2003 VALUATION ACTUARY SYMPOSIUM
September 11–12, 2003
San Diego, California

Session 24TS
Drawing Appropriate Statistical Inferences

Instructor: DOUGLAS L. ROBBINS

Summary: Stochastic testing has become an integral part of much of what valuation actuaries do. But can the appointed actuary be sure that they are making the correct inferences when looking at the results of a large number of equally likely trials or scenarios? In this session, the instructor provides some simple tools and guidelines to ensure that correct conclusions are in fact drawn. This material is presented in an easy-to-follow manner without using complex mathematical formulas. Topics include: applying the assumption of a normal distribution—when it is appropriate and (sometimes more importantly) when it is not; drawing inferences about, and creating confidence intervals around, a population mean; drawing inferences about, and creating confidence intervals around, tail percentiles; and certifying anticipated mortality under XXX in a mathematically concise way.

MR. DOUGLAS L. ROBBINS: I've decided to try to seriously work out a presentation that I could do straight out of my head. I have not gone to any textbooks at all for this. There are not very many formulas that you're going to have to worry about. I want to give you some basics of things. If you can remember these techniques and tips, they will help you to draw statistical inferences when you've been given a body of data, and especially as valuation actuaries working more and more with large sets of scenarios. I think it's a good idea for people to have those concepts.

I'm going to start out with two sections drawn from the title of the presentation, "How normal is normal?" and "What if it's not?" Then I'm going to go into statistical applications parts one and two. I guess you'll see what that means when I get there. I'll close with miscellaneous issues and a summary.

How was the normal distribution taught when you were in school? I'm not really talking about actuarial exams here. I'm talking about college, like cookbook statistics. There are anecdotal arguments that many things in life kind of have a

bell shape, at least this is how I was taught. You take a quick look at the probability density function in tabular form, and then you're introduced to the table from which you can quickly draw stochastic inferences about anything with roughly a bell shape. You get questions on your exam like, "If the amount of rainfall in Washington State over a given month is normal distribution, then what's the 95 percent confidence interval?" There is no evidence that it is so. Well, consider the following. For how many of these distributions would you feel confident estimating the 99th percentile using a normal approximation with a mu and sigma? First, for the number of accidents on a typical freeway into a large city, in morning rush hour, I assume the mean is two. Second, consider the result of a single randomly rolled fair die. Third, consider the number of days late a given Valuation Actuary Symposium session will be filled with speakers.

Well, maybe normal isn't so normal after all. Here's how you might do. If the true average number of accidents on the freeway is two, then you might, and I'll explain why later, use normal tables and estimate a bit over five. The true 99th percentile is in fact just over six, so that's an undershot on your part of about 20 percent of the amount of your guess. The mean of a roll of a single die is 3.5; the standard deviation is about 1.7. Using the normal approximation, we would use 2.33 sigmas to shoot for the 99th percentile and we would guess 7.5. We've overshot a bit wildly. On the other hand, it turns out that although the average lateness of these sessions to be filled might be a week, the 99th percentile is an estimation you wouldn't want to make based on the whole body of data. Sessions are usually either on time or may not get filled until the week before the symposium, as I know well. Researching the past records of given sessions could make your guess a shoo-in. That's a little lesson in bipolar data.

Generally, when is a normal approximation a bad idea? Well, it is almost always a bad idea when your data is somewhat scarce or bipolar in nature. Much of the time it is a bad idea when making tail inferences on an underlying distribution, given a distribution in trying to make an inference on, say, the 99th percentile or the 95th, in other words. Sometimes the non-normality of a body of data is clear from the sample data's shape. If you think about exponential distributions like the life of a light bulb, you can look at that and say you don't want to do that using a normal approximation, but how about one like the Cauchy distribution—one that's really bell-shaped? This distribution represents something like if you had sitting in space a big spinning circular paintbrush and paint was flying off it in all directions. In one direction there's a wall not far from the brush, so a lot of the paint goes right in the middle, but every once in a while (this is in zero gravity, by the way) a bit of paint flicks off almost parallel but not quite parallel to the wall and goes almost forever before it hits. If you start sampling from this distribution, you're going to find that your sample mean never gets very close to zero for long. You'll get samples from the distribution that are three, minus one, 10,000, minus two, something like that. In fact, what you have is a distribution with no mean. It is pretty weird to think about something that looks like this not having a mean. It does have a median right in the middle. The probability is equal on either side, but there is no mean. I

think the denominator of the PDF has pi in it, which relates well to this circular phenomenon that I was talking about. I only talked about that that much because that distribution violates one of the key criteria to use normality at all, even to try to work on the mean of it. I'll get to that in a second.

When is it okay to assume normality in a distribution? First, it is correct to assume normality under typically sound conditions—those conditions are a finite mean and variance, which Cauchy does not have. You need to have a sum of a large number of independent identically distributed (IID) random variables. Any time you sample a bunch of times from a distribution and the samples are all independent and distributed identically, and you either sum them or take the mean (which is the same thing), it's the sum divided by n. When you sample more and more of these, you quickly get a situation where the sum or the mean converges to normality.

