2004 Valuation Actuary Symposium*

Boston, MA
September 20-21, 2004

Session 35TS
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 appointed 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 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 analysis.

MR. DOUGLAS L. ROBBINS: Welcome to Session 35, a teaching session on drawing appropriate statistical inferences. I am the only speaker, the teacher so to speak. My name is Doug Robbins. I've been a consultant at Tillinghast-Towers Perrin for almost 10 years. Before that I had just completed a master's degree in statistics at Georgia Tech, and through the work I've done at Tillinghast, I've managed to hold on to some of that. A couple of years ago, as a member of the planning committee for the Valuation Actuary Symposium, I volunteered to teach a little session to refresh other people's memory on how to use statistics appropriately in their actuarial work. I've been doing it now for a couple of years. This may be the last time I do it. I've tried to keep the presentation current, so I've added some things toward the middle and end that are topical right now, but I'll start it out about where I started out the past couple of years. Hopefully most of you are new and haven't attended this session before, or you'll see some familiar material at the beginning.

I've entitled my session "How Normal *is* Normal, and What if it's Not?" That, of course, refers to the normal distribution, and that's where we head first. I have put together this entire presentation, and this is my promise to you, without ever

*Copyright © 2004, Society of Actuaries

cracking a textbook. All of this material is as I remember it, and I did it that way on purpose to keep myself from getting too technical. I've tried to put together points that are easy to remember, easy to follow and easy to use without your having to go back and crack a textbook unless you want to get more deeply into it.

I've put this together in five basic categories. "How normal is normal" is the first one, "what if it's not" is the second, a couple of statistical applications that I think are interesting to talk about comprise the third and fourth parts, and I'll finish with some miscellaneous issues and a summary.

First, how normal is normal? In this session I'll be covering a couple of topics, at least in the first couple of sections: "how normal is normal" and the bell curve, and "what if it's not" and the non-bell curve, which is the distribution that we're getting more used to seeing in some of our actuarial work.

How was the normal distribution taught when you were in school? Do you remember it? If you took a cookbook statistics class, here's how it was taught. You were told some anecdotal arguments about the weather and about traffic and told that many things in life have roughly a bell shape. You probably took a quick look at the probability density function of the normal distribution, which also has roughly a bell shape, and then had an introduction to a neat table at the back of the book where you could draw stochastic inferences of anything with roughly a bell shape on the assumption that it was normal.

I'll put forth a few things. Consider the following: For how many of these distributions that I'll show you would you feel confident estimating the 99th percentile using that normal table? First is the number of accidents on I-95. That comes up to Boston from Connecticut and Rhode Island and moves around, but it's a typical freeway in a large city. What would you say in a morning rush hour, assuming the mean is two? What about the result of a single, randomly rolled six-sided die, or the number of days late a given Valuation Actuary Symposium session will be filled with speakers? Maybe normal is not so normal after all.

Here's how you might do it. If the true number of accidents is two, you may think about it a little bit and say that maybe this is Poisson because you remember something about a discrete number of incidents in a given discrete period of time. Let's assume the mean is two and therefore the variance is two. You'd use the normal tables and estimate a bit over five, but the true 99th percentile, if it is Poisson, is just over six, so you've undershot by about 20 percent of the amount of your guess, which is not too good.

The mean of a roll of a single die is 3.5, and you can work out that the standard deviation is about 1.7. If we shoot for the 99th percentile using the normal approximation—2.33 sigmas—we get 7.5, and the die has only numbers one through six, so there we've overshot quite badly.

On the other hand, although the average lateness of Valuation Actuary Symposium sessions might be one week, the 99th percentile of the lateness of one given session is an estimation you would not want to make based on the whole body of data. Most sessions are filled on time, and a few have been filled this week, so it depends. If you do a little study into which sessions have which recruiters, you'd know which is which. I'm to blame this time, I think, but at any rate, that's a little lesson in bipolar data. Sometimes there is no good way to estimate them.

Generally when is it a bad idea to use a normal approximation to make distribution assumptions? Most always, when the data are somewhat scarce or bipolar, and much of the time, when you're making tail inferences on an underlying distribution. Sometimes you can get this from the sample data's shape, but not always. Even some distributions that appear bell-shaped can have tail features that are amazingly different from normal. Has anyone heard of the Cauchy distribution?

Some of you may have looked at the Cauchy distribution in college. It would be the distribution of light where you have a lamp sitting in one place, and there's a flat wall in front of the lamp. The light hits at various points on the wall. You're going to have a lot of light that hits near dead center from where the lamp is, but you're going to have some rays that are bent further out toward the parallel that hit quite a distance from the center. What happens is that you get a distribution that is based on things like trigonometric functions because you're dealing with angles. The interesting thing about it is that there's no mean and no standard deviation of this distribution because it's so widely disbursed.

How many of you have ever, if you were bored, sat at your desk and done the NORMDIST function in Excel with a random number and recalculated it over and over to see how big you could get the number to come out? I've done it before, and sometimes you'll get a three and think that's pretty big. I don't know if Excel does this for you, but if you do it with Cauchy, you'll hit 1, -2, 2.5, 10 zillion, 3 and -1. You'll get wild results. What that does is make the distribution so disbursed that you never converge to a mean because every time you get one of those 10 zillions, it throws the whole thing off, and it does it ad infinitum. That's interesting. All that is to say that you can have a bell-shaped distribution and could try to do some things with normal, and you'd be way off on the tails.

When is it okay to assume normality in a distribution? Under typically sound conditions (meaning it's got to have a finite mean and variance), the sum of a large number of independent, identically distributed random variables will converge to normal. That's the central limit, which you probably remember. You probably also remember the abbreviation "IID." That's the abbreviation for variables that are independent of each other and distributed exactly the same way. So those are the conditions you need. For instance, if you're rolling two dice and work out the distribution, which becomes a little more bell-shaped, although very linear-looking, you get an estimate of about 12.6 for the 99th percentile using the normal approximation. The real answer is more like 12, but you're down to an error of only

5 percent, which is a lot better. The more dice you put into the mix, which are all distributed independently, the less error you would have.

