2005 Valuation Actuary Symposium*

Orlando, Fla.
September 22-23, 2005

Session 24 TS
Drawing Appropriate Statistical Inferences

Instructor: Douglas L. Robbins

Summary: Stochastic testing has become an integral part of the valuation actuary's work, but can the actuary be sure of making the correct inferences when looking at the results of a large number of equally likely scenarios? In this session, the instructor provides simple tools and guidelines to better ensure that correct conclusions are drawn. This material is presented without the use of complex mathematical formulas. Topics include the appropriateness of the assumption of "normality," including discussion of the distribution of a conditional tail expectation, drawing inferences about population means and tail percentiles, and uses and pitfalls of linear regression analyses. At the conclusion of this session, participants have an understanding of the conclusions to be drawn from the sometimes-arcane language of stochastic modeling.

MR. DOUGLAS L. ROBBINS: I'm a Pacific Life actuary in the annuities and mutual funds division. Before I started doing that, which was last November, I spent about 10 years with Tillinghast-Towers Perrin, and it was during that time that I came up with the idea of doing a session on drawing appropriate statistical inferences. As a consultant, it was never a big seller. It was always something I did because I thought it would be useful to the other members of my profession to have a session on what I think is one of the intuitively most valuable and useful concepts you can have in your hip pocket.

In terms of how I designed this session, I always thought of it as something that I designed without opening a textbook. I had an advanced degree in statistics before I started work, so I did have some knowledge in the field, but at the same time, I have not gone deep into any formulas in developing the presentation, and I never wanted to go deep into any formulas. I wanted to give you concepts that would be easy for you to remember and easy to apply when you're back at your desk.

I've divided the presentation into five subsections. The first two are "How Normal is

Normal?" and "What If It's Not?" There are a couple of sections on statistical applications that you might use beyond the realm of normal distribution theory, and the last section is on some miscellaneous issues.

I want to open up by discussing how the normal distribution was probably taught to most of us when we were in school, especially in "cookbook" statistics. There were some anecdotal arguments that a lot of things in life that you've seen with a probabilistic structure have roughly a bell shape. I'm sure you heard about rainfall, highway accidents and things like that. These are things that, over time, have roughly a bell-shaped probability. Sometimes you have low, sometimes you have high and most of the time you have in the middle.

You then took a quick look at the probability density function, which you were taught, but then you were also taught that you couldn't do much with it because you couldn't integrate it or something like that. Then you were given an introduction to a neat table in the back of your book, from which you could quickly draw stochastic inferences. That's the way that I was taught and where it ended up. You did your sigmas and your mus, and then you said, "That's 33.3 percent likely" or something like that.

But consider the following. I'm going to give you some distributions and ask you how comfortable you would feel trying to make an estimate of the 99th percentile of a distribution using a normal approximation with some mu and sigma. How about the number of accidents on I-5—it's in Los Angeles, but think of it as any typical freeway—in a morning rush hour, if the mean is 2? How about the result of a single, randomly rolled, fair die with one through six spots on it? How about the expected cost in basis points of a variable annuity guaranteed living benefit (VAGLB)? How many are familiar with the last example? Not too many, so maybe that's a bad example. I'll spend a little more time talking about what that is when we get to the small case study on it.

I say that maybe normal isn't so normal after all, because here's how you might do. You may have thought that the number of accidents on I-5 with a mean of 2 must be distributed roughly Poisson, because we've been taught that if you've got a discrete distribution with a low mean that could be zero or that could be a large number, but most of the time it's roughly some average number, that the distribution could be Poisson. You might use the normal theory, using a mean of 2 and a variance of 2, which is the Poisson mean and variance. If you use the normal tables, you'd estimate a bit over 5. If you use the true Poisson distribution, the 99th percentile would be just over 6. So you would undershoot about 20 percent of your guess, which is not very good.

We know, just from doing the math, that the mean of a single fair die with one to six spots is 3.5. We know that the standard deviation is 1.7. You can work it out with a stubby pencil. Using the normal approximation with mu of 3.5 and sigma of 1.7, we'd shoot for the 99th percentile by adding 2.33 sigmas to the mu. We would

get 7.5, when the die has only a one through a six, so this is a significant overshot.

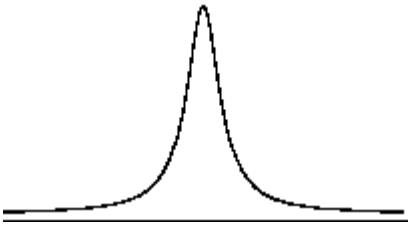
Maybe normal is not so normal after all. On the other hand, depending on the mu and sigma that we're talking about, the VAGLB estimate might just be doable. We'll see why later on in the session.

When is a normal approximation a bad idea? What conditions make it okay? Let's first talk about when it's a bad idea. Let's talk about the things that you would not want to try to approximate using normal theory if you know mu and sigma. It's almost always a bad idea when the data are somewhat scarce or bipolar in nature. Bipolar would mean that you've got two separate modes, if you will, about which the distribution tends to hover. It might be one or it might be the other, depending on conditions. A normal distribution doesn't even look like that, so it's not going to work. If the data are scarce, you may not have enough data to where things are even going to converge.

Also, we're making tail inferences much of the time, especially if the distribution is fat-tailed. In some distributions with which you work, especially if you've done any kind of health or casualty work, you might know that the distribution is fat-tailed. It might be something where a cost could start at zero but could be out to infinity, and you know that the shape is fat-tailed as opposed to humpback-shaped like a normal distribution.

But also, there are some distributions that could appear bell-shaped, but the tails in both directions can still be significantly different from a normal distribution. Chart 1 is a graph of the Cauchy distribution.

Chart 1
Cauchy Distribution



It looks higher in the top than normal, which is true. The shoulders are narrower. You might think that means there's more probability in the center, but there's actually a lot more probability in the tails, also. As an example of a Cauchy distribution, say that in zero gravity you had a wet paint roller. Say that it's sticking up, spinning rapidly and throwing off paint in every direction. Say that there's a wall some distance from it. Most of the paint is going to hit closest to where the paint roller is. But occasionally, you're going to have a flick of paint that comes off almost, but not quite, parallel to the wall, and it's going to hit it about 15 miles away from the paint roller, if it's in zero gravity. That's the Cauchy distribution. It's circular like that, because there's an arctangent even in the denominator of the probability density function or the cumulative density function (I said that I didn't

open a textbook; it's one or the other).