For instance, if we're rolling two dice instead of one, our 99th percentile estimate becomes 12.6. That's not a mean; that's a sum. If you sum a whole bunch of dice, like six or seven dice, you're going to find that it's extremely well fitted to a normal distribution because each die has a mean and a standard deviation. They're all identical. Your error there is only 5 percent. The more dice you roll, the less error you would get.

For certain kinds of defined distributions, if one or more parameters are large enough, common examples exist. Take the example I gave earlier of car accidents. I guess I should go back and say the reason I guessed a little bit over five with a mean of two is because I was assuming the number of car wrecks on the freeway is Poisson. If we look at the whole State of California instead of just the freeway—if you sum up the whole state and the mean is 300—then the normal approximation for the 99th percentile would be about dead-on at 340.

In the valuation actuary world, when is it absolutely correct to assume normality? One time is when you're looking at the mean of a sample of scenarios, as long as you have a reasonably large amount of scenarios—over 40. Usually 30 or 40 is the number you're given in statistics. Perhaps you should use a bit more if the underlying distribution is quite skewed. We'll actually see an example of that a bit later. Add to that the requirement that the mean and the standard deviation of your distribution of possible results are finite, and sometimes you can't look at a body of data and say, "Obviously my mean and variance are finite." It does turn out to be true most of the time, however.

When looking at such a mean of your sample, it's somewhat irrelevant whether the scenario results themselves look to be normally distributed. That's really important. There's a difference between trying to get to a tail of a distribution like the distribution of your results versus the distribution of the mean of the results. When I'm talking about a mean, that's when I'm saying it's somewhat irrelevant anyway about the shape of the distribution that you're sampling from.

An example would be the stochastically derived cost of a guaranteed minimum accumulation benefit (GMAB) after 10 years. Many of you probably sell a benefit just like this at your company. The policyholder buys a variable annuity, pays $10,000 or so for the premium in year one. At the end of year 10, you guarantee they're going to get $10,000 back, and you sample this over say 100 scenarios. After 10 years you're going to get a cost distribution where most of your results are zero, but in your tail you have some fairly significant costs. The funny thing about that example is the shape of the data makes it obvious that the underlying distribution is not normal—it looks pretty darn skewed. However, I know beyond a shadow of a doubt that that distribution meets the criteria for the mean to be distributed normally if there are enough scenarios. Why? It meets the criteria because my cost is bounded between $0 and the premium that was paid—$10,000 if they paid $10,000 of premium. I'm trying to say that if they paid $10,000 of premium, your cost can't be greater than $10,000. Although that's a lot of money, it can't be more than that. Because of that, the distribution has to have all finite moments. The only way you can have a non-defined mean or standard deviation is if the distribution itself is infinite, and this one is not.

My practical conclusion is that you can probably assume in most testing that you do as valuation actuaries that your scenario mean is distributed normally if you have enough scenarios. What does that mean I can do? It does not mean that for my GMAB, I can take my scenario mean plus two standard deviations and get the 97.5^{th} percentile of the underlying distribution. That won't work. It does mean that I can take whatever my mean is and draw symmetric or non-symmetric confidence intervals around that mean. That is my confidence interval then for the true mean of the benefit. The standard error of the mean, by the way, which I would use to do that, is the standard deviation divided by the square root of the number of scenarios. I think for this to be unbiased, there's supposed to be an "n-1" in there somewhere, but I'm not going to look in a textbook. I promised you I wouldn't do that. It's either in the derivation of the standard deviation, which is where I think it is, or in the number of scenarios. I think it's the first.

Let's look at our GMAB example. Let's say that I ran 100 equity scenarios to cost out this GMAB and I have 90 that produce no cost; five that produce $5 apiece; three that were $20 apiece; one that was $41. There is a final big one that cost me $74 per $1,000—7.4 percent, which is quite a large cost, as I noted. The mean cost is clearly $2. You can work it out for yourself if you like. The sample variance is $81; therefore, the standard deviation is $9. I'd get a slightly different answer if I did add minus one, I guess. Dividing by 10, I get a standard error of the mean of 90 cents. What I can say then is that I'm 95 percent confident given the normality that my population mean is between 20 cents and $3.80—plus or minus two standard errors. I'm getting really close to zero on the bottom end. There's not much room there for 2.5 percent of confidence. For this distribution, because it's skewed, it may be true that even 100 scenarios aren't quite enough for the conversion to normality to be perfect. It's probably a little bit off, but it's probably not off that much. This confidence interval is probably pretty decent. Actually the

problem is that the true confidence interval would still not be symmetric. So given that we've moved both ends of it a little to the left on my distribution, it's probably adequate to say that close to 95 percent of the probability is in that range.

Does what we just did imply that if I run 100 new scenarios under the same conditions, but with a new random seed, I should be within 20 cents and $3.80 95 percent of the time? No, it doesn't quite imply that, because that new sample will have its own error and that error could compound the error in my first sample. The inference isn't on a new sample; it's on the actual population mean. The two confidence intervals that I get though will probably overlap pretty well. They'll probably be real close, and the new confidence interval that you would get from looking at all 200 scenarios together is obviously better than either sample by itself.