There also are certain types of distributions if one or more parameters are large enough: Poisson, the sum of Poisson, gamma is the sum of alpha (I think), and for other reasons lognormal works this way. If you have a large enough parameter, all of a sudden the distribution itself becomes the sum of a distribution and converges to normal. If you looked at the whole state of Massachusetts, and the Poisson mean was 300, the approximation would give you 340, which is just about dead on.

In the valuation actuary world, when is it absolutely correct to assume normality? The main time is when you're looking at the mean of a sample of scenarios. Maybe I shouldn't say "scenarios" because you might have other uses, but let's say we all work with a lot of scenarios now in the valuation actuarial world. I think we do, so let's say "scenarios" for now. As long as there are a reasonable amount (maybe over 40 or perhaps more if your distribution is skewed) and you met your boundary conditions that I mentioned earlier (as long as your mean and standard deviation are finite), you get a situation where the sampling error and the sample mean can be used to construct a normal distribution once the sample gets high enough. It's marked out as a T distribution, if you remember, but by the time you're up to 40 or more, it's so close to normal you can't tell the difference. That's one time. Central limit theorem says you can do that.

That's the key time for valuation actuaries. When you're looking at such a mean, which is distributed normally, it's somewhat irrelevant whether the scenario results themselves look like they're normally distributed. I won't say it's totally irrelevant because it affects how many scenarios you need (I mentioned that a second ago), but it doesn't affect the fact that it converges to a normal distribution over time the bigger the mean gets. It does that because each scenario is distributed some way, and running more of those scenarios as long as they're distributed identically creates this IID condition (see Chart 1).

I want to put out an example: the stochastically derived costs of a guaranteed minimum accumulation benefit (GMAB) after 10 years. GMAB is a guaranteed return of premium on a variable product. No matter how your funds perform, you can have equity guarantees. Therefore the funds could fall, and you guarantee that people get their premium back. The cost of providing that benefit over 100 or 1,000 is going to look like that green line at the bottom where most of the results are zero, but a few of them have a significant cost. It's heavy-tailed. Even though the shape of the data makes it clear that the distribution is nothing like normal, I know that my central limit theorem works.
Why do I know that? My cost is bounded between zero and the premium that's paid. You can't lose more than the premiums. If the policyholder paid $10,000, the most he can possibly lose is $10,000. If that's true, any distribution that's bounded and not truly infinite must have all finite moments. Therefore this has to apply under the central limit theorem.

The practical conclusion for all of us as valuation actuaries is that we can assume in most testing we do that our scenario mean is distributed normally. You'd have to work out an unusual example to where your possible losses were infinite. Maybe not in the casualty actuary world, but I think for us it's going to be unusual. You can assume that your scenario mean is distributed normally.

What does that mean we can do? One thing it doesn't mean is that the scenario mean plus two standard deviations gives you the 97.5 percentile of the underlying distribution. That's not what we're saying. If you go back to this GMAB distribution, it wouldn't be true that you could take the mean plus two standard deviations and work out the 97.5 percentile of the cost. The inference is on the mean, not the distribution.

What it does mean is that you can draw symmetric or non-symmetric confidence intervals around any scenario set mean using that mean and its standard error. The standard error of the mean is the standard deviation divided by about the square root of a number of scenarios. I think if you want it to be unbiased, it's got to be the number of scenarios minus one, or the standard error has to have the minus one in it. Again, I didn't look at any textbooks. It's one of those two, but at any rate, when you're talking about a number like 40, 100 or 1,000 scenarios, the minus one doesn't mean a whole lot.

In our GMAB examples, let's say I ran 100 scenarios, which nowadays people would say is not nearly enough, but I didn't want to redo my example. We've got 90 scenarios out of our 100 that produce no cost, but because we invested in equities, five of our scenarios cost $5 a piece, three scenarios cost $20 a piece, one scenario cost $41 and one cost $74. The mean cost is $2. You can work that out for yourselves if you like. For this sample variance, the calculation is $81, so the standard deviation is $9. I divide by the square root of 100 and get a standard error of my means of 90 cents, so I've got a mean of $2 and a standard error of the mean of 90 cents. That becomes the piece of it that's distributed approximately normally. Even here, because this benefit is so skewed, the approximation isn't that great (I'll show you why in a second), but it's reasonably good. It's much better than the single die or the two accidents on I-95.

What I can say if I'm going to use the normal distribution is that I'm 95 percent confident that my true population mean for this GMAB is between 20 cents and $3.80, plus or minus two standard errors. Does that imply that if I run 100 new scenarios with the same scenario distribution that I should be within 20 cents and $3.80, 95 percent of the time? No, it doesn't imply that, because each sample has got its own sample error, and the two sample errors might compound. The inference is on the actual population mean only, so whatever that population mean is, I'd be 95 percent confident that it would be within my new confidence interval if I ran another 100 scenarios.

Note two things about the mean plus three standard deviations. One is that it only gets you to almost $29, and 2 percent of my sample, two scenarios, are well above that point estimate. You can see that the underlying distribution is extremely non-normal. Does everyone understand what I'm saying? I'm saying, "What if I was going to try to draw an inference on the population distribution, not the mean?" If I tried to do that, I'd still be way off.

The other thing about how skewed it is that you might notice is that my 95 percent confidence interval is almost down to zero, and if I tried to do a 99 percent confidence interval, it would go to a negative number. That's a sign that you're not to perfect normalness yet. If it were truly normal, you'd be far enough from zero so that three or four standard deviations would not get you to zero because the true shape of normal is infinite in both directions.

I've got this distribution and I want to make inferences on the tail because I'm a valuation actuary, and I'm not so concerned about the mean cost, although I'm interested. I'm also concerned about the tail cost, and I'm saying I can't do it because it doesn't work to assume normalcy.

