the FX rate, which would be the one divided by the other. And the FX rate would come out to be a 15 decimal point number, and then everything would be fine.

And then we'd give it to the banks, and they would complain. And they would say, well, our system only handles six decimal places or eight decimal places in the trading system. And when I do that, and I put six decimal places in the FX rate, it changes my euro notional. It gets a little bit off.

And so we ended up having to put notionals in that were round numbers of dollars. And that came out as an exact round numbers of euros using a four decimal place FX rate, or a six decimal point FX rate.

And that turns out to be quite tricky because you need to round all these numbers. And you can't just take a solution and just round it to the nearest ten. So if I have a party, which they've got a-- and I'm selling eight units of something to one person, and I've offset that by buying four units from the second person and four units from a third person, everything is flat.

But if I round everything to the nearest 10, then I've rounded the 8 to 10. And I round both 4's down to zero. So now I'm no longer flat. So you can't do this rounding as a post-processing step afterwards because it breaks the flatness in the whole system. So you have to do it as part of the optimization. If you do it as part of the optimization, that makes it into a mixed integer problem. And that makes it considerably, considerably harder.

People also want-- by nice, they want a small number of trades. They don't want to have to trade thousands and thousands of trades in order to get some of this margin saving. They'd like to have a small number of trades. We talk about it as notional efficiency.

And there are other advantages of CCPs, I mentioned earlier, in addition to just the netting. And so there's some advantage of saying, well, I want to put-- I'd like to put more trades into the CCP. And so you might do some-- that might be a nice thing to have for other reasons not directly related to initial margin.

So you can add all these niceness features into the system. The problem with adding the nicest, things like round numbers, small numbers, like trade counts, these kind of things, they all make basically continuous problem into a mixed integer problem, which is hard to solve.

And then fairness is-- well, the first question is, what does it mean? What does what does fairness mean in an optimization problem? So you have this example. Hopefully, you can see this. This is a four-- back to the triangle system-- four-party system. It's kind of symmetric between A and D.

There are three possible things you can do, all of which have the same overall saving here. Essentially, you can ignore D, and solve the A, B, C problem, in which case D gets no saving. You could ignore A and get the same saving. Or you could kind of split the saving between A and D.

So they are the three different scenarios that you can get right. So I put them down in the table at the bottom. So scenario one is the ignoring D. This is the saving that you get. A gets saving of 4. A, B, C, will get a saving of 4. D gets nothing.

So this is-- you can argue that's not fair because D gets nothing there. You can argue scenario 2 is not fair because A gets nothing. You can argue that of these scenarios, scenario 3 is the fairest, just intuitively. It still might not be the fairest because, well, why did A get 2 and D get 2, but B and C got their full 4 allocation? So the concept of fairness is not a well-defined one.

One thing you can look at is-- whoops-- have a two-step system. So you, first of all, figure out what is the best possible saving that any one party can get. So the best possible saving that all of these people can get is always 4. And so you do a whole bunch of optimizations where you ignore everybody else and just try and optimize for A. Then just optimize for B. Then optimize for C. Then optimize for D. Now I get the best possible thing that each one can be, subject to all of the constraints still being there.

When I know that, now I'm going to say what I'm actually going to solve, is how do I minimize the square of the difference from my actual solution to those things? So I've got a least squares problem that's in there, which is this problem here. And that's the thing. So I'm going to do one more optimization at the end when I solve that problem. And I'm going to say that's the answer that I've got.

And you can see if I did that, in this case, clearly with the squared scenario, 3 is going to definitely come out as being the best solution because B and C will be zero in those things. And then for scenario 3, I'll get 4 plus 4. Whereas, in scenario 1 and scenario 2, I will get a 16 from the square.

So if I were to do that on this case, this would be an algorithm that would deterministically give me scenario 3 as the fairest. But there are other ways of defining fairness. And I don't think it's a well-established kind of a thing yet. I think it's the single most interesting problem that we've got in the optimization space.

And then I'll skip over. This is just some pictures of what we've kind of really got in terms of saving. The more interesting one is really this, which is showing the kind of network effect. So you can expect, obviously, that as I put more and more banks or parties into the system, the savings should go up. That's kind of a reasonable thing.