Also note that the mean plus three standard deviations, which would be that percentile if it were normal, only gets you almost $29. Fully 2 percent of my sample that I developed is well above that point estimate, so you can see that the underlying distribution is very non-normal. We should not try to estimate the 99.87^{th} percentile this way.

What can I do if I want to make inferences on the tail instead of on the mean? That gets me to the second subpoint of the presentation—what if it's not? What if I can't make inferences on the tail based on a normal approximation, what can I do? The answer to that lies on the realm of non-parametric statistics, and we'll explore that shortly. First, I want to tell you about an idea I came up with back in the late 1990s during the Web craze called the e-lottery. In this idea, I get together a huge crowd of one billion people and each one of them gives me $1. They get to each pick and record a number between one and one billion. They can't conspire with each other. I get a trusted firm to certify and record the number that I pick, and it's a lottery at that point. Each of the people puts in $1. If there are a bunch of winners, they split the pot. If there's one winner, he or she gets the whole $1 billion. If there are no winners, I get to keep the $1 billion. What are my odds of keeping it? Does anyone have a guess? It looks like there are no guesses. To keep the money and retire, I need everyone to guess wrong. The odds of one person guessing wrong are [(n-1)/n]=[1-1/n] (where in this case n is one billion). If they're all independent, which we said they are, because none of them are conspiring with each other, the odds of all of them guessing wrong are [1/1/n]^n. As n gets very large, this approaches 1/e. That's a pretty decent shot for me; it's almost one in three. So now you know why this is called the e-lottery. What's the point? It turns out that roughly the same odds apply to the underlying 99^{th} percentile in my 100-scenario GMAB example. Remember, the chance of any individual scenario being less than the true (where true is the unknown thing that we're trying to sample) 99^{th} percentile is clearly 99 percent—the odds of any one scenario coming in below that. Therefore, the odds of all 100 of them being less than the 99^{th} percentile is 0.99 to the 100. The 100 is not that large an "n," but the approach to 1/e is from the underside, which is good from the perspective of trying to limit my probability, so it's actually less than 1/e. So in 100 scenarios, there's only about a 37 percent chance that all my scenarios are less

than the actual 99th percentile of the underlying distribution. Looking at it from the other side, that means that I have a 63 percent chance that my worst scenario limits the 99th percentile of the underlying distribution. Out of my 100 scenarios, remember I had one where I lost $74 out of my $1,000 of premium. There's a 63 percent chance that the true distribution—the 99th percentile—is no worse than that.

Better yet, if I'm doing the sort of testing where I only care about the 95th percentile, my odds go from 1/e to (1/e) ^5. So that makes me about 99.5 percent confident that my worst scenario limits the 95th percentile of the underlying distribution. Now I'm getting somewhere. I mean 99.5 percent confidence is a really good confidence. If it turns out that I'm willing to lose $74 per $1,000 95 percent of the time, then I'm 99.5 percent confident that that is, in fact, going to work out with my sampling. If you only care about the 90th percentile, it's (1/e)^10, and the worse scenario is almost absolutely certain to be a rock-solid boundary.

What if I do care about the 99th percentile? In that case, 63 percent confidence is pretty poor. No one wants to be only 63 percent confident. The answer is yes, it is, but then what you probably need to do is just 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 worst back at (1/e) ^10. Again, it's almost rock solid at that point.

Do you always have to use the worst scenario? The answer is no; it's just the simplest case. Remember, I'm trying to give you something simple enough so you can take this home and a year from now remember oh yeah, there's something about if I lost $74 in my worst scenario and as a company we're okay with that being our somewhat worst case, then there's a way I could think about how many scenarios I'd run. Maybe you can look back at this presentation and get an idea just how confident you are that that scenario limits what your worst case actually is.

Any time you're looking at a percentile though and running a scenario, the idea is that you're really just setting up a Bernoulli trial every time you run a single scenario. At the 95th percentile, either a scenario is bigger, 5 percent chance, or it's not, 95 percent chance. For any percentile of the underlying distribution that you're sampling on and any ranked scenario, you can work out the odds using binomial theory. I'm not going to go into it, but you can. That methodology can be used to construct any size confidence interval you want based on any percentile, or with completed scenario results you can test a stochastic hypothesis and how that would work.

So you've run 1,000 scenarios to make an inference on the total surplus of your company, and you want your surplus level to pass 99 percent of all possible scenarios. You do not want it to pass 99 percent of the scenarios you ran, but 99 percent of all possible scenarios that could be run using your set of assumptions.