That brings us to "what if it's not." What can I do if I've got a distribution, want to make a valid statistical inference on the tail and can't do it based on the normal approximation? The answer lies in the realm of non-parametric stats, and we'll explore that shortly. First I want to tell you about an idea I have to get rich, because I want to ease off on my job even more than I have been and kick back. Of course, I need to find a way to get a billion dollars. Let me tell you about this game I invented called the E-lottery.

I get together a huge crowd of people, perhaps a billion people. I randomly pick and record a number between one and a billion. I get some trusted entity to certify and record my pick. Each of the people puts in a dollar and then without conspiracy — it's important that they can't all get together and make sure one of them picks one of my numbers—each one has to independently pick a number between one and a billion. If there's one winner, he gets the pot. Multiple winners split the pot, but if there are no winners, I keep the billion dollars. Who wants to have a guess at my odds of keeping the billion dollars? Let's work it out.

To keep the money and retire, I need everyone to guess wrong. The odds of one person guessing wrong are [(n-1)/n] because that one person has to pick a billion minus one over a billion of the possible number. He's got to miss the one number. If all are independent, which they are because they can't conspire, then the odds of all n of them guessing wrong are [1-1/n]^n, and I happen to remember that as n gets large, this approaches 1/e. I've got a 1/e shot of keeping the billion dollars. It's a pretty decent chance.

What's the point? It turns out that roughly the same odds apply to the underlying 99th percentile in my 100-scenario GMAB example. They aren't exactly the same odds because it's only 100 and not a billion, but the odds are roughly the same. The chance of any individual scenario being less than the true 99th percentile is clearly 99 percent. If all 100 scenarios are independent, the chance of all 100 of them being less than the 99th percentile is 0.99^100. The approach to 1/e is from the underside, not the high side, which is a good thing. The odds are less than 1/e.

I've got less than a 37 percent chance that all of my scenarios are less than the actual 99th percentile of my underlying distribution, or to put it another way, this means that I have a 63 percent chance that my worst scenario is worse than the 99th percentile of the underlying distribution. Even with just 100 scenarios that's true. Almost two-thirds of the time, my worst scenario is going to be worse than my 99th percentile. That gives us the ability to draw a statistical conclusion about the true 99th percentile, which is that we're 63 percent confident that it's less than the worst scenario.

What if I'm doing some testing where I care more about the 95th percentile? Now my odds go from 1/e to (1/e)^5. Now I'm 99.5 percent confident. If I only cared about the 90th percentile and ran 100 scenarios, I need to widen my cell to put it that way to catch where it turns from zero into a positive number regarding the odds that possibly a scenario might not be worse than the 90th percentile.

What if I do care about the 99th percentile? If that's the case, 63 percent certain is pretty poor, right? Nobody wants to be 63 percent confident that his worst scenario is worse than the percentile he cares about. Then you just need to run more scenarios. If you run 1,000 scenarios, you're back to (1/e)^10. You're back to being almost certain that your worst scenario limits the 99th percentile.

Do you always have to use the worst scenario? No. It's just the simplest case. I think by doing that I give you something easy to remember, something you don't even have to look up. You'll remember, "If I run 100 scenarios, my 99th percentile is 1/e that I'm worse. If I run 1,000, it's (1/e)^10." That's easy, but a lot of times you're not going to care about that, and a lot of times your worst scenario is not going to be good enough. This works only if you're happy with your worst scenario. It happens sometimes, but not always. This gives you something to keep in your back pocket.

Any time you're looking at a percentile and running a scenario, you're just setting up a Bernoulli trial. At the 95th percentile, you've got a 5 percent chance that a scenario is bigger and a 95 percent chance that it's not. For any percentile of the underlying distribution and any ranked scenario, you can work out the odds using binomial theory. You can use this methodology basically to construct any size confidence interval on a percentile that you might want. If you've got completed scenario results, you can test the stochastic hypothesis. If you remember hypothesis testing, you'll remember that it's always closely akin to setting up a confidence interval. In this case it's similar to the one we're used to with the mean plus two standard deviations, but you use non-parametrics instead. Here's an example.

Suppose I've got 1,000 scenarios and want to make an inference on the total surplus of my company. I want my surplus level that I'm testing to "pass" 99 percent of all possible scenarios. I run my 1,000 scenarios, and five fail or have negative ending surplus. If 10 of my scenarios failed, my point estimate would be that I am passing 99 percent of all possible scenarios. You might think that only failing five is easily good enough. If you set up your null hypothesis, it's that your *true* 99th percentile result (not the 99th percentile of the sample, which is scenario 10) would fail. If that's true, the chance of any success could be no more than 99 percent. It could be worse, but it could be no more. Therefore, a chance of a failure could be no less than 1 percent.

Under the null hypothesis, the highest possible chance of zero failures would be 0.99^1000, or 0.00004. The chance of two failures would be a lot greater at 0.0022, and so on. If I add up my probabilities for five or fewer failures, I get 6.61 percent, which is a high chance. Often we test at a 5 percent significance level. Here, we could not reject our null hypothesis, therefore we could not conclude that the 99th percentile result is a failure, at least at the 5 percent significance level. That's even though our point estimate could pass scenario 10 and could be passing by a wide margin. The tail could be so steep beyond that that you still couldn't conclude that you pass.

This is like the situation where you're testing a mean, and your sample mean is on the right side of the hypothesis you're testing, so it's better than what you want, but not by enough standard error—not by two, for example. In that case, often what you do is run your hypothesis test with more samples to narrow your standard deviation and therefore narrow your confidence interval. Maybe now you can reject, and maybe you can't. Similarly, if we had had four or fewer failures, we could have rejected, but because we had five, we can't. The solution isn't necessarily that we need to increase surplus. It could be just to run more scenarios. If we run 10,000 scenarios, and the failure rate is 60, which is worse than 5 in 1,000, we could still reject easily because the increased number of scenarios has narrowed our confidence interval.

