2002 Valuation Actuary Symposium
September 19–20, 2002
Lake Buena Vista, Florida

Session 46TS
Drawing Appropriate Statistical Inferences

Instructor: Douglas L. Robbins

Summary: In this session, attendees are provided with some simple tools and guidelines to ensure that correct inferences are drawn. This material is presented in an easy-to-follow manner without using complex mathematical formulas. Topics include application of a normal distribution—when it is appropriate and when it is not, drawing inferences about and creating confidence intervals around tail percentile and certifying anticipated mortality under XXX in a mathematically concise way.

MR. DOUGLAS L. ROBBINS: I'm going to go through some statistical concepts that I think can be of use to you all as valuation actuaries. I'll start with the section, "How normal is normal?" and "what if it's not?" There is a third and fourth section with some basic statistical applications of other sorts that might help you. Finally, I have some miscellaneous issues and a summary. I was standing in the park with my kids the other day right before this started. I was thinking to myself, why is it that a Frisbee seems to get larger and larger as it's moving towards you? Then it hit me. That's not what I want to happen to any of you guys who have drawn some kind of statistical inference and then realized, as it moves toward you, that the basis of the foundation for it was improper.

Note: The chart(s) referred to in the text can be found at the end of the manuscript.

How Normal is Normal? What if it's Not?

In this session, we'll be covering two topics. How normal is normal and what if it's not? Basically, how was the normal distribution taught when you were in school? I'm not really talking about your actuarial courses as much as your basic statistics course that you first sat down for. There are many things in life with kind of a bell shape.

That happened to a lot of us, I think. Consider the following for probability distributions in general: for how many of the following distributions that I give you would you feel confident estimating the 99th percentile using a normal distribution knowing mu and sigma? (a) The number of accidents on a typical freeway into a large city on a morning rush hour, assuming the mean is 2. (b) The result of a single randomly rolled fair six-sided die. (c) The number of days late a given val act session will be filled with speakers. (Maybe normal isn't so normal after all.) Here's how you might do. If the true average number of accidents on our course is two, you might, on your own, estimate that the variance is also two, get out your normal tables and estimate a bit over five. The true 99th percentile, given the distributional assumption I'm making, is just over six. That's an error of about 20% of your guess.

The mean of a roll of die is 3.5 and the standard deviation is about 1.7. Using the normal approximation, if you're going to use your textbook tables, you shoot for the 99th percentile by adding 2.5 sigmas to the mean when you get 7.5. The biggest number on the die is six—so you've overshot by quite a bit. On the other hand, although the average lateness of valuation actuary sessions might be about a week, the 99th percentile of the lateness of one given the session is really hard to estimate since most of the sessions are filled on time and a few don't get filled until the week before the session is about to happen. Research into the past records for a given session, if there are any, might give you a better estimation technique than looking at the statistics. That's a little lesson in bipolar data.

Generally, when is the normal approximation a bad idea? It's a bad idea almost always when the data are somewhat scarce or bipolar in nature. This is especially true if it is bipolar or oddly shaped and when you are making tail inferences on an underlying distribution. Sometimes this can be clear from the sample data shape, but not always. I present the Cauchy distribution to you as one that appears very bell-shaped if you glance at it, but it has tail features that are amazingly different from normal.

Here is an example of the Cauchy distribution. Say you're in outer space where there's no gravity and you have a round, spinning paintbrush that's sopped with paint. Then, some distance away from it, there's a wall. The paint is either going to fly off perpendicular to the wall and hit or it's going to fly off at some angle (of course, half of it will go the wrong way and not hit the wall at all). Most of it will go off at some angle to the wall, and some of it is going to hit. If it goes off almost at 90^o, it's going to hit quite far away, right? That is the shape of the probability density function for Cauchy. The interesting thing about it isn't my paint brush example. It's the fact that that distribution has no mean or standard deviation. It's actually based on trigonometric functions as you might imagine since I gave all this degree talk. If you start drawing samples from a Cauchy distribution, you'll get something like +3/-2; +5/-1,576,000. It will be some number like that. No matter how long you sit there drawing samples, you will never get a mean that is tending to converge toward zero if that was the center point of the distribution. You're not going to get a sample mean that makes any sense. It's just going to bounce all over the place forever. That can happen with real fat-tailed distributions, which is a reason that the normal approximation would be really bad.