Five of your 1,000 scenarios fail (negative ending surplus). The question is if your null hypothesis is that your true 99th percentile result—not, again, the percentile of the sample, but the true 99th—would fail, then can you reject that? The way you test it is this. If the null hypothesis is true, the chance of a success can be no more than 99 percent. The chance of a failure can be no less than 1 percent. Here's a one-sided hypothesis test. Under the null hypothesis, the highest possible probability of zero failures then would be precisely four in 100,000. For one failure, the figure would be about four in 10,000. The highest possible chance of two failures would be about 22 in 10,000 and so on. So you can get a cumulative probability of five or fewer failures, and remember five failures is the result you got. Most conservatively that cumulative probability is 6.61 percent. If the significance of your test was 5 percent, you could not reject a null hypothesis based on this, because the chance of you're having five or fewer failures is bigger than your willingness to be wrong, or 5 percent. If that's true, even though your point estimate of the 99th percentile is a success by a wide margin, remember we're sampling on the 99th percentile. The 99th percentile of the sample is 10 failures and we only had five. Doesn't that indicate a problem with your testing, that you didn't reject, even though you were so far on the correct side of the parameter that you were trying to sample for? The answer is no. It's really a corollary to the situation where you're testing for a mean. Everybody remembers how to do that using the normal table. You get a mean; you get an estimated variance; and if you're not on the correct side of the value you're sampling for by more than the right number of standard deviations, you can be on the side you want to be, but not by enough and still end up not being able to reject. This is the same situation.

If we had only had four or fewer failures in this case, then we could have rejected at the 5 percent significance level. Does that mean then that we need to increase our surplus? No, that's probably not the right thing to do. The right thing to do is probably to run more scenarios, because remember we did only fail 0.5 percent, which is much better than the 1 percent that we want to fail at a global universal level. We can always refine our estimate by running more scenarios. If we ran 10,000 scenarios and our failure rate even went up a little bit, but we got 60 failures, we'd reject easily.

I'll make some final comments on confidence statements, and then I'm going to editorialize just a little bit. In this section we've made several statements of probability or confidence levels. Remember that these statements that we're making are only based on scenarios. We've run 100, 1,000 and 10,000 scenarios, and we're saying that we are 99 percent confident that if you ran a zillion scenarios that 95 percent, say, would pass whatever testing we're doing. We assume all your other model assumptions are correct. For many actuaries, that might not be a problem because you're willing to live with some fluctuation in lapses and mortality and other assumptions like that, but in the long run you feel like they're more predictable. You can rely on them more than economic conditions, but you still need

to be aware of the basis for your confidence statement, and you may want to sensitivity test other assumptions.

Editorially, all I want to say is this. There is a tendency, I think, in written work by actuaries to state that 95 percent of their scenarios passed the testing they were doing, maybe 95 out of 100 or something like that. Sometimes we state that we feel like this is the 95th percentile of our sample, therefore it's an estimate of the 95th percentile of the universe. What I'm trying to give you is a methodology of quantifying how much potential error there is in that. I think this was mentioned last week at the stochastic modeling symposium on a different topic. I think if a point estimate is given to a senior actuary, or worse yet, management, and it's not quantified, they're going to think that it's more reliable than it is. They're going to say this is the number, this is our 95th scenario, therefore, this is the 95th percentile and they're going to rely on that. The truth is, if your distribution is very skewed, your worst scenario—the one we talked about, which limits the 95th almost all the time—could be like five times as big as the 95th. How many of you have seen that when you're testing with a variable annuity benefit? If the distribution is that skewed, they just might not realize how bad that error is. I think we need to get better about providing confidence limits or confidence intervals around estimates that we give.

The other editorial comment is that I think that the lack of people doing that up until this point has led to some silly things being said in official guidance that actuaries are given. The C3, phase two guidance for instance, kind of tells you that you need to run 1,000 scenarios. It just says that; that's just a number. It doesn't say why; it just says you need to run 1,000. For some things that you might do, it's clear that that could be too many. We've seen that that might be the case. If you have a benefit that easily passes in all of 100 scenarios, you might be quite happy with leaving it alone at that point and not need to spend the run-time to do the other 900. The other dangerous thing is you could be in a situation where you really ought to be running 10,000 scenarios, and that might be true even not just to pass C3 phase two, but for some purposes of your own. The C3 phase two is kind of designed to prevent your company from going insolvent, and for that a CTE 90 standard might be what you want to shoot for, and 1,000 scenarios could be enough. On the other hand, you could be designing a new benefit where your personal bonus is going to be dependent on it, in which case you might want a much higher standard than just the solvency of the company.

I'm going to go into some statistical applications now. The first part deals more with generation of scenarios, now that we've talked about utilizing scenarios. There are a few key issues I think everyone should think about when they're going into generating stochastic scenario sets. The first one is the mean return. I think most of you have an idea what I'm talking about when I say the mean return of a stochastic scenario set, but there's still a lot that can go into understanding exactly what's meant. Volatility of returns is another key one. The benefit of diversification is another. I just want to go into what's tricky about each one.

First I'm going to talk about the mean return of a stochastic scenario set. Often you might hear a quote like this. "We studied the S&P. From 1970 to 2001, the S&P 500 index (net of dividends) had an average return of almost 10 percent." What can we make of that? I actually did this, so I'm not making this part up, but if you look at the actual annual returns—January 1 to January 1—the arithmetic average was 9.75 percent. In other words, I'm taking each year's return, summing them, dividing by n and getting the average. However, the S&P's value at year-end 1970 was about 92. At year-end 2001, the value was 1,148. If I take the return on the index to the 1/ 31st power, I get a geometric average annual return of only 8.5 percent. So that's different by 1.25 percent from the figure I gave you the first time. What causes that difference? Annual volatility in the index creates that phenomenon. Many of you have thought about this countless times, and this is nothing new to you. However, in case there's anybody who really hasn't given this any thought, clearly if you think about just two years, a 0 percent return on an index that, say, keys your annuity product, followed by a 20 percent return or vice versa, only gets you to a total return of 20 percent, whereas two 10 percent returns running gets you to 21 percent (1.1 times 1.1 minus one) return. The effect is more pronounced the higher the volatility, as is the rate of the change.