Let me give you some final notes on confidence statements. In this section we've made several of those. They were based on scenarios. They were based on running 100, 1,000 or 10,000 scenarios. In other words, they assume all your other model assumptions are correct. Actuaries tend to do that a lot. We tend to assume that we've got a good handle on decrements and what those will be, such as mortality and lapse. We know what our expenses are going to be, but we don't have any idea what the economic conditions are going to be.

That could be true, but you've got to be aware of the basis for your confidence statement, which is, given all our other model assumptions are true, we're 90 percent confident. You want to be aware of that basis and may want to sensitivity test your other assumptions if you're not going to make them stochastic. That concludes our "how normal is normal" and "what if it's not" sections of our lesson today.

That brings us to the first statistical application I'm going to talk about. I'm going to shift gears a little bit and talk to you about the use of general linear regression. Why? I think it's probably the statistical tool that's used and occasionally misused by more actuaries than any other. Out of all the statistical tools you get in a program such as Excel, I think it's the most popular. One reason is it's common. Almost every spreadsheet packet I've ever seen has it. Another reason is it's easy to remember how it works. I'm trying to take my X's and get a linear relationship between those and my Y, which is some variable I'm trying to predict. It's easy to remember what the solution set of parameters means, too. You've got your alpha and then your B1, B2, B3 or whatever. It's also fraught with pitfalls that need to be avoided if you're going to get truly meaningful results. We're going to work through an example and talk about those.

The goal of linear regression is to get a linear relationship worked out between your independent and your dependent variables by minimizing squared error or squared differences. You take that, estimate coefficients that you're going to use to put your line together and then you generally want to use the relationship to predict other data points, sometimes future ones, sometimes interior ones. If it's future data points, sometimes it becomes more of an econometrics problem or forecasting. But still, often linear regression is used for that, and I've seen it used even at my company.

I've put together an example. You notice I have a heavy sweater on for September because I'm in Boston, and it's like not August or July. I don't like the cold, so I'm hoping that the Valuation Actuary Symposium people who decide the future sites read my speech, and I've put together an example of a regression of valuation actuary attendance against latitude. I'm trying to demonstrate to them that they get better attendance if they put it farther south. The meeting two years ago was at Disney World, and that was well-attended, and last year it was in San Diego. It was working, and this year the SOA put it in Boston, so we'll see how this goes.

You can see the linear relationship in Chart 2, although you can also see that it's not intense. I've drawn the red line through my data, with latitude getting bigger as you go further north and attendance getting bigger as you go further south in latitude. It's real. It's based on the data of the Valuation Actuary Symposium since it started back in the 1980s. My X-coefficient is -17, and my T-statistic is -2.16 with a probability value of under 5 percent. At a 5 percent significance level, I've got a meaningful regression here. My R-squared statistic is 21.5, and we all remember that R-squared measures the amount of explanation we've done at the error. If we had error worth 100, 21.5 of that error has been explained by my regression. As I said before, the relationship is not intense, but you're explaining something. It's better than nothing, and it is significant. I could quit for the day, go home and be satisfied.

What do you think Chart 3 shows? It's the same thing. It's Valuation Actuary Symposium attendance against calendar year. The relationship looks better. I've drawn the red line through the data in Chart 4. That's disappointing because that might show that my other regression is coincidental, or it might not. We'll see. We'll talk about it.

What happens if I start with my first regression but then don't change and just add calendar year to the regression as a second independent variable? Now I've got X-1 and X-2 trying to predict Y. I get a coefficient on calendar year of 33.85 with a T-statistic of 6.13, which is huge. Think of the T-statistic in terms of standard errors with normal. You remember that 2 is a lot, so 6 is almost unheard of. My coefficient on latitude goes to 0.82 with a T-statistic of almost zero, so it's no longer a useful variable. My R-squared value jumps up to 76.5, and that's important, but it's also important for you to realize that any time you add another X variable, that's going to happen to some extent.

The fact that it's a big jump probably does mean that calendar year is meaningful. In fact, if I had done the opposite and started with calendar year, the number would have shown up as 76.5. When I added latitude, it would have gone up but by less than 0.05 or whatever is half of that. In other words, it would round to the same number, even though it is going up. It's true, and it's important for you to understand that any time I add another X variable, with one exception, I'm going to make my R-squared bigger because I'm going to explain the data a little bit better.

What mathematical pitfalls are within this example? First of all, the issue I already raised, which is selection of inferior predictive data, is not a mathematical error. That's a judgmental error, perhaps, or maybe it's not an error at all. Maybe it's all you've got. If I didn't have calendar year data and all I knew were the latitudes, it would be better if I was trying to predict attendance than nothing. There's something there. It's important to remember, though, that correlation and causation are not the same things. If I were trying to pick a Valuation Actuary Symposium out of a hat and guess the attendance, it would help me some to know latitude. That's a fact. There's no disputing it. If I didn't know calendar year, I would be better off using latitude than nothing.

There are two important mathematical pitfalls that many of you probably recall about regression. The first one is non-constant variance, or the outlier effect. The second is extrapolation beyond the range of data. We didn't do the extrapolation one, but we could do it based on these data. I'll show you how. The outlier effect is there. Recall the 2001 data point on the most recent chart. It's where the big dip is. It's easily the biggest squared. It's not even close. It's easily the largest squared error of the set. The year 1994 was in the ballpark, but less than half. Is that point truly indicative?

Who remembers what happened to the Valuation Actuary Symposium in 2001? It got moved to Florida because of the World Trade Center incident, so a lot of people that were planning to go had to cancel their plane tickets, companies probably got frustrated and people didn't even want to travel by December or late November when it took place. It's not a great data point, and you need to consider this in drawing conclusions. If I do include it, it's definitely going into my regression. It's having an impact, and you have to decide if that impact is warranted if you're trying to draw conclusions. That point had a relatively small squared error when you did latitude.

The second issue is extrapolation beyond the range of data. If I wanted to use a linear formula to predict the attendance of a past symposium between 1984 and 2002, I would want to use calendar year. It's better. You can see that from the T-statistic. Interestingly, despite how things look, you could find that predicting attendance in 2010 could be better done with latitude. Why is that?