The point is that it's a distribution with $\pi/2$ and arctan and stuff. You get crazy results. If you start sampling from it, you get 2, –1, 10 million, 3, 4, -2 and so on. In fact, you never converge to a mean. That's the crazy thing about it—there is no mean of this distribution. How many of you looking at that would assume that there was no mean? Nobody. You would look at it and think that the mean was right there in the middle. The mode is in the middle, and the median is in the middle. There's equal probability on both sides, and there's clearly a mode. It's counterintuitive, but there's no mean. If you integrate the probability density function times X or whatever, it's not integrable; the answer is infinity. You're never going to be able to make any kind of accurate inferences on that kind of distribution using normal theory.

When is it okay to assume normality in a distribution? I say that it's okay under typically sound conditions (a finite mean and variance, which means that the mean and the variance have to exist) and there's a sum of a large number of independent, identically distributed (IID) random variables or samples from the distribution. Independent and identically distributed means that there are a bunch of trials, they're all independent of each other, and they all come from the same distribution. If you add them up and get enough of them, you start to get something that's going to converge to normality under these conditions. Say we roll two dice. Two dice are not a large number, but they're more than one die. Already you get an estimate of 12.6, where the true 99th percentile is obviously 12 (two sixes). So your error is much lower than it was for one die, and more dice would mean even less error. In fact, if you've ever played Dungeons and Dragons (or any game where you roll a bunch of dice) and thought about the probability distribution, it starts to become bell-shaped; you can see that it's becoming normal.

There are also certain kinds of defined distributions, like lognormal, gamma and Poisson, where if one of the parameters underlying the distribution is large enough, it becomes normal-shaped. So if instead of one interstate going into Los Angeles, we were looking at the entire state of California, and the mean was 300, the normal approximation for that Poisson distribution would be 340. That's almost dead-on.

In the valuation actuary world, when is it absolutely correct to assume normality? The answer is, when you're looking at the mean of a sample of scenarios. The reason is that when you're getting a mean of a sample of scenarios, that by definition is a large number of identically distributed, independent samples from the same distribution, whatever the distribution is that underlies your scenario generator. I'll say that's over 40, but maybe more if the underlying distribution is quite skewed. Use your judgment on that. Most actuaries have agreed, in my career, that 40 is probably enough, but we never use just 40 scenarios, anyway. The fewest we've ever used for any work is 100. If you do that, and the mean and the standard deviation of the distribution of possible results are finite, you've got the conditions you need to prove that under the central limit theorem, it converges

to normality. So if you're sampling for a mean ROI for pricing, you've got it. For valuation actuaries, if you're sampling for the mean present value of ending surplus, it's the same thing. As long as you're looking at the mean, that is something that is distributed normally almost all the time.

My next point is important. When you're looking at such a mean, it's somewhat irrelevant—I don't say that it's completely irrelevant, because, remember, it does affect the number of scenarios you need to run—whether the scenario results themselves look like they're normally distributed.

For example, let me talk about stochastic costs of a guaranteed minimum accumulation benefit (GMAB) after 10 years. They would typically look something like Chart 2.

Chart 2
Stochastic Costs of a GMAB

You would have most of your results, from about the 20th percentile on, that would come up with a cost of zero. You've got several costs that are low, and then in the beginning of the distribution, you're going to have several, maybe three or four, costs that are extremely high. That doesn't look normal at all. In that example, the shape of the data makes it clear that the underlying distribution is not normal, but I know beyond a shadow of a doubt that I've got the conditions I need for the mean to converge to normality.

Let me first explain what a guaranteed minimum accumulation benefit is. How many variable annuity actuaries are here? Just a few. How many of you have looked at the literature enough to have some idea of what I'm talking about? Most of you. Let me explain quickly then. I purposely chose the simplest living benefit, because I knew the audience was just everybody at the Valuation Actuary Symposium (Val Act). It's a benefit that says you're going to buy a variable annuity and put your $100,000 premium in, and I'm going to guarantee that no matter how you invest your money, subject to the conditions in the contract, even if you put it all in large-cap stock, you will get at least your premium back after 10 years, as long as you persist the entire 10 years.

That's a risky guarantee for an insurance company to make. You don't know what the stock market is going to do over the next 10 years. Since we expect a positive bias in the stock market returns, you think that we're probably going to be fine, but if you run stochastic scenarios and are reasonably conservative, you're going to come up with some scenarios that go down over 10 years and produce a big cost to the company. But it's probably going to be only a few with big cost, a few more

with a small cost and then most with zero cost, which is what we saw.

I know that my mean is going to be distributed normally if I run enough scenarios, because the cost is bounded. Even though it seems like a lot—the person puts in $100,000—what's the most that the company can lose? There's no way that the company can lose more than $100,000. If the cost is bounded (if it's not truly infinite), by definition the mean is finite and the variance is finite. That's all I need to know to know that I'm subject to the central limit theorem, and, if I sample enough times, my mean is going to become normal.

The practical conclusion for valuation actuaries is you can probably assume, in most of the testing that you do, that your scenario mean is distributed normally. What does this mean that you can do? It does not mean that your mean plus two of your standard deviations gives you an estimate of the 97.5th percentile of the underlying distribution. I want to be careful about that because I've seen internal memos in my old company that seem to indicate that somebody thought you could do that. It's probably out there. What it does mean is that you can draw symmetric (or nonsymmetric, if you want) confidence intervals around your mean and get a confidence interval for the true mean of your distribution. Remember that you're taking a sample. A sample is the best you can do, and you hope that all your assumptions are correct. Subject to that sample, you can believe that your mean is within plus or minus so-and-so of the sample mean that you've got. You can use normal theory to do that.

You do it using the standard error. The standard error is the standard deviation that you get from looking at your scenario results, divided by the square root of the number of scenarios. There may be an n-1 in there if you're trying to be unbiased. Most of the time, n is large enough that it doesn't matter. You've probably heard that before and might even remember it from when you took statistics.

Let's look at our GMAB example. Let's say that I ran 100 equity scenarios. One hundred is still not much nowadays, but let's say that's all I did. I had 90 that had no cost because my equity scenario ran well enough that even after I took out my contract charges to the policyholder, he or she still had more than the premium at the end. But I had five that cost $5 apiece, so that's $25. Three cost $20 a piece, which makes $85. One cost $41, so that's $126. One cost $74 dollars, so that's $200. This is all, say, per $1,000 of premium. For the last one, I shelled out 7.4 percent of the premium that got paid. There's no way that I made money in that scenario. The company lost money on a present-value basis, once you account for the commission. Even with all the charges, if you're taking another 7.4 percent out at the end, you know that it's hopeless. That's a bad scenario. That's a total of $200 costs over 100 scenarios, so the mean cost is $2. That's the mean of my scenario costs. My best estimate of my expected cost is $2 per $1,000.