Try it yourself with negative 10 or negative 20 or 40 or whatever. You'll see that even though the average continues to be 20, the compounding effect produces a situation that's much worse. I think I should have said 10. In fact, a decent approximation for the difference between an arithmetic and geometric average for a scenario set will be sigma squared over 2, where sigma is the annualized volatility. Going back to the S&P 500, I had a difference there of 1.25 percent, and actually I can't work out in my head what that equates to. I think it's about a 16 percent volatility that would tend to produce that much difference, between an annual average or an arithmetic average and a geometric average, which is the long-term one.

What are some important things to realize? Well, first of all you're going to generate a stochastic scenario set. You sit down to do it; you study your data. The mean return that you build into your stochastic scenario set, if you want it to replicate history—be a realistic set of scenarios—you want it to shoot for a mean in your scenario set that's based on the same statistic as the mean that you generated from your historical data. So if you did the 9.75 because you did it arithmetic on the S&P, then you should be shooting for an arithmetic average return of 9.75 on your scenario set, not a geometric average of 9.75, which would inflate results quite a bit.

The return in a levelized-separate-account environment represents the geometric return with regard to fund growth, but generally equates to an arithmetic average with regard to profitability. If you think about this for a second, you can see why. If the distribution of returns each year is lognormal, the possibilities, if you sum all of them for where your M&E charges, say on a variable product, end up, is what gives

you the average profitability over that year. If you work it out using the lognormal distribution, you can see, at least in that case, that it comes out to the average scenario return producing the same as a level scenario rather than the geometric combination of all those. If you do that over an entire scenario set, say for 30 or 50 years, the effect actually is the same. It's the arithmetic average that produces the same profitability as a level scenario.

Next I'll discuss volatility of returns. Generally people talking about volatility in economic scenarios are referring to the standard deviation of the annual returns. That's certainly important for valuing return of premium in ratchet benefits, where big sudden drops cause a lot of the cost. It's less important for roll-ups or benefits with a substantial waiting period. Why is that? Because if you have a substantial waiting period, the effect of volatility tends to mute itself over time, especially if there is mean reversion in the scenario set.

For the latter benefits, some measure of long-term volatility, like the C3 phase two guidance for calibration, may be needed rather than just looking at your annual volatility. The C3 phase two is just a measure of what the Academy considers reasonable tail results for various durations. What are the correlations of returns between successive years? You can have 20 percent volatility on one scenario set and get costs for a GMAB that are much higher than a different scenario set over here with the same volatility and mean if, in this scenario set, you have mean reversion built in. Mean reversion creates a negative correlation of returns between successive years. In other words, the scenario that starts out good will tend to turn bad over time; but more importantly, if a scenario starts out bad, it will have a tendency to recover over time, which means for something like a GMAB, the results will tend to be better than they would be in the first scenario set.

I'm not going to get into theoretical basis for mean reversion, but there was a paper presented recently at the stochastic modeling symposium that tends to indicate that most models don't do well in the long term, even if they do well in the short term, unless there is some mean reversion built into the model. That's as far as replicating historical S&P results. That's kind of interesting.

The third tricky issue is benefits of diversification. Let me give you an example of the basic concept of diversification and what it does for you. Let's say you have a return-of-premium guaranteed minimum death benefit (GMDB). In other words, it's a benefit where, if a person paid $10,000 of premium, they're guaranteed to get that much money back if they die, and this is on a variable annuity. Of the fund value the company has sold, 50 percent has been invested in bonds and 50 percent has been invested in equities, and that's all you know. After a year, equities have gone down 10 percent and bonds have increased 15 percent, net of M&Es and other fees. Are you in trouble on your return-of-premium GMDB? The answer is that it depends. If all of your policyholders are invested 50/50 in bonds and equities, then each one of them has a total return on their fund of 2.5 percent, so your GMDB is okay so far; you're not in the red. However, if half of your policyholders are

invested in 100 percent bonds and half in 100 percent equities, then the bond policyholders are 15 percent to the good on their money, but the equity policyholders are down 10 percent on their 50 percent of your total fund value. In total you're at risk for almost 5 percent of your total fund value. In other words, if the equityholders die, you're going to be paying out money.

That was a pretty simple example, but it just illustrates the point that, in theory, diversification helps you. The more of your policyholders that are mixed between two or more funds, the better off you are. It depends on the mix of funds you offer, though, and how they're held. In the long run you can say that the mean return of a mix of funds, say you have a mix of bonds and equities, will be the weighted average of the funds' mean returns. In the long run the volatility will depend heavily on how the funds are related to each other. Statistically this is the key issue for diversification.