Even in 2010, warm temperatures will be what they are now. It's warmer in Florida. It's going to be warmer in Florida. Future attendance could flatten out. There could be a pattern emerging that hasn't fully come through yet in terms of calendar year, but you can see that possibly we've reached the pinnacle of where we're going to get with valuation actuaries, and the data are going to flatten out. You're not getting that by looking at these data and trying to predict ahead that far. It's true that if I try to fit a parabola by using calendar year and calendar year squared, it fits even better than calendar year by itself, so that would indicate a bowed shape. On the other hand, using latitude to predict more than 1,200 attendees in Quito, Ecuador, even if it was next year, would be much worse because none of your companies would pay for you to go there.

There's one more mathematical pitfall possible concerning multiple linear regression, and this one is often missed. I've seen it often missed in work that people have done, and it's colinearity of data. Typically, and I said this almost always happens, adding a new variable to regression provides more explanatory power. I want to emphasize explanatory power as measured by R-squared. Any new line that you add, as long as it's not basically parallel to one of the other lines, will help you at least a little bit to explain the Y data that you've already got. That's a fact.

The one exception would be adding a variable that's a multiple of a previous variable plus a constant. If you think of it in only two dimensions, that would be basically another line that's parallel to the line you've already got. The closer that is to being true, the less new explanatory power a new variable adds. If the lines are not colinear but almost colinear, you're going to get little new explanatory power.

It turns out that calendar year is a fairly good predictor of latitude. In other words, they are almost colinear. A regression between the two finds a relationship even at the 2 percent significance level. They're very colinear. What that means is the SOA has already been going my way on this. I think the first Valuation Actuary Symposiums were all in Chicago, and it's had some in D.C. and Boston, but there have been more in Florida and the southwest. The crux of the matter is that despite increasing R-squared, which increases explanatory power, adding non-useful variables or colinear variables to a regression can reduce its predictive power, and that's the key.

The number of degrees of freedom that your regression has is the number of data points minus one, minus the number of independent variables. Each time you add an independent variable, you reduce your degrees of freedom. Each time you reduce your degrees of freedom, you make the design or the experiment more wobbly or more variable. If you add too many, you reduce the power of your test so much that even useful variables have confidence limits so wide that they don't appear useful. The point estimate becomes unreliable, so unreliable that you can't say for sure whether that's a relationship or not.

Even in this calendar year example, if I added four or five more variables that were more or less colinear, even though I used to have a T-statistic over 6, I could get to where my T-statistic is meaningless, and that's scary. That shows you that you're doing the experiment wrong because there's an obvious conclusion that you're missing. The bottom line is that extraneous variables muddy the water, so you don't want them there because you won't know what you've really got.

Sometimes you may want a good, strong, explanatory model, though. You might want to increase the number of variables and not know how. I'm going to give you a few variables that may be useful to explain things and that are never colinear with your first variable.

The first—and I think the most important—is higher powers. You remember looking at my graph against calendar year and seeing a possibility of an X and an X^2 being a better fit than X by itself. That was true. Any higher power is always non-colinear with the first because the regression is still linear in the coefficients, even though the X variables are not linear any more. Regarding other possible transformations, natural log is one, as are exponential functions, trigonometric functions and so on. Hopefully that's a little bit you can keep in your hip pocket about regression. The key item to remember because it's the one that's missed the most often would be the colinearity.

In statistical application Part II, I'm adding material, as I alluded to at the beginning of the session, to try to be topical. Again, I didn't look at a textbook at all, but some of what I've learned in the past that has gone into this I will credit to the stochastic modeling symposium held last year around this time. The application is statistical considerations under C3 Phase II. I'll be using the paper that was presented at that symposium that I thought was done well when we get to Conditional Tail Expectation (CTE) 90.

Is anyone not familiar with C3 Phase II? I'll explain quickly. Up until recently, variable products, or at least variable annuity products, have had low required capital on a regulatory basis. When all you had was a product that has been sold, the commission has been paid, and you are going to charge your mortality and expense, basically it equates to a spread on a fixed product, if some of you don't work with variable. It's a charge that you apply to the account value, and it's where you get your revenue as a company. You're just charging that and getting expenses. There's no C3 risk in that. There's no risk that your assets are not going to be sufficient to cover your liabilities, so required capital was low.

However, on variable annuities, first we added, as an industry, death benefits, followed by guaranteed death benefits, enhanced guaranteed death benefits with roll-ups and ratchets, and then living benefits. It got to the point where regulators saw a need to establish higher capital levels. The living benefit requirement lately has been onerous.

Can I just take the scenario set I have—my real-world scenarios—and stretch it out? In other words, can I take what I've generated and then pull my tail scenarios down in some reasonable fashion and make those worse until they calibrate? I've read the verbiage in the guidance, and it doesn't seem to forbid that approach, but I think as valuation actuaries, you all need to decide whether that meets the spirit of the guidance or not.

Another question is, how about other asset classes besides equities? The guidance implies that you can come up with your own calibrations, but be consistent. There are probably several variations on that. Some might think that for bond classes the volatility historically is weak enough that you don't need to do anything to your real-world generators; other people might think that you have to do something to be consistent with the fact that for equities you have to do something.

Let's go on to the second statistical consideration. That was more theory and an attempt to provide you with some insight, but on CTE 90, I'm going to talk specifics again. The CTE 90 is the average of the worst 10 percent of possible outcomes estimated using stochastic scenarios. Once as valuation actuaries we might have been happy with value-at-risk measure, like a percentile for required capital. I've talked earlier about how you might not want to estimate that with a point estimate. You might want to make sure you're conservative enough that you cover the 95th percentile with reasonable confidence, but now, with heavy-tailed benefits, the industry seems to need a more robust measure than just a value-at-risk measure.