The sample variance is ($5^2*5+20^2*3+41^2+74^2)/100 - $2^2, or approximately $81, and so the standard deviation for my sample is $9. Therefore,

my standard error of the mean is $0.90, or 90 cents per $1,000. What I can say is that I'm 95 percent confident that my population mean is between $0.20 and $3.80 per $1,000, plus or minus two standard errors. I'm using normal theory because I've run 100 independent, random, identically distributed samples from my distribution. I know that under central limit theory, my mean is distributed normally. There is some skewing involved. You'll notice that if I go much further than two standard deviations, I'm going to be below zero, which is an impossible result. My confidence limits should be shifted a bit to the right, but, in general, this is close enough. If I ran another 100 scenarios or so, I'd probably be to where it's truly looking normal.

We're talking about scenario mean costs. I realize that we're valuation actuaries here, and valuation actuaries usually care more about the tails. When might mean costs matter most to a valuation actuary? The best modern application is market-consistent valuation. We're talking more and more about valuing benefits on a market-consistent basis. We want to know not what we think as actuaries are the potential losses, but how much it would cost to hedge away this risk. If this benefit were traded on the open market, what would it cost to trade it? In fact, FAS 133 requires that treatment for contracts with derivatives. Some of you are familiar with that.

Guaranteed minimum accumulation and most withdrawal benefits are currently covered. If you want to know which withdrawal benefits might not be covered, we can talk about that at the end. A guaranteed minimum income benefit (GMIB) means you annuitize at the end of the period. That's generally not considered a derivative, because you can't net settle it. However, if there becomes a secondary market that allows net settlement, that could fall under FAS 133. I don't want to get too much into FAS 133. I just want to say that if you value policy options using risk-neutral scenarios to do FAS 133 reserves, you get your costs from the mean. The tail results of a set of risk-neutral or market-consistent scenarios don't mean anything. You get the mean cost, and that becomes your estimate of the market cost. It would be good to know how much statistical error might exist. If you want to know for sure that you're within so-and-so, or that your charge would cover the 95^{th} percentile, say, of possible mean costs, you would want to know your confidence limits on the mean results of your scenarios.

Does this imply that if I run 100 more scenarios on the same benefit I should be within that range 95 percent of the time? No, not quite, because that sample is going to have its own error, and the second sample's error could compound the first sample's error. If the first sample was low, and the second one is high, you could end up outside of the range more than 5 percent of the time. The inference is on the actual population mean, which means that if I ran a gazillion scenarios, I would expect to be within that range 95 percent of the time.

There is one more thing to note. Remember that I had a result of $41 and another result of $74, so that's 2 percent of my sample that's higher than $29. This is,

again, just to point out that you can't take the sample mu plus three sample sigmas and draw any inferences on the tail. It's not saying that the sample is distributed normally; it's saying that the mean result is distributed normally.

What can I do if I want to make inferences on the tail? That brings us to the "what if it's not" part of this session. What if I have a distribution I'm sampling from that's not distributed normally, and I want to make tail inferences? The answer lies in the realm of nonparametric statistics. That's how I'm going to treat it. We'll explore that shortly.

First I want to tell you about something that I came up with during the Web craze that I called the "e-lottery." It's a way I came up with to make some money. I get together a huge crowd of people, say 1 billion. I randomly pick and record a number between one and the number of people (so between one and 1 billion). I get a trusted firm to certify and record my pick. Each of the people puts in $1 and then, without conspiracy (none of them can look at each other's choice), every one of them picks a number between one and 1 billion. All the winners split the pot, but if there are no winners, I get to keep the $1 billion. What do you think are my odds of keeping it? Raise your hand if you think that I've got a good chance. About one-third of the people. Raise your hand if you think that I've got almost no chance. Nobody thinks that. Let's talk about it.

To keep the money and then retire, which is the point, I need everybody to guess wrong. The odds of one person guessing wrong are (n-1)/ n. There are n-1 numbers that are wrong and n numbers. The chance is (1-1/n) if I simplify, where n is 1 billion. If they are all independent and can't look at each other's choice, the odds of all n of them guessing wrong are (1-1/n)^n. As n gets very large, the limit of that formula approaches 1/e. My chance of winning is 1/e, which is a pretty decent shot. It's about 37 percent.

What's my point? It turns out that roughly the same odds apply to the underlying 99th percentile in my 100-scenario GMAB example. What do I mean by that? The chance of any individual scenario being less than the *true* 99th percentile (remember, I'm talking about the true 99th percentile, which nobody knows) is obviously 99 percent. That's the definition of the 99th percentile. Assuming independence, the chance of all 100 of them being less is (0.99)^100, which is (1-1/ n)^n, where n is 100. The 100 is not that large an n, but this is conservative, because the approach to 1/e is from the underside, not the high side. So in 100 scenarios, I've got only about a 37 percent chance that all of my scenarios are less than the 99th percentile of my true distribution. This means that I've got about a 63 percent chance that my worst scenario, which was $74 per $1,000, limits, is higher than the true 99th percentile of the underlying distribution. There's a two-thirds chance that my true underlying distribution is not worse than 7.4 percent of my premium at the 99th percentile.

Better yet, if all I care about is the 95th percentile, my odds go from about 1/e to

(1/*e*)^5. Now I'm 99.5 percent sure that that 7.4 percent is worse than the 95th percentile of my underlying distribution. I'm almost positive. If all I care about is the 90th percentile, it's (1/*e*)^10, and it's many significant digits before I get to a number that's bigger than zero on that probability that I'm wrong.

What if I *do* care about the 99th percentile? Then 63 percent certain is pretty poor. Nobody wants to tell the boss, "Oh, I'm about two-thirds sure that we're not going to bankrupt the company." But then you just need to run more scenarios. If you run 1,000 scenarios to make an inference on the 99th percentile, that puts the odds of your worst scenario being less back at (1/*e*)^10. Assuming the scenarios are correct, which nobody knows for sure (we all do the best we can to have scenarios that are conservative enough), if your company is willing to live with 7.40 and that's the worst out of 1,000 scenarios, that is almost 100 percent sure to limit your 99th percentile, if you've run 1,000 scenarios.