Let me give you the following simple guidance that will at least allow you to get a feel for the volatility of a mix of funds like equities and bonds. If you mix two funds given certain initial weights, in the short run you can say this. If they're uncorrelated, the total variance will be the weighted average of the variances, and the volatility is just the square root of the variance. For funds that are 100 percent positively correlated, the total volatility will be the weighted average of the volatilities. That illustrates the futility of mixing highly correlated funds for variance reduction, which many people do by modeling, say, large cap, small cap and international equities. In good markets, there is some degree to which those are not 100 percent positively correlated. However, studies have shown that in big bear markets, the world tends to all go in the same direction; and large and small cap tend to go in the same direction, in which case you're getting much closer to 100 percent positive correlation. You may find that you haven't reduced your combined variance at all even if people are diversified.

For funds that are 100 percent negatively correlated, the volatility will be the difference in the two volatilities, and that is a very useful result indeed if you can find it. You have two funds both with a mean return of 8 percent. The volatility of the first is 20, and the volatility of the second is 10. If they're 100 percent positively correlated, the volatility of a 50/50 mix will be 15 percent, just as a weighted average. That's not very good. If they're independent or uncorrelated, the volatility will be the sum of the weighted variances to 0.5, or take the square root in other words, which gives you 11.2 percent—a pretty nice savings in volatility. If they're 100 percent negatively correlated, the volatility of a 50/50 mix will be only 5 percent, because whenever one moves one way, the other's automatically going to move the other way. Anybody would love to have that if they could get it. In reality, you don't find that very often in variable annuity funds. I've seen negative correlations as high on an absolute basis as 0.2 and that's about it. Still, even when you can get a negative 0.2, that's a great result if people are diversified between those funds. Short-term volatility will affect long-term considerations if other conditions allow—things like mean reversion and whatnot.

Earlier I said that based on a mix of funds, like 50/50 equities and bonds, you could calculate the weighted average volatility in the short run. When I said that, I didn't mean that the theory changes in the long run, just the weights, because unless you're assuming your policyholders constantly rebalance back to 50/50, your projected fund weights will tend to drift over time, and that may be something your scenarios should be taking into account.

To sum up diversification, there can be substantial savings from including a mix of funds in valuing certain blocks of business, especially variable annuities with guarantees. The value of diversification is highly dependent on your funds actually being diversified within individual policies, not just across the company by a bunch of people with all their eggs in one basket, even if their baskets are all different sizes and shapes. The value of diversification is also dependent on funds being relatively poorly correlated, so testing two similar large-cap-fund styles may provide little value. In other words, to put it in a way that could actually happen, if you somehow develop a scenario set for an indexed fund versus a managed fund, there is some difference that may unfold. If the two are modeled correctly, they'd be very highly correlated, and diversification wouldn't provide you much value.

Statistical applications are part two of my presentation. I'm going to veer a little bit from scenarios, and I'm going to talk about another subject—the use of general linear regression. I think this is probably the statistical tool used and sometimes misused by more actuaries than any other. Why is that? Well, it's common in spreadsheet packages. I know Excel has a really powerful one. It's also easy to remember how it works and what the solution set of parameters means. However, it's fraught with pitfalls. Maybe that sounds a little too scary, but there are pitfalls and they need to be avoided if you're going to get truly meaningful results from a regression. The goal of linear regression, you might remember, is to demonstrate a linear relationship between two or more bodies of data. You want to minimize square differences, estimate co-efficients and generally use the relationship to predict other data points, often ones that lie somewhere out in the future.

I've done a regression example. This is a body of data that dates back to some time in the 1980s, maybe 1983, to show the attendees and the locations of symposiums. It covers the places they've held it on the X-axis and then attendees on the Y-axis. I did a regression. I'll discuss the actual results. The regression line shows a relationship that I want, so that's good. I got an X co-efficient of minus 17. The X co-efficient is important, but you don't know what it means without the t-statistic. The t-statistic was –2.16, with a probability value of 4.57. In other words, at a significance level of .05, I can reject the null hypothesis that there's no relationship and determine that I believe there is a relationship and it's a negative relationship. The lower the degree of latitude goes, the higher the number of attendees gets. It's demonstratively true.
The R-squared statistic is 21.5 percent. The R-squared is how much of the variance in the data is explained by the regression. At first, we had just data points all over

the place. If I shove them all to the left hand side and didn't have latitude and all I had was attendees, there would be this distribution of dots kind of bunched up around 600 and then spread out. What I'm explaining is how much variance there is between those dots that I can explain by knowing latitude. Almost 80 percent of the variance is not explained, but still that's not too terrible of a result. I can quit for the day.

But I thought about it some more. What do I need to chart? I need the data for attendance by calendar year. I think we can detect a stronger relationship there. What happens if I add calendar year to my regression as a second independent variable? Well, I get a co-efficient on calendar year of 33.85 with a t-statistic of 6.13, so that's a probability value on the order of 10^-5, so that's really good. My co-efficient on latitude goes to 0.82 with a t-statistic of 0.15—it's -0.15 standard deviations, in other words. That's very, very poor. That's not the significance level. The significance level would be almost a half. It's so close to zero, it's almost indistinguishable. It's no longer a useful variable.