There are benefits that might not have any surplus at all at the 95th percentile. In other words, that could have a major impact deeper in the tail, so major that they wipe out everything from 90 percent to 95 percent. If that's true, the CTE 90 is going to pick that up, and the 95th percentile wouldn't pick it up at all. By taking all of our 10 percent of worst outcomes into account, CTE 90 helps us account for the heavy-tailed risk.

What are our basic statistical considerations for CTE 90? First it's important to note that we're estimating two things in a CTE 90. It might not be apparent at first. You might think we're just estimating the expected value given that the scenario is higher than the 90th percentile, and it's true that you are doing that, but first you're estimating the 90th percentile. We pointed out earlier that your point estimate is not right by definition. It could be. It's continuous, but there's a very small chance that it's right. It's probably too high or too low.

Second, given that the 90th percentile is right, you're estimating the conditional distribution of losses. This makes the variance of a CTE more volatile than that of a typical main result. More important, if I say that this is my estimated CTE 90, given my scenarios, it's only asymptotically unbiased. What do I mean by that? I mean it's unbiased only if you run an infinite number of scenarios. If you run less than infinity, you not only have statistical error, which could go either way, but you've got a bias in one direction only. The worst thing about that is that the bias is unconservative. In other words, it's downward-biased.

Let's talk about why that is. For a correct treatment of order statistics, if you look at the 90th percentile and estimate it according to the textbooks, you always get an unbiased percentile estimated for the 90th percentile, although there's statistical error in it. However, combining the two estimators produces bias in small samples, and it's downward, which is unconservative. I'll tell you why it is. If you're looking at order statistics and trying to estimate percentiles, your worst estimates are always your worst scenario and your best scenario, especially if you've got the distribution with one heavy tail, which you do a lot when you're valuing these guaranteed benefits. You don't care about your good tail. It's not even really tailish. It ends where all your scenarios are profitable.

On your bad side you've got this estimate that's out there somewhere, but the true distribution, although it's not infinite (we talked about that at the beginning) is so close to infinite it might as well be, and if I run 100 scenarios, the real worst possible outcome is always worse. If I run 1,000 scenarios, the real worst outcome is always worse, although I might be closer. If I run a million scenarios, the real worst outcome is still always worse. I'm always going to get closer, but it's never close enough. It's never close enough to be unbiased unless I run an infinite number.

When I'm estimating the mean of my scenarios, the actual mean around the middle, I've got this bias, but I've got it at both ends, so that's okay. When I estimate the CTE, I've got the bias only at the top end, so that's why my CTE bias is always downward, but it decreases as sample size increases. The closer I get to infinity, the better off I am. There's also bias in the standard error estimate, and I think it's also unconservative, but it's probably not a big deal. It's a lot less important. Using the standard error estimate, you can come about it the same way, although its properties are different. You can form a confidence interval on your CTE given a large enough sample size. Remember central limit theorem? We need to be a little conservative about sample size, both to get our bias down to where it's less important than our statistical error and because our statistical error is fairly large. It's a skewed distribution. It's much more skewed than just your distribution if you're testing a variable annuity guarantee because you're looking just at the end, where all the skewness is.

If you're estimating the CTE of a nearly normal event, you might get by with a relatively small sample, but most of the distributions you're dealing with are so fat-tailed that I would say you don't need to be thinking about having your actual sample that you're using be 40. You need to get it to at least 100, and so for that reason I would say your big end, the number of scenarios you run, should always be at least 1,000. I'm not a person who always says that. I sometimes argue with other people in my company and tell them that for a lot of applications they're doing, they need only 100 scenarios, so stop overdoing it. In this case, you really need 1,000 because of the bias and the extreme skewness.

I'm going to cover a couple of miscellaneous issues and then summarize. This is another application. It's part statistics and part stochastic, but I think it's still important. A lot of you may not have heard of it before. It's a method of speeding up run time on Monte Carlo simulations for XXX testing. I'll credit it again to my dad. He credited a *Transactions* article that's probably more than 20 years old. Is everyone familiar with X factor testing under XXX?

When you set X factors, which are basically applied to your valuation mortality to try to minimize deficiency reserves under XXX, you have to test those statistically each year. That's where the statistics come in. You have to test them using some sort of analytical tool, although it doesn't have to be a confidence test or a hypothesis test. You have to use analytical tools to justify the levels and show that it's reasonable to have your X factors low enough. I'm not going to go into all the rules, but there are multipliers that are less than 1 that are applied to valuation mortality that lower your valuation net premiums for the purposes of comparing to your actual premiums. You've got to justify analytically that they're correct or that they're not unreasonable.

At some point you may have a close call requiring extensive numbers of trials. Let's say we're in the situation where we feel that for the 10,000 policies that we have, because we're testing losses in terms of face amount on a gross basis, we have to do Bernoulli trials to get our distribution, and we have 10,000 Bernoulli trials. We feel we need that many in our program to get sufficient numbers of trials to have high enough confidence that our X factors are legitimate. When we do that, think of it in terms of a spreadsheet, although you wouldn't do that type of spreadsheet.

You're creating a grid of 100 million independent zeros and 1s, where for each of the 10,000 policies, you run 10,000 trials. Zero is no death, and 1 is death. You get your zero or your 1 via some random number, and if it's lower than the mortality over the test period, you credit that as a death or otherwise no death. You sum up your total losses for each of your 10,000 trials down the columns where each trial goes across each row. Each policy—10,000 trials—goes across each row, and at the bottom you get your total face amount of death for each of the 10,000 trials. The typical methodology that I've seen, because it's the easiest to think about, arrives at those entries one at a time, with one random number drawn for each, but that's not the only mathematically correct way to fill in the grid of 100 million zeros and 1s.