Do I always have to make these kinds of inferences using my worst scenario? That seems to be a pain. Often my worst scenario is really bad, and I might have a result with which we're not comfortable, but intuitively I think that this product should be okay. We shouldn't have to pull it, so something is wrong. The answer is no. The worst scenario is the simplest case, and I wanted to give you something that you'd remember, because what I'm about to tell you, you won't remember. It gets more complicated. With this you can think, "Okay, so my worst scenario isn't even going to bankrupt us. What does that tell me?" You can take that home, remember it and maybe not even have to look up the slides on the Internet.

However, any time you're looking at a percentile and running a scenario, you're setting up a Bernoulli trial. Each scenario is a single Bernoulli trial. If you run enough scenarios, Bernoulli becomes binomial. You create a binomial distribution by running 100 or 1,000 scenarios. You can work out the odds and create your own confidence intervals by using binomial theory. You can say, "Given that I'm trying to prove that my benefit is not going to bankrupt the company more than 1 percent of the time, then my *p* is 1 percent and my *q* is 99 percent." You can work out a range of probabilities around the estimate of the percentile at which you're looking.

You can also do hypothesis testing. I'm going to run through one quickly. It gets a little hairy as far as looking at the detail, but bear with me. If I wanted to do a one-sided hypothesis test using nonparametric statistics, here's what I would do. Say I've run 1,000 scenarios on the total surplus of my company, including this guaranteed accumulation benefit, which seems very risky. I want my surplus level to "pass" 99 percent of all possible scenarios. By "all possible," we're talking about the universe; we're talking about what we're sampling on that we don't know. Out of my 1,000 scenarios that we do know (this is my sample), five out of 1,000 fail and have negative ending surplus. That seems like a good result because five out of 1,000 is only 0.5 percent, and I'm happy if I can prove that only 1 percent of my scenarios fail. Let's see. My null hypothesis in the hypothesis test is that my true 99th percentile result would "fail," because you're trying to disprove your null

hypothesis. If the null hypothesis were true, the chance of a success would be no more than 99 percent, and the chance of a failure could be no less than 1 percent. Let's be as conservative as possible and assume exactly 99 percent and 1 percent. The highest possible probability of no failures (all the scenarios "passing") would be 0.99^1000, or 0.00004, which is tiny.

For one failure, the binomial distribution result would be 0.99^999*.01 *1000, or 0.0004, which is 10 times as big. The highest possible chance of two failures is 0.0022, so now we're getting to be a real number. If you add up the first six results and get up to a possibility of five or fewer failures, the cumulative probability is 6.61 percent. At a 5 percent significance level, I cannot reject my null hypothesis. Even though my result that I tested was five scenarios out of 1,000 and that's only 0.5 percent failing, I don't have enough evidence at the 5 percent significance level to reject my null hypothesis.

If you remember sampling on means and doing hypothesis tests, it's kind of like getting a result where your mean is on the right side of what you're trying to prove, but not by enough standard deviations. In this case, it's not by enough probability, in the binomial sense. If we had four or fewer failures, at 5 percent significance level, we could have rejected. That means we need to increase our surplus, right? No, not necessarily. The result was in the direction we want, just not by enough. We can always refine—look for more evidence—by running more scenarios. It turns out that if we run 10,000, and our failure rate is worse at 60 out of 10,000 (so it's 0.6 instead of 0.5), now we're rejecting the null hypothesis easily; we've proved what we want to prove. The answer often is to run more scenarios if you're not sure and if you can't quite reject.

I have some final comments on confidence statements. We've made several statements of probability or confidence levels. Remember, as the actuary, this is about scenario testing. The statements are based on your scenarios only. In other words, they assume that your lapse assumption, your mortality assumption, your expense assumption and other assumptions are correct. Most of the time, as actuaries, we're comfortable with that because we feel like we've got a good handle on that stuff, but just be aware of what you're saying. If you're making a statement to your chief financial officer (CFO), you may want to couch it and say, "This is assuming all of our other assumptions are correct." You may want to sensitivity test other assumptions. If mortality doubled because of a huge calamity, obviously that changes what could happen to our surplus.

I'm going to go now to some statistical applications. We've talked about sampling theory long enough. I want to give you some hip-pocket notions to carry with you to some other areas in your job such that you might suddenly decide that you could use statistics there and how you might want to think about it.

The first one is use of general linear regression. How many remember general linear regression? How many have used general linear regression in their jobs? It

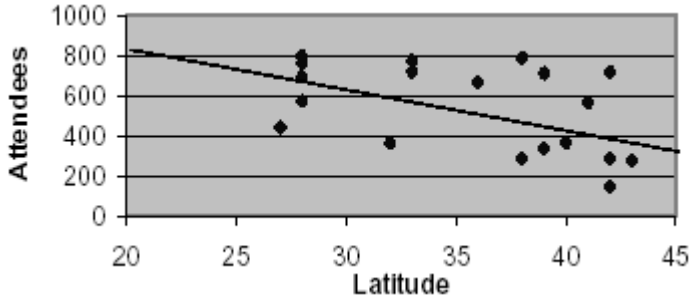
looks like about a third, at least. That's good. It's probably the statistical tool used, but sometimes misused, by more actuaries than any other, as far as statistical tools in Excel or other spreadsheet or any other kind of packages you might have. Why? It's common; it's in almost all of them that I've ever seen. It's easy to remember how it works and what the solution set of parameters means. Everybody knows that you're looking for the beta1, beta2 or whatever. You want to know what your estimators are for the coefficient on your independent variable to get your dependent variable. But it's also fraught with pitfalls, and I'm going to try to teach you how to avoid some of those.

We all remember the goal of linear regression. It's to demonstrate that there is a linear relationship between two or more bodies of data. You minimize squared differences, you estimate your coefficients, and you're done. You're basically trying to predict other data points, often future ones. Seldom do you do a linear regression just to explain the relationship; usually you're doing this because you want to know what's going to happen next.

The last time I remember using this is when I was consulting for a company that wanted to know, before valuation date, a good estimate of what its funds were going to be valued at (the funds held in the variable accounts). It wanted to know, based on the Standard & Poor's (S&P) and the NASDAQ, the probable change in its funds, even though no fund is as broad as those two indexes, and there's going to be some variation. But the regression was good, and you could use it well to do a prediction like that, even when it was a future event, and you didn't know what *y* was going to be. So that's what you're trying to do. It's not just to explain the relationship that is there already, because you already know all those data.