When is it okay to assume normality in a distribution? Under typically sound conditions (that's a word that's thrown around a lot in statistical textbooks, but in this case, I mean finite mean and variance), it is a sum of a large number of independent identically distributed random variables. Anyone remember IID from your statistic text? That's what that means. Do you have random samples from a distribution that you pull over and over, and they're all independent? For instance, say we're rolling two dice instead of one. We know the 99th percentile is about 12, the highest number you can get. The normal approximation will be 12.6, so that has reduced the error from 20 down to 5 (more dice will produce even less error). On certain types of defined distributions, one or more parameters might be large enough. Part of the reason for that is some of these distributions that I list as examples are, in fact, sums of other distributions. Gamma is the sum of exponentials. If the parameter that represents how many exponentials is put together to make the gamma, and if that's large enough, then that starts to be exactly normal. Poisson is the sum of Poisson. It's the same idea. Normal is a little different. It's not necessarily the sum of other distributions. If the parameterization is right (in other words, if the hump is not too close to zero), then lognormal stuff should be very normal, and you can draw inferences as though it were.

Let's say the number of car accidents in a day was Poisson, for example. We're looking at the whole state of Florida instead of just I-4 in the morning. The mean is 300 instead of 2. Take the square root of 300, which gives you the standard deviation, and that would produce a normal approximation for the 99th percentile of about 340, which is almost exactly right given that it is Poisson.

In the valuation actuary world, when is it absolutely correct to assume normality? You'll often hear discussion of the normal distribution related to drawing inferences on a set of scenarios. As valuation actuaries, we're often asked to produce a set of scenarios results for a given run, and then make distributional inferences. That's what we're really getting at. Suppose we're looking at the mean of a sample of scenarios. We hope they are reasonable (over 40). It really depends on how skewed the underlying distribution is, but 40 may be a good number of a lot of what we do. Then when you look at such a mean, it doesn't even matter whether the scenario results themselves look like they're normal. The mean itself is distributed normally because what is a mean? It's the sum of 40 or more independent samples divided by N 40. The fact that you're dividing by N doesn't affect the fact that you have a sum of IID random draws. Thus, the more you have, the more the distribution of the mean has to become normal over time according to the central limit theorem.

For example, let's look at the stochastically derived cost of a guaranteed minimum accumulation benefit (GMAB) on a variable annuity after ten years. You've all heard of a guaranteed minimum accumulation benefit. After ten years, the policyholder is guaranteed to get his premium back even though it's a variable annuity and he has probably invested that premium in equity funds that may have dropped. He is guaranteed to get the premium back and the company has to foot the bill at the end of the tenth year. Most of my results are quite good; they're exactly zero. However, a few of them are bad. The last few at the bottom of the distribution are very, very bad.

The funny thing about that example, the shape of the data, makes it quite clear that the distribution underlying the numbers we're getting is not normal; it's very, very skewed. Most of the results are zero, and a few of the results are not. Those have a very fat tail. However, I note, in this case, beyond a shadow of a doubt, that the mean meets the criteria I gave you to be considered normal if you run enough scenarios. This pertains to the IID and the central limit theorem stuff that I mentioned. The reason is because the conditions I gave were a finite mean and finite variance. In this case, even though my cost can get really large, when looking for GMAB, it's bounded at the premium paid at the start of the model. Actually, it's probably bounded at an even lower number than that because you probably have lapses and debts that occur in your model, to some extent, even if the benefit is in the money. It's going to be less than the premium paid, but it will be somewhere between zero and the premium paid at the start of the model. Since it's bound by both ends, it can't go to infinity. I know that it has a mean and a standard deviation. Any bounded distribution must have all finite moments.

What's the practical conclusion? In most of the testing that you do as a valuation actuary, you can assume that your scenario mean is distributed normally. What does that mean you can do? Here's what it means you cannot do. You can't take the mean plus two standard deviations and get the 97.5th percentile of the underlying distribution. Doing that would be assuming that the underlying distribution is distributed normally, and we see that that's not true. What it does mean is that you can draw a symmetric or nonsymmetric confidence interval. You could have just a one-tailed confidence interval. You can draw a confidence interval around any scenario set mean using that mean and its standard error. The standard error of the mean, since it involves N random draws and N random samples from your distribution, is the standard deviation divided by the square root of N. It's the standard deviation divided by the square root of the number of scenarios.