If you have any sequence of independent trials with equal chance p of success or q in the case of death, there's a probability distribution for the trial number where you get your first one, and that probability distribution function (PDF) is (1-p)^(n-1)*q. You could in your mind substitute q for p there, but I've used the more common formulation of this distribution. For instance, let's keep thinking of it in terms of death. If your chance of death on your first trial or for one policy is 1 percent or 10 deaths per thousand, the chance that your first one is in the first trial is 1 percent. The chance that it's not in the first trial but the second trial is 99 percent x 1 percent because you've got to have a nondeath in the first and then a death in the second. The chance that it's in the third trial is 1 percent x (99 percent^2) and so on.

Now, p, is that expected mortality? You can create a cumulative distribution function (CDF) for that cell—a CDF for the probability of the first death being on trial number n. If you do that, all you have to do is run your tests across the row. Remember that for each policy, you've got a row of 10,000 items. They're all distributed the same way, and you create a geometric distribution. I then sample from the CDF, so that CDF has 0.01 and then almost 0.01 and so on as we saw, and I fill in "n-1" zeros for that row and then a 1. I then draw a new random number and so on until that row is complete.

Let me back up. It may be true that if I get a random number of 0.04 or 0.05—and I'm pulling this out of my hat—for a 1 percent death cell, that could be a situation where the first death occurred in cell number 90, so then I would fill in 89 zeros and then a 1 in row 90, and to do all of that I just had to draw one random number. I would start with cell number 91, and let's say my new random number was 0.015. That would be a 2, so I would put one zero and then a 1 on that row. Keep doing that across until that row is complete. At the end if you've got five cells left and draw 110, you just fill in the last five cells as zeros. You're always going to get some point at the end where usually your random number doesn't get you to the exact last cell. It puts you somewhere beyond, and then you fill in zeros the rest of the way. You have to start the test over for each new row. You can't carry down and back to the left because the mortality rate will change, so it's not an identical independently distributed function anymore.

The end result is an expected reduction in the run time of your program, approximately in proportion to your average expected mortality rate. In other words, if your average mortality rate were 1 percent, you could expect your program to run in a hundredth of the time that it used to take. This could be offset a little bit by some added programming time on the front end, but I would argue for me that's always worth it. How many of you have ever not done added programming on the front end of a project and then pulled your hair out because you kept trying to run stuff or make little changes or things didn't work quite how you wanted, and each run now takes a whole day and maybe it would have taken a couple of hours? I've done it. It's common. I would argue it's worth it if you think this could ever come up.

Once that grid is filled in, you tally the results down the columns the same as before. The important thing is that the stochastic implications are identical, so I'm not giving you a shortcut approximation. I'm giving you a shortcut that's exactly the same as far as the stochastic implications.

I'm going to summarize what we've talked about today. First of all, normal is somewhat normal, but you have to be careful about approximating a distribution this way. It's much better if you keep your normal statements of confidence to the mean of what you're drawing from, and that's especially true if the sample is big. Sometimes for skewed distribution, it has to be very big, and the CTE 90 is an example of a skewed distribution where it needs to be very big.

When a distribution is not normal, you can draw statistics from the sample, such as a bunch of scenarios, and make inferences on the tail percentiles using order statistics. When generating stochastic scenarios, you want to be clear about what the stochastic implications really mean. This comes into play for you now when you're generating calibrated scenarios for CTE 90 because I think that's mostly what we've talked about today that applies there.

Finally, I think regression is a useful tool, but it's important to realize that increasing explanatory power is not the only goal. If you want to increase your predictive power, you need to keep the number of variables as low as you can while adding only useful variables to the mix.

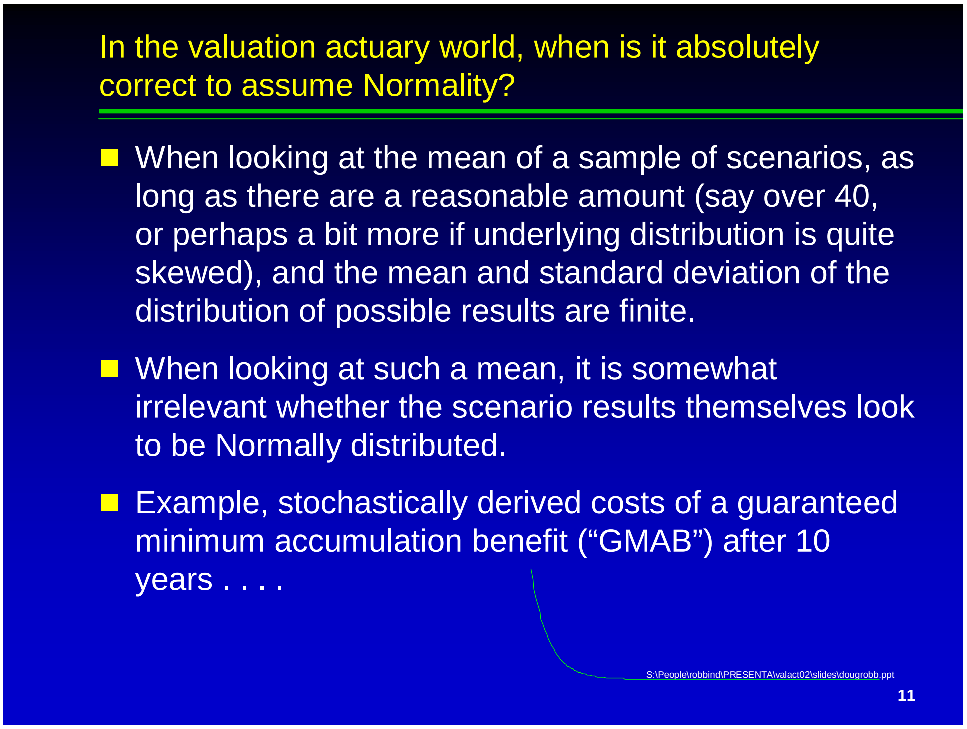
MR. ERIC SCHUERING: Could you explain in a little more detail exactly what happens in the calibration exercise that you talked about earlier?