I've done an example. I knocked off in about 2002, partly because I felt like I had made my case, and we'll see that later. I did a regression of Val Act attendance versus latitude, as shown in Chart 3, hoping that the SOA people would look at this and draw their own conclusions and go my way. Obviously you're getting further north as latitude gets bigger. It's colder in late September, and I don't like it when it's below 70 degrees. I've tried to show that attendance was higher when you were in the lower latitudes, like here in Orlando. A couple of years ago, San Diego was a good choice, I thought. So I did my regression.

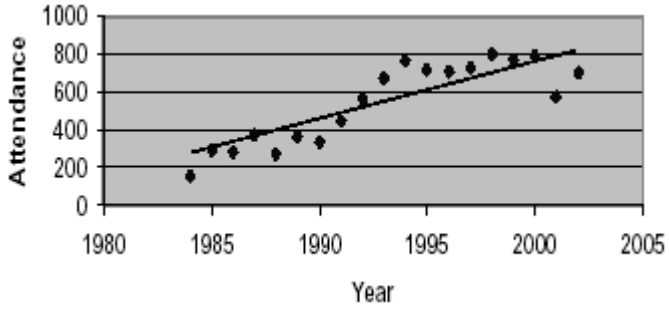
Chart 3
Regression of Val Act Attendance vs. Latitude



Let's look at what I got in my results. I got an X-coefficient of –17. I wanted a negative result, because the higher the latitude, the less attendance you get. I got a t-statistic of -2.16. If you think of it like normal, a standard error over 2 is a pretty good result, because I have a probability value of 4.57 percent, so at the significance level of 5 percent, I can reject and all that. My R-squared is 21.5 percent, so I've explained 21.5 percent of the variation. Before, attendance was all over the place, but if I put in this predictor, latitude, I explain more than 20 percent of the variation in the different attendance levels. That's not too terrible. I could quit and say that I've proved my point.

But now look at Chart 4. It is a regression of Val Act attendance versus calendar year. There are a couple of outliners, but it's a good result.

Chart 4
Regression of Val Act Attendance vs. Calendar Year



What if I add calendar year to my regression as a second independent variable and do a multiple linear regression with X1 and X2 data? I get a coefficient on calendar year of 33.85 and a t-statistic of over 6. Six sigma is generally beyond-a-shadow-of-a-doubt-type proof that something is significant. The probability value is 10^(-5). In other words, it's zero. There's a 0 percent chance that I'm wrong about this. My coefficient on latitude goes to 0.82, with a t-statistic of 0.15, which is almost nothing, so latitude in this regression is no longer a useful variable at all. My R^2

value jumps up to 76.5 percent. I'll point out to you that this happens almost any time you add another explanatory variable. You get more explanatory power, and R^2 goes up, but not nearly as much for a poor variable. If I would have started with calendar year, I would have already been at 76.45, and adding latitude would have made it approximately 76.51. There would have little additional explanation.

What mathematical pitfalls are within this example? First of all, the issue I've already raised, which is selection of inferior predictive data (selection of a bad variable), is not a mathematical pitfall; that's just something that happens. It's an issue, but not a mathematical issue. Often you might think that there's a relationship test for it and not find a relationship. The math is all good. You've indicated to yourself that you don't have a good predictor. In this case, if latitude was all you had to go on, it would be much better than nothing. That's why it tested well when you didn't put in calendar year. If you didn't know that there were calendar years and all you had were data on latitude and attendance, going on latitude would help explain some of the variation and might help you predict. Sometimes you don't have perfect data.

However, there are two important issues here that you may recall. One is the nonconstant variance/outlier effect, which is the fact that you're going to have some data that are kind of crazy. The other issue is extrapolation beyond the range of the data.

Regarding the outlier effect, I'll bring up 2001 on Chart 4 that you saw earlier. It is easily the largest squared error. The year 1994 is at least the same number of significant digits in terms of squared error. Is that point truly indicative? Does anyone remember what happened to the Val Act in 2001? It was scheduled for September 12 or September 13. It was right on the tail of 9/11. It got postponed, and a lot of people cancelled their plans altogether. A lot of people were still afraid to travel, even when it was held in December. It's not truly indicative, and you need to consider this in drawing conclusions. It might even be worthwhile throwing that data point out before you did a regression. Interestingly, when you looked at latitude, it had a relatively small squared error.

The other pitfall that you probably would remember is the difficulty if you try to extrapolate beyond the range of the data. Remember, in these data we had latitudes between maybe 25 and 45 and calendar years between the early 1980s and 2002. If you wanted to use a linear formula to predict the attendance of a past symposium during those calendar years, you'd want to use calendar year as your predictor. Latitude would not be as good. However, despite how things look on that chart, you could find that for predicting attendance in 2010 or so (estimate at 1,214), calendar year is less useful than latitude. It's because you're going beyond the range of the predictive data, and shapes can often change if you go outside of that range.

What has probably happened is that the Val Act grew in popularity over a set of

years, but if you look at data since then, you can already see it starting to tail off. In fact, a parabolic regression (a regression with X and X^2) would fit the data much better than what's there. I'm not indicating that it's going to go down. A parabola would first flatten out, and then even a parabola would start not to fit. Maybe you'd want to use a cubic or something to where it could maybe go up gradually again. We don't know what kind of regression would fit it the best. The point is that we've probably reached the peak of where we've got almost all the valuation actuaries in the country more or less interested. Some will come one year, and some will come another year. But it's unlikely that we can grow too much more, just because of the quantity of potential attendees.

Even in 2010, warm temperatures are going to be what they are now, unless there's a huge climatic shift. Future attendance could flatten out, but you could still have better attendance in Orlando than you would in Pittsburgh . On the other hand, if you used latitude to predict 1,225 attendees in Quito, Ecuador, because you have a zero latitude, that's probably not going to work, either. It would probably be much worse. Again, you're going outside of the range of the data that you have, and you ought to be cautious when you're doing that.

There's one more often-missed mathematical pitfall with multiple regression. You must be careful of colinearity of data. I said earlier that adding variables to a regression almost always provides more explanatory power, as measured by R^2. It can't really *not* do that, because you're adding new ways to look at the data each time you add a set of predictors. Whether it's a good predictor or not, it's going to provide some explanation that the first set didn't. The one exception is if you add a variable that is a multiple of a previous variable, plus a constant. In other words, it is linearly related to the first set of predictive data. If you do that, you get no additional explanatory power. In fact, what happens mathematically is that you get a matrix that you can't invert, and you get no solution to the regression. If you put in something as X2 that is equal to X1 times a multiple, plus a constant, you get no solution. The closer that is to being true (if it's almost true), you start to get a weak solution, and you get almost no new explanatory power from your variable.