Let’s look at the GMAB and the kind of distribution you would get if you were testing the cost of a GMAB. I have 100 equity scenarios at a $1,000 initial premium. At the end of the runs, I have 90 scenarios that produce no cost at all, five scenarios that produced a cost of $5 apiece, three that cost $20, one that cost $41, and one that cost $74. So it’s pretty fat-tailed and pretty skewed. The mean cost is clearly $2.00; you just add up these numbers and divide by 100. The sample variance that I’ve worked out is 81 and the standard deviation is 9. It’s the standard deviation of your sample. I used N, not N –1. You could use N –1, but allow me to use N here.

Dividing by the square root of 100, we get the standard error of the mean of 90 cents. The sample mean is two bucks and the standard error of the mean is 90 cents. I’m 95% confidant that my population mean is between 20 cents and $3.80. So that’s the true expected value of the loss for this GMAB. It says nothing about the tails; that’s what I’m trying to make clear. We’re just trying to estimate the expected cost if we could do this as many times as we want forever.

Since I’m doing a 95% confidence interval, I’m looking at two standard errors, and I get down to 20 cents. It’s clear that if I went much further I would be estimating that my cost could be negative, which we know is impossible. That’s a sign that I might not have run quite enough scenarios. If it starts to look like even a reasonable tail percentile gets you all the way to zero and beyond, you might need more scenarios to get a true normal shape. It’s not going to be a terrible estimate of the 95th confidence limits on both sides.

Does this imply that if I run 100 new scenarios under identical conditions I should be within 20 cents to $3.80 95% of the time? No. That’s because this sample has some sample error and that sample will have its own sample error. In some cases, the two cases will compound each other. However, the inference is on the actual population mean. The real inference is, if I ran an infinite amount of new scenarios, as I said before, then you would expect that 95% of the time it would fall within that interval.

I’m going to talk about the tail of the distribution. I’m not talking about the mean any more. Note that the mean plus three standard deviations was $9.00 each. If it was normal that would be the 99.87th percentile. If you look in your table, that only gets you to almost $29. I think there was $45 and $74. Only 2% of my sample is well above that point estimate. We can see why it's a bad idea to try to make inferences on the tails of the distribution, assuming it's normal when it's not normal. It's just one more example.

What can I do if I want inferences on the tail? That gets us to what if it's not; what do I do if my distribution is not normal? If I can't make inferences on the tail based on the normal approximation, what can I do? What if I don't know the distribution? The answer I'm going to go over with you lies in the realm of nonparametric statistics.

First I want to tell you about the e-lottery. This is a game I invented during the Web craze. I thought—wow, there's lots of opportunity out there to make money and here's a way. I get together a huge crowd of people, say a billion. I randomly pick and record a number between one and the number of people (1 billion). I get a trusted firm to certify and record my pick. Each of these people puts in a buck, so I have a billion dollars in the kitty. Then, without conspiracy (that's why it has to be a billion people and not a million picking 1,000 times each), they have to independently, without talking to each other, pick a number between one and a billion and try to guess correctly. If there are multiple winners they split, but if there are no winners I keep the one billion dollars. What are my odds of keeping it? It is pretty good actually. To keep the money and retire, which is after all the goal, I need everyone to guess wrong. What are the odds of one person guessing wrong? Well, 999,999,999 in a billion because there's that many numbers they could pick and be wrong. So that's N – 1/N or a better formulation for what I'm about to do is 1/1 – N, where, in this case, it's a billion. If all the guesses are independent, the odds of all N people guessing wrong are 1 – 1/N^N. As N gets very large this happens to approach 1/e, so that's a pretty decent shot. 1/e is better than one in three – it's about 36% or 37%. Now you know why this is called the e-lottery because the odds are 1/e of me winning.