MR. ROBBINS: In a calibration exercise, what's going to happen is you're going to look at a set of real-world equity scenarios that have been generated in some way, probably using a spreadsheet. This is the way I would do it. A lot of people would get one of your students to program this. You're going to look at all the returns for the first year of the scenario set and create a distribution where you know the entire range of first-year returns. You're going to create a cumulative return through the fifth year of each scenario set and create a distribution of that, and then you're going to create a cumulative distribution through year 10 of each scenario in your set and look at the distribution of that.

For each of those—the one, five and 10—the guidance is going to provide you calibration points that your scenarios have to meet, and your initially generated scenarios may not meet those. The first one is pretty weak. It's around the 10th percentile, and the next one's the fifth. I think the next one is the first, although I'm not 100 percent sure. I'm sure there's one at the 0.50 percentile and there may be one at the 0.25 or the 0.1 of a percent percentile.

There are also calibration points on the high end, but the guidance makes a lot of sense here. It's perfectly clear that you don't have to calibrate the high end of your scenarios if you're testing a guarantee on how bad things can get before the company goes into the money. If you were testing a guarantee that said, "If your fund does really well, we're going to bonus you 10 percent of the gains," or something like that, you would need to calibrate on the high end as well, so you'd have to look at the 90th percentile, the 95th, the 99th, etc. Using some methodology, and this is somewhat left up to the user, you're going to have to make your scenarios a little bit more volatile or make the returns a little bit worse possibly in order to calibrate this type of a deal.

Chart 1



In the valuation actuary world, when is it absolutely correct to assume Normality?

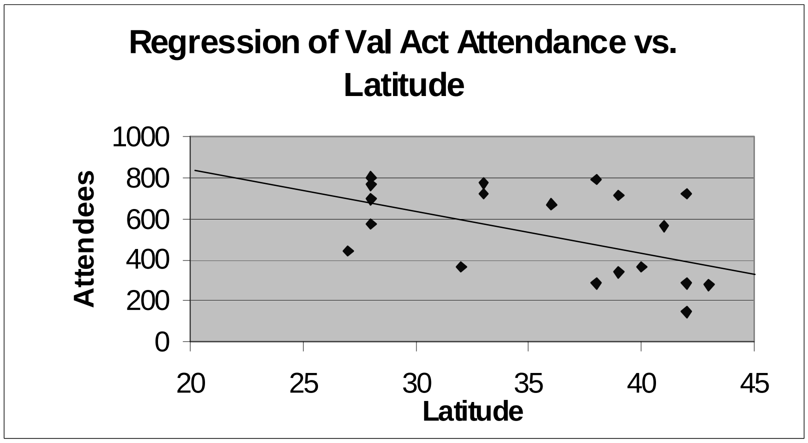
- When looking at the mean of a sample of scenarios, as long as there are a reasonable amount (say over 40, or perhaps a bit more if underlying distribution is quite skewed), and the mean and standard deviation of the distribution of possible results are finite.
- When looking at such a mean, it is somewhat irrelevant whether the scenario results themselves look to be Normally distributed.
- Example, stochastically derived costs of a guaranteed minimum accumulation benefit (“GMAB”) after 10 years . . . .

S:\People\robbind\PRESENTA\valact02\slides\dougrobb.ppt

11

Chart 2

Regression Example

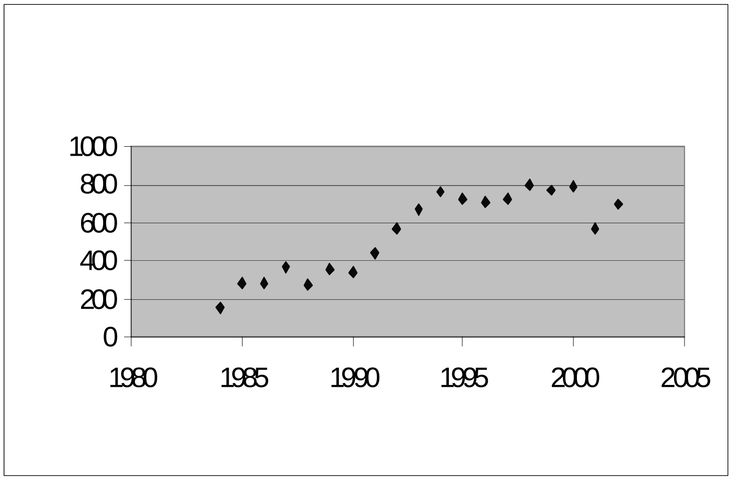


S:\People\robbind\PRESENTA\valact02\slides\dougrobb.ppt

32

Chart 3

Regression Example (continued) - What do you suspect that this is a chart of?

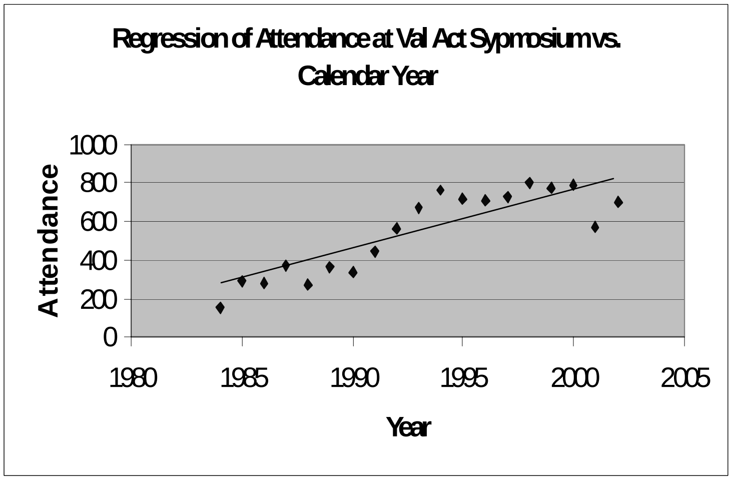


S:\People\robbind\PRESENTA\valact02\slides\dougrobb.ppt

34

Chart 4

Regression Example (continued) - That’s Right!



S:\People\robbind\PRESENTA\valact02\slides\dougrobb.ppt