It turns out that calendar year is a fairly good predictor of latitude. Regression between the two is more powerful than the first one we looked at between latitude and attendance. That shows that the SOA has already been going my way on this, so I win. I think that last year the meeting was in Boston, but this year's meeting is here, next year is Scottsdale and the year after that is Austin, I think.

The point is that by adding calendar year to latitude, I got two variables that were closely related. The crux of the matter is that despite increasing explanatory power, which is measured by R^2, adding nonuseful variables to a regression can reduce its predictive power, which is the power to predict new data points that might come up. This is because the number of degrees of freedom in the regression is the number of data points minus 1, minus the number of independent variables. When I added X2, if I have the same number of data points, I reduced my degrees of

freedom by 1. If you think of degrees of freedom as ways of fixing something in place, the less you have, the more wobbly it is. The fewer degrees of freedom you have, the more wobbly your estimator of future data points. If you add too many variables, and they're not useful variables, you can muddy the water so much that you end up not knowing what you have, and your point estimates of future data points become unreliable. The sigma is big, because you're dividing by such a small t-statistic.

Let me go over a few variables that may be useful to you if you're doing a multiple regression instead of a single, and they're never colinear with the first one. One is higher powers of the independent variable. If X1 was X1, X2 could be X1^2. That's always going to work. The regression is still linear in the coefficients (the beta1 times X1, plus beta2 times X2, et cetera), but you get a quadratic (or a cubic or whatever) formula on your X variable. That always works. It's never colinear; it can't be colinear. Recall that I added an X^2 on attendance versus calendar year, and the parabola fit better, for instance.

Other transformations you could use are natural log of the independent variable, exponential functions, trigonometric functions and so on. Trigonometric could be useful if you're doing a seasonal regression, and you've got the winter, then the summer, then the winter, then the summer, and you're trying to fit that in.

Now let's go to statistical applications, part two. We finish regression and move on to something called C-3 Phase II. This is going to be especially helpful to the variable annuity actuaries. It's an area where we're really getting to use statistics. We're going to have to do some sampling based on a defined distribution of scenarios to get a defined result, which is also a statistic. If you're not interested in this subject, please bear with me. I think it's interesting because statistics are getting more and more into the valuation actuary realm in this area of C-3 Phase II.

The two new statistical issues are calibration of scenarios and conditional tail expectation (CTE) 90 measure for risk-based capital (RBC). What should we know about both? If you're doing a valuation of a variable block that has guarantees, C-3 Phase II says that you have to calibrate your equity scenarios to fit a certain distribution. Then you take your results, look at only the worst 10 percent and average the loss over those 10 percent, present value back. I'm simplifying a lot to make it simple. That becomes your total asset requirement, and you figure out your required surplus based on that, minus the reserve that you're already holding.

What should we know about both? The calibration points for the equity scenarios are based on statistics. They are measured according to historical rate movements in the market. But if you think about it, how many independent 10-year historical holds do we have in the stock market? One of the statistics is to calibrate your equity scenarios for the hold over the first 10 years, and it has to be at least this conservative at the 10^{th}, the 5^{th} and the 1^{st} percentile. What do we know about 10-year holds? The S&P 500 index only goes back to something like the 1920s. That's

only about 80 years. How about for the modern-era S&P 500 index? Was it the same back in the days of the stock market crash in the 1930s as it is now?

How about for more recent indices that we've added? The NASDAQ hasn't been around that long. We don't have many independent 10-year holds. Even though you can measure 10 years many ways, you realize that if you measure 1980 to 1990 and then 1981 to1991, there are a lot of similar data in those two periods; they are not at all two independent examples. By the time you get to 1985 to 1995, you've got five years of dependence and five years of independence. It's not until you go to a new decade that you're completely independent. We have very few samples, so one has to develop distant (say 10-year) calibration points using some theory about distributions. You can't do it just by sampling on what has happened.

Here's the gut check. Who developed the regulatory scenario set? Regulatory people. How would you describe regulatory people? Conservative. What kind of distribution do you think they chose to develop the calibration points? A conservative one. That means your results could turn out to be very conservative results, as far as the surplus you have to hold.

How can you calibrate your equity scenarios? By using some fairly conservative assumptions of your own. I found that you probably cannot assume mean reversion, at least when I've tried it. Maybe some of you could come up at the end of this presentation and say that you've done it with mean reversion, it worked, and here's how. I'd be interested. But I found that you can't do that. Mean reversion is when you assume that if a scenario starts out poorly, there's a tendency for it to come back up because assets and the economy are growing at a certain rate, and even if valuations change, it should come to the same place after enough time. I don't think you can assume that for a 10-year hold under these calibration points.

You probably should have a fairly diffuse distribution, with fairly wide parameters, in other words. The trick is that you want to try to fit all your calibration points. You've got to fit them all, or at least be confident that you're within materiality fitting them all, without badly overshooting on some of them. That can be a handy trick. Simply lowering your mean return or inflating your implied volatility could cause overshooting on some of your points. You may want to use a combination of the two or come up with your own alternative approach.

Can you just "stretch out" your real-world set that you've run? Can you run your set how you like, and then, if you don't fit, push some of your outer scenarios out in some way? The verbiage in the guidance seems to allow that approach. I read it, and I think it does. It's up to valuation actuaries to decide whether this is acceptable to you, though, and whether it produces adequate capital. As far as other asset classes besides equities, the guidance says that you can come up with your own calibrations, but you've got to be consistent. That's all it says. There are probably several variations you could come up with on that.

What is CTE 90? I've gone through it a little bit. It's the average of your worst 10 percent of possible outcomes estimated using your stochastic set. At one time in the valuation community, for getting required capital or for estimating target surplus on a new type of block where we didn't have a formula, we may have been happy with a value-at-risk measure. For instance, say our surplus has to cover the 95th percentile of results. I can tell you that my company has some heavy-tailed benefits that are just like this. For heavy-tailed benefits, we needed something more robust than a value-at-risk measure. The reason is that there can be benefits like the GMAB example that I gave you, where at the 95th percentile, the cost is low or even zero, but going on out beyond that, the costs are highly significant. We needed a measure that included a feel for what the total tail costs were, out to the worst possible. By taking all of the worst 10 percent of possible outcomes into account, CTE 90 helps account for risk and gives you a better surplus measure.

What are the basic statistics in a CTE 90? The first thing to realize is that we're estimating two quantities in a CTE 90, not just one. The first one is the 90th percentile. We're getting an unbiased estimate for the 90th percentile using our scenarios. If we run 1,000 scenarios, somewhere around the 900th scenario is an unbiased estimator of the 90th percentile. Once we know that, we can estimate the conditional distribution of losses, given that they exceed the 90th percentile.