What's the point? It turns out that roughly the same odds apply to the underlying 99th percentile in my 100 scenario GMAB example. You remember the 99th percentile was $74. It depends how you define percentile actually. There are a couple of different ways the textbooks tell you to do it. The unbiased estimator would actually put it a bit higher than that. It would be around it. You wouldn't be far off of $74. Anyway, roughly the same odds apply to that in this way. The chance of any individual scenario being less than the 99th percentile is clearly 99%—that's the definition of the 99th percentile. What that means is, in this distribution of cost for the GMAB, there is the true 99th percentile. Every time I run a scenario, my chance of being lower than that is 99%.

Let's assume independence of all of my runs. If our scenario generator works correctly, we can expect our random seed is good. Assuming that's true, the chance of all 100 of them being less than the 99th percentile is 0.99^{100}. One hundred is not as large an N as a billion, but the approach to 1/e is from the underside so this actually works out conservatively for us. What this means is, in 100 scenarios, I have only about a 37% chance that all of my scenarios will be less than the actual 99th percentile of the underlying distribution. I have about a 63% chance that my worst scenario, the $74, limits the 99th percentile of the underlying distribution. In other words, the true 99th percentile is only 37% likely to be higher than that. I have a 63% chance that that's my confidence limit on the 99th percentile of the underlying.

If I'm doing some kind of target surplus testing where I'm concerned about the 95th percentile, my odds go to about $1/e^5$, so that makes me about 99.5% confident that my worst scenario limits the 95th percentile of the underlying distribution. That's very good. If I have enough like asset fees or whatever to cover my $74, then I have enough to cover what I'm almost certain is my 95th percentile. If you're doing cash-flow testing or something else and your boss tells you, you care about the 90th and you run 100 scenarios and don't fail any, you can be rock solid certain that your worst scenario limits were the 90th percentile of where your actual distribution could be. You can be that certain that you won't fail at the 90th percentile.

What if I do care about the 99th percentile? Then I've only got that 63-64% confidence, and that's pretty poor. Then what you need to do is just run more scenarios. If you run 1,000 scenarios, and you want to make an inference on the 99th percentile, this puts the odds of your worst scenario being less back at $1/e^{10}$. It's so far out—probably ten significant digits. You're just about rock solid certain again. That's based on your worst scenario not failing. You're absolute worst one is not failing. You can do this stuff with e that I'm talking about.

Do I always have to use the worst scenario? No, that's the simplest case. I wanted to give you something you could hold onto with the $1/e^N$ where N is the number of scenarios past your tail. I wanted to give you that to keep in your back pocket because you might even be able to remember it next time this comes up. There is something you can do with nonparametric statistics. Any time you want to look at a percentile, you set up a confidence interval based on the fact that each scenario is really a Bernoulli trial based on the percentile you're trying to estimate. At the 95th percentile, every time you run a scenario, you're either going to fail or be at a higher than 5% chance or not higher than a 95% chance.

For any percentile with the underlying distribution and any ranked scenario, you can work out the odds of a binomial theory. You can use that to construct any size confidence interval on a percentile that you might want. Most of these actually are one-tailed, so when I say any size, I mean you're going to put the top limit on it and not worry about whether you pass your way off on that side.

You can also construct a stochastic hypothesis. I run 1,000 scenarios to make an inference on the total surplus of my company not just target surplus. So I want to be 99% sure for some reason that I pass as a company. I want to be sure that I pass 99% of the global scenarios out there. I run my 1,000 scenarios, and five of them fail. If ten of them failed, that would be right on the cusp. My point estimate would be the 99th percentile for ten failures, but I only failed five so that sounds good. I decide that I'm going to do a hypothesis test and my null hypothesis is that my universal 99th percentile would fail. I want to reject that hypothesis and come up with the conclusion that it passes. If that null hypothesis were true, the chance of a success could be no more than 99%. It could be less, but we're going to get conservative and assume that it is 1%.

Under the null hypothesis, the highest possible probability of zero failures would be 0.99^{1000}. With this many significant digits, there's zero probability of no failures. For one failure, I start to get my binomial distribution out because I've got 1,000 Bernoulli trials. This becomes a binomial distribution, and I get a probability of 0.00004. Precisely two failures would be 0.0022, and so on. Then I accumulate up to five or fewer failures, and I end up at 6.61%. Thus, at a 5% significance level, I cannot reject the hypothesis that my 99th percentile results in the universe of scenarios as a failure. I can't reject it, so I'm stuck with it. That doesn't mean I've proven that it is a failure. It just means that I can't assume that it's not at a 95th confidence level or at a 5% significance level.