The R-squared value goes up to 76.5 percent, which is a big jump. I'm explaining over three-quarters of the variance now between the dots just by knowing calendar year and latitude. In fact, this value goes up any time I add a new variable to my regression, but not nearly as much for a poor one. If I had started with calendar year alone, I would have gotten an R-squared that was a little smaller than this but rounded to the same number, if I just looked at one digit past the decimal point. Then I would add latitude, and you wouldn't even notice a difference even though you would have one.

I've given you a regression example where I have a couple of independent variables under my deep-end at very low attendance. What mathematical pitfalls are actually within my example? First of all, it's not the issue already raised—selection of inferior predictive data—in other words, latitude. That's not a mathematical problem; that's just a judgmental/selection problem. In fact, if latitude was all I had to go on, it would still obviously be much better than nothing. If I knew latitude and I didn't know anything else and I needed to predict attendance, I would do better by using latitude than by just guessing 600, because it's about the average. Sometimes you don't have perfect predictive data. Sometimes you have to make do with the best you have, but there are two important flaws that many of you may recall from looking at this stuff. The first is non-constant variance or the outlier effect. The second is extrapolation beyond the range of the data.

Don't forget the outlier effect. The 2001 data point dips way below the line and then in 2002 it's the last data point calendar-year-wise. Who remembers what happened in 2001? We had to reschedule the meeting. Nobody wanted to travel. It's fairly easy to determine the largest squared error is the 2001 data point. Only the 1994 data point is really in the ballpark of how big that squared error was. But is that point truly indicative? As a statistician you have to make that call. This could be biasing your regression, and if you have any other year that you're trying to

predict and there is no huge calamity caused by terrorists, then you might be better served not using that data point to set up your regression. You need to consider that in drawing conclusions, but it's just minor interest. With latitude, that point was pretty close to the regression line. It had relatively small squared error.

The second flaw would be extrapolation beyond the range of data. If you wanted to use a linear formula to predict the attendance of a past symposium between 1984 and 2002 (because you don't know it), you'd clearly prefer to use calendar year. In other words, I ask you what the attendance was in 1990 and you have no access to the data, but you know the regression formulas that I gave you—one for calendar year, one for latitude. The calendar year's clearly going to be the better one to use. Almost every data point was very close to the line, but interestingly despite how things look there, you might or might not find that if you use something for predicting attendance in 2010 or so, it's less useful than latitude. The estimate that you would get from using calendar year is about 1,214. If you think about it, even in 2010, warm temperatures and nice climates are going to be then what they are now; there's going to be no change in what latitude means. Latitude is still going to be within the body of data, but for calendar year we're going to be extrapolating. We're taking data that stops in 2002 and trying to project it eight more years, but future attendance could flatten out. In fact, we know from other symposiums and meetings that that does happen. You get to a point where, unless the amount of actuaries grows, you hit a critical mass and that's about how big the meeting gets. In fact, if we start to get more and more meetings that draw people off, and their company only pays for them to go to two a year, attendance could go down. A parabolic regression—in other words, quadratic—fits the data for calendar year even better.

Starting in 1985, if you think of a bow that levels off around the last few years, you could see how that could fit the data better. I think I got an R-squared of almost 90 percent by using a parabola instead of a line. You don't want to extrapolate and say, "Well, that proves that we're going to lose attendees over the next few years." It's just one possible shape the data could have. On the other hand, if I use latitude to predict that if we have the next one in Quito, Ecuador, I'm likely going to fare worse. That would be extrapolating too, and we want to use this regression to predict attendance somewhere in the continental United States. Even Hawaii may be a bad idea, or maybe everybody would then go.

There is one more often missed mathematical pitfall concerning multiple regression. I've seen many people use regression and just add independent variables rather indiscriminately. Co-linearity of the data is the problem. When you add variables to a regression, it always provides more explanatory power of the data you have as measured by R-squared, but the exception is when you add a variable that's a multiple of a previous variable plus a constant. In other words, for every data point, multiply a previous set that you had by two and add 15 or something like that. If you add a variable like that, it doesn't actually provide anything. It makes the matrix that is used for the multiple regression non-singular, and it means that you

can't solve it. The closer that is to being true for any new variable you add, the less new explanatory power a new variable adds. If you have a variable that's very close to a multiple plus a constant, it will add some explanatory power, but very little. It turns out that calendar year is a fairly good predictor of latitude. The regression between those things is actually much more solid than the one between latitude and attendees.

The crux of the matter is this—despite increasing R-squared or despite increasing the amount of explanatory power you have over your data set, adding non-useful variables to a regression can reduce its predictive power. That's really the key thing. The reason is because in any regression, the number of degrees of freedom you have for your t-statistic is the number of data points minus one, less the number of independent variables.

Most of you remember what the t-distribution looks like. The fewer degrees of freedom you have, the more it starts to bow out and look really funny. Instead of close to two being the number of standard deviations to get to 97.5 percent, it can get way worse than that. It's harder to use it to predict anything. That's what happens to your regression the more independent variables you add. You reduce the number of degrees of freedom. If you add too many, you reduce the power of your test or of your calculation in parameters so much that even the variables that were useful can have confidence limits so wide that they appear non-significant, and the point estimate becomes unreliable. That sounds ridiculous, but I promise, if I added 10 more variables that had no real use, even the calendar year predictor would start to look not significant. It's really weird, but it's true, and it's because of this effect. The bottom line is that extraneous variables muddy the water and you won't know what you really have. You want to keep yourself on a regression to variables that have some meaning.