That's why it's called a "conditional" tail expectation; it's a conditional distribution, given that it's worse than the 90th percentile. It's a conditional mean, but it is a mean. It's a mean with a variance that's more volatile than a typical mean. A typical mean is a CTE 0; it's the mean given that your losses are worse than nothing. CTE 0 is the whole distribution. CTE 90 is a mean, but it's a conditional mean.

Here's why the CTE 90 variance is more volatile. One reason is that the estimator is only asymptotically unbiased. What do I mean by that? If you ran a gazillion scenarios—whatever a gazillion is—it would be unbiased. But if you don't run a gazillion scenarios, it's not unbiased; it's biased, and it's biased in the wrong direction, because you're estimating the 90th percentile.

As I said earlier, if you do a correct treatment of order statistics, around the 90th percentile of your sample is an unbiased estimator of the percentile. So a value-at-risk measure is always unbiased. But when you combine the two measures, here's what happens. You estimate the 90th percentile, and then you're also putting in kind of an estimate of every percentile above that. The further up you go, the less unbiased the estimator from that point on is going to be, once you've set your 90th percentile. Think about what the worst statistic in any scenario set has to be. It's always your maximum: your worst cost or your worst surplus result. If you think about it, if you run more scenarios, it's always either going to be the same or worse. It's the one thing that can never go the other direction. The 90th percentile, if I run more scenarios, could go either way. It could actually go down. It's unlikely, and the further it is from the median, the less likely it is, but it could happen. The

maximum is always going to get worse or stay the same, with the more scenarios you run. If you have an infinite distribution, you can sample all you want, but it's always going to get worse and worse and worse, forever. You're never going to get a scenario that produces the maximum.

The problem with a small sample size estimating the CTE 90 is that the fewer scenarios you run, the worse that maximum is, and, in fact, all the other numbers beyond the 90th percentile, but not as bad as the maximum. You're always going to be biased downward, and you're always going to be biased unconservative. You need to watch sample size. There's a similar bias in the estimator of the mean standard error, but it's small.

If we use the mean and standard error estimate, since this is a mean, we can build a confidence interval around our CTE. That's the good news. The bad news is that there's bias, and it's volatile. The good news is that it is a mean, so we can use the mean and the standard error to get an estimate, and that gives us confidence so that if we don't run a gazillion scenarios, we can put confidence limits on where our true CTE is.

What should we do? Be conservative about sample size. If you're estimating a CTE of a nearly normal event (as an actuary, you need to make that determination), a relatively small sample could be sufficient, and even a CTE 99 could be measured accurately. For most of the distributions, you're going to be doing CTE 90, but you're going to be doing fat-tailed stuff. For those, the bias isn't so bad a problem if you run maybe 1,000 scenarios. However, the combined variance of the two parameters could be a problem. The bottom line is that the papers I've read on the subject say that you have to use at least 1,000 scenarios if you're doing CTE 90, and the reason is the volatility of the parameters.

I'd like to discuss some miscellaneous issues and then give a quick summary. I find that there are many times that you have a limited sample size. We talked about one example where we didn't have enough data. You don't have enough data, and there's no way to get more, so what do you do? When I say that, it's about the S&P 500 history. You don't have a lot of sample size right now. The question is, Can you do anything with statistics there, or do you just throw up your hands? Does anyone else have any other examples of where they've had limited data but needed to draw conclusions somehow?

FROM THE FLOOR: Lapse sensitivity.

MR. ROBBINS: Lapse sensitivity is a good one. Say we don't have enough years yet in a product. We want to try to draw some kind of conclusion, but we're not sure what we can do. Expense studies could work the same, as well as mortality studies. Anything else?

FROM THE FLOOR: Interest rates.

MR. ROBBINS: Interest rates are similar to equity rates, right. Yes, that's a good point. Do we throw up our hands and surrender, since we don't have enough data to converge to anything? My answer is no. My answer is that it depends on what your null hypothesis is. If you look at S&P 500 history, what we're trying to do is determine what we can say about the probability of a negative 10-year gross total return.

A lot of people in the past, probably in the 1990s, designing the original living benefit structures that we've had, have said, "Even in the stock market crash of the 1930s, a negative 10-year hold on the market has never happened." That's a fact. It has never happened. If you look at the index, there's a 10-year period where it has happened. But if you factor in dividend returns, it's unlikely. It's interesting. Dividend returns buy you more index when the index is down; it's kind of like dollar-cost averaging. There's a little bit of added insurance there than on a variable product where you get dividends. But to say that it has never happened carried some weight, and I think that intuitively people thought the stock market crash of the 1930s was as bad as it could get. It has never happened here. I think that what started to scare the markets and some people is Japan, because I think it has happened now there, right? Am I correct in saying that? I think that we've gone 10 years, and it's still not close to where it was at the beginning of the 10-year period. It's potentially possible in the market to have a negative 10-year hold on stocks. What can we say, though, about the United States markets if it has never happened here? We think that maybe we have something like eight independent 10-year periods at which to look. Can we draw any statistical conclusions from that?

It depends on how reasonable our null hypothesis is that we're trying to disprove. Remember, you set a null hypothesis. As far as what I can do with a sample size of eight, it depends on the conclusion that I'm trying to reach.

What if I'm trying to say that I think the chance of a negative 10-year hold in the future is not greater than 1 percent? That's a tough conclusion. Things are not so good in that case. If it were true that the market could not go down over 10 years more than 1 percent of the time, eight in a row times that it didn't happen, that would happen 92.3 percent of the time. If something happens 92.3 percent of the time, yes, of course that would happen. It's not proof of anything. You can't reject that hypothesis and say that no, you don't think so.

What if I'm trying to say that I think there's no more than a 10 percent chance of a negative 10-year hold? Now we've got an argument, but it's not a strong argument. Again, that's like saying 0.9^10. If it should be 90 percent likely that we get a positive, 0.9^10 is about 43 percent. That's only 57 percent, as 1 minus that probability. Forty-three percent is still not that strong an argument. It's like saying that it's less than half likely, so that's good. There's more evidence than not, but not enough evidence to form a strong conclusion.