This is kind of like when you're hypothesis testing a mean and you want to prove that it's bigger than some number. You might get a number that's bigger, but not big enough. It's in the right direction, but it's not conclusive. If we had had four or fewer failures instead of five, then we could have rejected. The answer to the puzzle is we wanted four, but we got five so we're out of luck. Do we need to increase our surplus? Are we stuck with increasing our surplus or finding a way of selling bonds, stocks, or whatever? Are we trying to find a way of getting ourselves in a better position? The answer is—no, not necessarily. Remember, the sample result was in the direction we want; it was much less than ten. It was only about 0.6%, which is much better than 1%. But we can always refine our estimate by running more scenarios. If you're going in the direction you want to be, but not by enough, that's probably the thing to do. If we run 10,000 scenarios and the failure rate is 60, we can easily reject and go celebrate.

Some final notes on confidence statements. In this section, we've made several statements of probability or confidence levels. Those statements are based entirely on the economic scenarios. I want to make sure that's absolutely clear. We're running 1,000 versions of the future economically. Under those we reject five or we fail five or something like that. That assumes that all your other model assumptions are correct. Actually, those are very unlikely to be correct, but actuaries are very comfortable with decrements over time. They just seem to be things like decrements, and things like mortality and lapse rates. They're comfortable that even though for a given issue we might get totally different results than this, they're comfortable that, if you sell enough business, these things are going to even out. The thing that we're really after is, how adversely could the economy affect us? That type of economy is going to happen only once. It could be a market that goes down and goes down forever with all our GMABs being terrible. That's what we're after when we do this kind of scenario testing. The same thing is true for the stuff you've been doing for years on your C-3 for cash-flow testing.

Whether you're comfortable or not with your other assumptions, you should be aware of and probably disclose the basis for these confidence statements that we're trying to make. You might want to sensitivity test other assumptions.

Statistical Applications: Part 1

Let's talk about generating stochastic scenario sets. There are a few key issues to think about. There's the mean return of the stochastic scenario set; volatility of returns; and the benefits of diversification. I just want to go through what's tricky about each. Any of you that has gone through testing on your variable annuities are going to be somewhat used to thinking about this. If you've been mostly working with fixed and other scenario testing on fixed products, some of this isn't going to matter so much to you, but it might be interesting.

Let's look at the mean return of a stochastic scenario set first. Let's say it's a large cap. We're going to create a scenario set, and we want to point it toward a certain mean return, even though we're going to do 1,000 scenarios. Often when you're generating something like that, someone in the company or somewhere will give you a quote like this. From 1970 to 2001, the Standard and Poor's (S&P) 500 Index is just the index net of dividends. It had an average return of almost 10%. What can we make of that? If we look at the actual annual returns from January 1 to January 1 of each year, the arithmetic average was 9.75%, so that would be a true statement. If you add dividends in, you'll get a total return, but the index return was about 9.75% as an arithmetic average. The S&P 500 value at year-end, 1970, was about $92 and at year-end, 2001, was $1,148. If I take the return on the index to the 1/31st power for 31 years, I get a geometric average return of only 8.5%, so where did the other 1.25% go? This is actually what happens any time you introduce volatility into a set of scenario returns. Most of you probably invest, to some extent, and your financial planners have told you that you need to diversify. You might think, why? I'm young. The market returns are going to even out over time. What they're kind of trying to say is if you can get a good market return, then by diversifying you'll reduce your volatility. It reduces the effect that we're talking about here. If you really could get 9.75% every year, that would be one thing. If you get that on average, but in reality you're getting some –10s and some +30s through time, you don't actually end up making nearly as much as you would with the 9.75% each year. This is what you need to think about when you're creating your scenario set.

If you get two 10% returns in a row, you get 21% total: 1.1^2 is 1.21. If you get a zero and then a 20, your total return is 20. With –10 and 30, you'll see that it's even worse. The bigger the volatility gets, the worse it's going to be. There's a decent approximation, although I'm not sure what the theoretical basis is, but it seems to work pretty well. A decent approximation for the difference between an arithmetic and geometric average for a scenario set is $\text{sigma}^2/2$, where sigma is the volatility. If I have a 20% volatility, what that's saying is 20^2 is 4, and 4 divided by 2 is 2. That means there would be about a 2% differential between the mean return on an annual average basis and the geometric return.