But more than that, the saving divided by the initial, that also goes up. So you get not just the total saving in the system goes up, but the total-- the efficiency increases, the more people you add into the system. I wouldn't put too much faith into whether it's really linear or whether it's a kind of not linear thing in there.

But there was clearly a drift upwards in these things, which basically says that if you want to do this, the most effective thing you can do is get big networks of people. And big networks of people are beneficial to everybody in the system. That's basically the story of that picture that's there.

I think that is the end. Yeah. So that's just a quick summary. We talked about value at risk and expected shortfall. Talked about the centrally cleared and bilateral. We talked about counterparty risk, what variation margin is initial margin. What we're really trying to do is optimize this initial margin.

That basically comes down to doing constrained convex optimization. The difficulty in constrained optimization is really all the numerical issues that you've got. And actually, the interesting problems are not-- it's so much solving the feasibility and optimality bit, but how do you add the niceness and the fairness on top of that? Hopefully, that is the end of your slides. There you go. Thank you very much.

[APPLAUSE]

That makes sense to anybody, or-- [LAUGHS] anyone have any questions?

PROFESSOR: Go ahead.

AUDIENCE: Yeah, I had a question about the SIMM that you mentioned.

JAMES SHEPHERD: Yeah.

AUDIENCE: It's going to-- like, this new thing that's coming out and--

JAMES SHEPHERD: Yeah.

AUDIENCE: Is that what you said?

JAMES SHEPHERD: Yeah.

AUDIENCE: It's going to be the first one where it's not using data from a stressful time, like COVID, for example?

JAMES SHEPHERD: Yeah, yeah, yeah.

AUDIENCE: I was wondering, what goes into that data cut-off, and then what implications do you think that will have for just overall behavioral [INAUDIBLE], or just affecting the economy [INAUDIBLE].

JAMES SHEPHERD: So I don't think it will necessarily-- well, the effect on the economy is going to be-- well, potentially, there is going to be a lot of margin released back into the system. So essentially, margin is money that cannot be used for trading. It's essentially locked away in segregated accounts.

So the release of all that margin into the system means there's more margin available for trading. That's kind of the effect that it would have on the system. In terms of the kind of choice of parameters, I think effectively, Monsieur, that really it's a roughly a two-year lookback, similar to what I was doing in the toy examples that I had earlier.

So this is-- even though it's looking at other different things, the stress ended roughly '20, '21, '22-ish, something like that. So basically, it's going to be looking at a region that's, say, from 2021 or 2022 beyond. And that's the reason that all these big shocks that have happened in 2020, they're too far in the past to be counted in the correlation these days.

Now, of course, there might be some more. This exact same thing was about to happen in 2018, 2019, just before that happened. And they would also have been about to say, well, let's just smooth it all out. And then what did happen was then all these shocks came in, and people decided that what SIMM was doing was massively underestimating the VaR.

And the fact that they were only recalibrating on a yearly basis was kind of problematic because they had to wait a whole year, and in fact, a year and a bit because of the timing they had in December and the way they came out before they could start, including these shocks that happened in 2020. So actually, as a result of that, they moved from an annual recalibration process to a semi-annual process, so that they could bring the effect of these shocks in more quickly.

AUDIENCE: Is "they" the regulators?

JAMES SHEPHERD: ISDA. It's the International Swaps Derivatives Association. So there's a panel. They chair it. It's ISDA's model. They're the ones who ultimately put that thing there. But there's a lot of collaboration with various banks and other interested parties. Yeah?

AUDIENCE: I've heard recently that there's the raising-- like increasing using a lever of data rather than just [INAUDIBLE] the pros and cons of that [INAUDIBLE].

JAMES SHEPHERD: It's a different beta.

AUDIENCE: OK.

JAMES SHEPHERD: Yeah. So that's a beta in terms of-- well, this beta here is just a confidence interval. So let's just say-- the beta is the, I'm 99% confident I won't lose more than this. Yeah.

AUDIENCE: Do you mind going back to the slide where you talk about back testing-- or I guess the limitations on back testing expected shortfall.

JAMES SHEPHERD: Yeah. Hang on. That one.