But if I'm willing to be happy with no more than 30 percent likely that we'll ever have a 10-year hold that's negative, now I can reject that at the 5 percent significance level. You can work this out in your head. If I wanted to say that eight times in a row conclusively proves that you wouldn't get a negative 10-year hold more than half the time, now you can just say, "Oh, yes—two, four, eight, 16, 32, 64, 128, 256. If it were a 50/50 shot of getting eight in a row of no negative 10-year hold, that can happen only one time in 256. Now I can reject that. There's no way. It's just a 50/50 shot." Somebody might say that he or she thinks that it's 50 percent likely right now that the market is going to go down over the next 10 years, and you could say that that's ridiculous. Based on the historical evidence, unless something is changed, there's no way.

Put that in other terms you might want to use, say the probability of the Saints winning the Super Bowl this year. If it were a decent shot, it should have happened before. So no, that can't happen. If you root for the losing team in every sport like I do (the Saints, the Cubs), you know what I'm talking about. You can draw conclusions based on not much data, if you're willing to pick something that is reasonably unlikely or a conclusion that's reasonably easy to prove.

If you need to prove something more difficult, you just need more data. If you need to prove something more difficult, and you don't have enough data, you're stuck, and you need to look for another way to design a product or something like that.

My point hasn't been to argue about how likely a negative 10-year hold might be, but rather to show that you can draw valid inferences if you're not too demanding on your data. By the year 2200, when we're looking at independent 10-year holds, we could know enough, if we haven't had any negative 10-year holds, to reject at 10 percent likelihood that there would ever be one at the 5 percent significance level. To get a feel for a 1 percent likelihood, we'd need 100 trials, and that's 1,000 years. I don't know if we'll have stock markets in 1,000 years. That, in fact, is why a 10-year put on the index does in fact cost something, even though it has never happened before that the put would have been in the money.

Let me summarize. To summarize my first two sections, normal is somewhat normal. You do find a lot of applications where you can use normal theory, but you have to be careful about approximating the shape of a distribution that way. It's mostly useful for approximating where your scenario mean has wound up and for getting a feel for the overall mean. In most of your distributions, as a valuation actuary, the mean will converge and become normal if the sample is big. When a distribution is not normal and you want to get a feel for the tail, you should be using some kind of binomial theory with nonparametric statistics.

When generating and using stochastic scenarios, it's good to be clear on what some of the statistics really mean, such as CTE 90 and tail parameters. Regression is useful, but it's important to realize that increasing R^2 is not the only goal; you want

to increase your predictive power as much as you can and not muddy the water with extraneous variables. Even limited data can be useful if you're not too demanding with your conclusions that you're trying to draw.

MR. BERNARD RABINOWITZ: I'm a health insurance actuary, so I'm not familiar with a lot of the terms you used. But can you explain calibration of the model? What does that mean?

MR. ROBBINS: It's calibration of the equity scenarios. The requirement states that if you're doing a model for equities, for funds that are generally invested in something like an S&P 500 index, after one year the 10^{th} percentile cannot be any better than, say, a 10 percent drop. The 5^{th} percentile out of all your scenarios cannot be any better than a 15 percent drop, and the 1^{st} percentile can't be any better than, say, a 25 percent drop (something very conservative). After five years, you also have a required 10^{th}, 5^{th} and 1^{st} percentile, and after 10 years, again, you have the same thing. You have to make sure your scenarios are diffuse enough—spread out enough—that you meet these requirements for them before you're allowed to use those scenarios. Does that answer your question?

MR. RABINOWITZ: Yes. I have another one. I may not have been exactly truthful when I said that I'm a health insurance actuary. I used to be a life and annuity actuary, way back in the 1960s and 1970s. Our company had a subsidiary in Canada, and we were selling equity-linked whole life insurance. We put a guarantee on it, and we tried to cost the guarantee. These are the days before you had the type of computing power that you have today. One of my colleagues came up with a method, and it involved looking at the daily swing in stock prices. Sometimes it went up, and sometimes it went down. You assigned that the probability of going up and going down was the same, so all he looked at were deviations. We built a model. It took three or four days to run on the company's mainframe. We'd do quite a few hours at night, and we crunched it through. We got some interesting loss scenarios. It's like saying that we haven't had a hurricane for 20 years, so it's not going to happen. This is what I'm hearing.

MR. ROBBINS: That's interesting. What you're describing is kind of a binomial lattice approach, where every year or every period, you have an equal probability of an up by sigma or a down by sigma?

MR. RABINOWITZ: Yes. When you run it yearly, you get entirely different results. But when you run it daily, the results show much more volatility.

MR. ROBBINS: Sure. If you run it daily, it sounds simple and could be scoffed at now, but it should produce good results. That approach is still the basis of option pricing theory. All Black-Sholes is is the continuous equivalent over time of a series of discrete jumps like that. If you do it daily to get results that don't matter daily, and you end up with a distribution at the end of a year, it would be a good model.

MR. RABINOWITZ: But using that method, there was a cost.

MR. ROBBINS: That's good. There probably should have been. I think that Canadian variable products, which were called segregated funds, have had things like guaranteed accumulation benefits a lot longer than they did in the United States. Companies used to think that the costs should be about zero, and then they brought in more option pricing techniques and found out that they were totally underpriced. The United States was probably just lucky that it didn't develop those guarantees as early, before the theory caught up.

MS. SHIRLEY HWEI-CHUNG SHAO: As we move toward the principle-based approach, are there going to be more reviewers, for example, the corporate actuary department that wants to review what the variable annuity people are doing ? There's also going to be a reviewing actuary, maybe an independent reviewing actuary. There will also be regulators.

MR. ROBBINS: Sure.

MS. SHAO: Do you have any suggestions on what kind of role these reviewers should perform when it comes to what you're talking about? How much should we be reviewing? Should we be reviewing whether the statistic inferences are made correctly and all that?

MR. ROBBINS: In a session just on statistics, as far as telling people what they should be doing as far as their job, I'm not sure that I have much to say. I would look for the basics, like what I've taught today. If people give you a CTE, are they just giving you the CTE from their scenarios, or are they telling you the confidence limits on the CTE? Say they've done only 1,000 scenarios. How do you know that CTE doesn't happen to be an unconservative sample and that it could be higher if you did a different 1,000 scenarios? Much simpler, if someone is telling you that the 99^{th} percentile cost of this benefit is such-and-such, did that person just pull the 99^{th} scenario out of 100, or did that person really figure out a bound on the 99^{th} percentile of the true distribution? It's that kind of thing. I would open up the questions like that, and I would probably be caught if someone asked me things like that at my company, frankly. Often, you take the simple road when you're trying to answer a question about a percentile or a statistic. You don't fully think it through and think what the worst case could be.