What are some important things to realize? The mean return you want to shoot for in your scenario set should be based on a statistic generated in a similar fashion. If you're using the 9.75%, and you're going to look at your scenario set, then you want to look at the annual average of your scenarios. You want to look at each return in each scenario one year at a time. Then average them and get to 9.75%. If you're using the 8.5%, then you can look at the geometric return over the life of each scenario and approach it that way. The bottom line is that actuaries don't grasp the return 100% of the time. The return on a levelized liquid account environment applies to a lot of VAs. You do your level 9%. That's geometric. That assumes no volatility at all. So it assumes the two averages are exactly equal. For that reason, the returns that you see in valuation or in pricing actuary work are actually higher compared to history than the people who are doing the pricing realize. That's why the guy's head is on the table. That could be the Frisbee, too. There's more and more pricing based on stochastic. If people don't make their return on investment, this is part of the reason.

Generally, when people talk about volatility in economic scenarios, they are referring to the standard deviation of the annual returns. That's mostly a pretty good definition based on the way people try to calculate volatility and based on the history of the S&P or whatever. Annual volatility is really important for valuing flat benefits, like return of benefit and ratchet benefits on a variable annuity where big sudden drops cause a lot of the cost. How volatile is my return going to be? How likely am I to go down to 40 next year?

It's less of an issue for roll-up benefits and/or benefits with a substantial waiting period. Take a ten-year guaranteed minimum income benefit (GMIB). The annual volatility is an issue, but not quite as much of an issue there as it is for an annual ratchet. For the later benefit, some measure of long-term volatility may be needed. In the Canadian standards, they've looked at one-, five- and ten-year returns. How bad they get at tail percentiles might be the way we end up going.

Another thing going on there is correlation of returns between successive years and how that affects long-term as opposed to short-term volatility. If you have really high volatility at 20-25%, but you assume that your returns are negatively correlated so that a market drop makes it more likely to later see a market rise, then that will tend to mellow your long-term volatility. On the other hand, if you assume the Doug Doll market-drops-are-succeeded-by-market-drops-because-everyone-panics type theory, then that makes your long-term volatility worse than it would be if everything was independent. You should understand assumptions when you're generating your scenarios and working with volatility.

The last tricky issue is the benefit of diversification. I'm going to do a quick example that you've probably all done in your head—return of premium, GMDB, on a variable annuity. Fifty percent of your fund value that you've sold the first year was invested in bonds and 50% was in equity. That's not meant to be realistic; it's just my example. After a year, the equities have gone down ten and the bonds have increased 15%, net of fees, so you must have had a big interest rate drop due to a flight to quality from the equity market or something. Are you in trouble on your GMDB? It depends. That is the point of the benefit of diversification. If all of your policyholders are invested and totally diversified, where every single person is 50/50, then they have a total return on the fund of 2.5% and your GMDB is fine. It's 2.5% out of the money. If, on the other hand, half of your policyholders are invested 100% in bonds and half of them are 100% in equity, then the bond people are 15% to the good, but the equity is 10% off. So that 50% of your total fund value that you start with has gone down to 45%, but the GMDB is still at 50%. If everyone were to die, the company is at risk for almost 5% of the total fund value that you have. It's important to see the difference there.

In the second example, it's not that the diversification didn't help you at all. You're better off than if everybody had invested in equity, but you're not nearly as well off as if everybody was mixed between the two. That was the simple stuff that you've probably all thought about a lot. Diversification helps you; it mellows your volatility a lot. How much does it really bear out in the long run? It depends on your mix of funds offered and how they are held. Let's look at the mean return of a mix of funds in the long run. You have five fund classes—large cap equity, small cap equity, international equity, general bonds, and some cash. The funds' mean returns and the mean return on your total fund balance will be the same as the weighted average of those funds' mean returns. However, the volatility of a mix of funds depends heavily on how they're correlated with each other.