AUDIENCE: Yeah, I guess it was-- I guess in my-- what are some ways that you can almost cut corners, like make some assumptions that allow you to calculate expected shortfall. What do you have to assume for it to be possible to do?

JAMES SHEPHERD: To be possible to back test?

AUDIENCE: [INAUDIBLE], yeah.

JAMES SHEPHERD: Well, you have to make some assumptions around the form of the tail. So if you were to say, well, the tail is kind of flat, or you have some structure in that thing, you could start to say, does it fit in there?

What normally happens in real life is that regulators have moved from VaR from saying, well, your actual initial margin and capital should be based-- that used to be VaR. Then it was stress-VaR, and now it's moved to expected shortfall. But they still require you to back test VaR

So what you generally do is you build a model, which contains-- which will predict both VaR and expected shortfall. You compute the number using the expected shortfall number. And then you back test the VaR because that's a little easier.

And then you say, well I've built the model consistently. And all I did was apply a different multiplier for the expected shortfall versus the VaR. So if the VaR back tests properly, then I'm prepared to accept the expected shortfall.

AUDIENCE: Right. If you assume a different shape in the tails, then theoretically that's not necessarily a valid statistical test. Is that [INAUDIBLE]?

JAMES SHEPHERD: Well, it is. But when you're back testing-- you're back testing, and you say, well, what happened on the 23rd of December 2015? Well, there was no distribution. That's just whatever happened on that date. So there's no distribution there.

So you can't say-- it's difficult to make an assumption about whether or not that was kind of predicted or not predicted. Because it just-- it's a single draw from the distribution as opposed to the shape of the distribution. It doesn't tell you anything. You can make any assumption you want about the tail. But because I've taken one number, it doesn't really show me anything. Does that make sense?

AUDIENCE: Yeah.

JAMES SHEPHERD: Yeah?

AUDIENCE: I guess building on from that, so you mentioned that you can back past on both the shortfall and bar on bar instead of a shortfall? And that's kind of like this-- and then helps you like predict more accurately. Are these models accurate enough to account for the difference [INAUDIBLE] of somewhere between bar and shortfall? Are they precise enough to you?

JAMES SHEPHERD: Are they-- well, it depends how you build up the model. So essentially, my model here-- the model wasn't really a VaR model or an expected shortfall model. The model that I proposed here was that everything-- that the daily changes are normally distributed. And in this case, that they're independent-- all the daily changes are independent. That was essentially my model.

Having made that assumption, I get a VaR number and an expected shortfall number, and I can back test the VaR. In real life, people make more sophisticated choices. They don't assume that it's normally distributed. You can have other distributions. You don't assume that all the changes are independent of each other, but the model is not really a VaR model necessarily, or an expected shortfall model. The model is, what did I assume about the distribution?

Having made those assumptions, I get a VAR number and I get an expected shortfall number. Does that make sense?

AUDIENCE: On this example, when you were explaining it, you said that-- let's see. I guess the actual shortfall was 5.2%, was it, through the mod-- or through the historical simulation.

JAMES SHEPHERD: Yeah. So what it said was that this came up-- so the VaR came up as 2.25, based on the model. If I go back to this table-- so this is the leftmost column here. In reality, we can see that there were 12 actual days when I exceeded that number. There should have been five. That was the intention.

So you say, but if I was trying to produce something that had a 1%-- I'm supposed to exceed VaR 1% of the time. That's the definition of VaR. So if I'm trying to come up with a model where I expect something to happen 1% of the time, or I expect the VaR to be exceeded 1% of the time, then I do 500 draws of that thing, and I say, well, how many times was it actually exceeded? And then the probability of that being 12 is 5%.

AUDIENCE: So you are being conservative. Like this--

JAMES SHEPHERD: Yes.

AUDIENCE: This VaR is conservative. Therefore, it's OK. Even though it's wrong in a sense, or at least it's not calibrated. But it's conservative.

JAMES SHEPHERD: It's not that conservative. A conservative VaR would be bigger than the historical VaR. So the actual thing is the historical VaR was 300 or something. Conservative would be a bigger number. But it's not overly aggressive, let's say. I think that's probably the way to put it.

PROFESSOR: Any other questions? Anyway, thanks very much, again.

JAMES SHEPHERD: OK. Thank you.

[APPLAUSE]