Let me tell you what you can do all the time. There are a few variables that may be useful for a multiple regression and are never co-linear with your first variable, so they never muddy the water this way. Use higher powers of the independent variable. Remember, the regression is still linear in the co-efficients, even if the data is a higher power of itself. You have A times X plus B times X squared plus constant. You still have linear co-efficients. It's still a linear regression, but you end up with a parabolic equation when you're done or a quadratic equation. Recall that adding X-squared to our calendar year versus attendance actually gave a much better regression than calendar year by itself. Other transformations that you can do on data besides higher powers include the natural log, exponential functions, trigonometric functions and so on—basically anything that's not a multiple plus a constant.

Finally, let's discuss miscellaneous issues. The miscellaneous issue that I've added to this presentation is speeding up run-time on Monte Carlo simulations for XXX testing. I'm going to give credit to Ed Robbins, my father. He in turn would credit a *Transactions* article dating back 20 plus years for this. At some point you may have

a close call requiring extensive numbers of trials. When you're doing your X-factor testing, you may find that, even with your 10,000 policies that you're sampling on and 1,000 Bernoulli trials that you've run, you have a situation where you're not sure whether to accept or reject your X-factor. You're afraid that if you run 10,000 Bernoulli trials, which creates a grid of 100 million ones and zeros, that your run-time is going to become enormous and you're not going to be able to complete your testing in time.

This isn't really a statistical issue; it's a stochastic issue, but it fits neatly into hip-pocket stuff that I just wanted to give you to think about. If you run 10,000 Bernoulli trials past 10,000 policies, you're really creating a grid of 100 million independent ones and zeros. If you think about this as something existing in an Excel spreadsheet, even though you wouldn't really do it that way, it will help you picture it. You put the 10,000 ones and zeros to the side of each of your policies, and you run random numbers where a zero equals life for the testing period that you're running over and a one equals a death. Then you sum up all your policies and face amounts and you get total dollars of loss for that period.

That's the typical methodology that most companies use. The problem with that methodology is the way that they fill in the grid, because that's not the only way you can do it. In other words, having each of those cells or each iteration of your program create one random number so that you have a total of 100 million random numbers is not the only way to do it. Instead, if you have any sequence of independent trials with equal chance "p" of success, there's actually a probability distribution for the trial number where you get your first success. You can actually work it out by hand. If the trials are independent, what you need is one minus "p," which is the probability of failure, and minus one times, and then a success on the next trial, so that would be your "p." That's why I say you can work it out by hand. For instance, if a chance of success in any trial is 1 percent, the chance that your first success is on the first trial is also 1 percent. That's obvious, but if you want the chance that it's on the second trial, you need one failure, 0.99, and then a success, 1 percent, so the probability is 0.99 percent. The chance that it's on the third trial is 1 percent times 0.99 squared and so on.

If your "p" above is the expected mortality rate for a cell in certification testing, you can create a cumulative distribution function for that cell. When I say cell, I mean the one life with whatever its face amount is. The cumulative distribution is that of the geometric distribution. All I then have to do is run this test that I've put together across the row of that cell drawing a random number comparing to the cumulative distribution function and getting a numerical result. I then fill in "n-1" zeros into that row, and then a one, and then I start over. I draw a new random number and so on until that row is complete.

Clearly, if I get to the end of a row, since the mortality rate's probably going to change on the next life, I have to start over. In other words, out of my 10,000 cells going across from my first one, I get to the $9{,}980^{th}$ cell and have a success, so I

have 20 cells left in my hypothetical spreadsheet. If the next sample that I draw from the cumulative distribution function gives me a result of 80, I just fill in 20 zeros and then stop. Then I start over on the next row where my mortality or my probability of success is different.

The end result is an expected reduction in run-time, approximately in proportion to your average expected mortality rate. In other words, if over all of your 10,000 lives, your average mortality is about 1 percent, then your run-time should be reduced by 99 percent, or it should be reduced to 1 percent of whatever it would have been. That reduction could be offset a bit by added programming time on the front end, but if you're going to do this a lot, it could be well worth it. Once your grid is filled in using this methodology, the results are tallied down the columns the way they were before, and the stochastic implications are identical as to what they would have been if you would have filled in the grid one single cell at a time.

In summary of all we've talked about going back to the normality thing, normal is somewhat normal, but you have to be careful about approximating a distribution this way. It's somewhat normal in that for all valuation actuary types, if we're going to look at a distribution and want to look at the mean result, then we can certainly approximate it that way. If a distribution is the sum of any IID random variables, we can approximate it this way. When a distribution is not normal, statistics from a sample (where a sample could be something like a bunch of random scenarios) could still be used to make inferences on tail percentiles. When generating stochastic scenarios—this is our statistical application one—it's good to be clear about what the different scenario statistics really mean, as well as the benefits of diversification. Linear regression is a very useful tool, and I encourage anyone to use it. It's important to realize that increasing R-squared is not the only goal. Finally, testing of X factors might be able to be sped up.