The following simple guidance will allow you to get a feel for this. Say you mix two funds given certain initial weights. In the short run, you can say that if they're uncorrelated, the total variance return will be the weighted average of the variances. If you have two independent funds at 20% volatility each, square it to get 4%, add the two together to get 8%, and take the square root. It's not 40; it's just the square root of the variance so it's just the sum of the two. If that was 50/50, it would be half of that, so it would be less than 20.

For funds that are 100% positively correlated, the total volatility is the weighted average of the volatilities, which means no gain or no volatility help at all. If you have a fund that is an S&P 500 index and a fund that also is an S&P 500 index, then that's obviously no good. Say your second fund is a Dow Jones Index. Even if it's very positively correlated, the mix there is not going to help you a lot even if all your policyholders are diversified like we talked about.

When you're looking at how to construct a diversified scenario, it's important to realize that. For funds that are 100% negatively correlated, the volatility is the difference in the volatility instead of the fund, and that could be a very useful result indeed if you're trying to get nice looking results and low volatility.

Let's take another example. Say you have two funds that both have a mean return of 8%. The volatility of the first is 20%, and the volatility of the second is 10%. You've got a 50/50 mix. If they are both 100% positively correlated, the volatility of the 50/50 mix is 15%. That's what I was talking about. You get no help there. You expect 15% not knowing anything, and you get 15%. If they're uncorrelated (you can be uncorrelated without being independent), the volatility of the 50/50 mix is going to be 11.2%, so now you have nice volatility savings at 4%. If you have an annual ratchet GMDB, you'd be surprised by how much of the cost that will knock off. If they're 100% negatively correlated, the volatility is only 5% because it's the difference between the volatilities. It's actually not just the difference. How do you know what to subtract? What if you get in –5 instead of +5. That doesn't make any sense. Actually, it's like the difference squared and square rooted. You're going to end up with +5 no matter which way you do it. The point is, you get a huge volatility savings.

Such things do exist, but you end up with a poor return when you do them. The correlation between a fund that is an index fund and then a fund that is a bunch of calls and puts that are bought at a certain strike price would be 100% negatively correlated, but you'd end up with the risk-free rate, which is what you get if there's no risk. That's why they call it the risk-free rate. There's not much out there that's 100% negatively correlated that pays well. Our scenario generator bases things like aggregate bond and large cap stock on having some negative correlation in some instances; we've seen that in today's economy. The terrible stock market has induced a huge flight to quality, and the yield is way down for Treasuries. Therefore, a bond fund would have gone way up because of that flight to quality. It doesn't always happen, but even a small negative correlation can produce some of this effect and help you a lot. The bottom line is, when you're constructing your scenarios these short-term volatility issues will affect long-term consideration if other conditions allow it to.

One more key issue with this is the way you run your scenarios. Earlier I said that, based on mix of funds, you could calculate this weighted average volatility in the short run. The only reason I said that is because it's not the theory that changes, it's the weights. If you start out 50% stock and 50% bonds, after one year or after one month for that matter, your weights are going to change because the two returns won't be equal. You can either assume your policyholders all rebalance when you're doing your scenario testing, or you can assume that the weights drift over time, and that gives you different parameters for your future statistics.

Let's do a summary on diversification. There may be substantial savings from including a mix of funds in valuing certain blocks of business, but the value of diversification is dependent on the funds actually held being diversified by individual policyholders, not just within the company as a whole. The value of diversification is also dependent on the funds being relatively poorly correlated; otherwise, it doesn't do you much good at all.

Statistical Applications: Part 2

General Linear Regression is the statistical tool that's probably still used by more actuaries than any other tool (bar scenario testing, but I'm not sure). I've seen regression used quite a lot. Why? It's common in the spreadsheet packages that we all love to use like Excel. It's easy to remember how it works. It's easy to remember what the solution means, why we were doing it, and what the basic goals are. However, it's also fraught with pitfalls that should be avoided if you, as the actuary, want to get meaningful results.

The goal of linear regression is to demonstrate a linear relationship between two bodies of data. You want to minimize the square differences (that's how you get your line). You want to estimate coefficients for one or more independent variables; and then you want to generally use the relationship you developed to predict other data points, often when they're somewhere off in the future.

I have a regression example. Chart 1 is the regression of attendance at the symposium versus latitude. Latitude refers to the globe. When they came up with the phrase, "Some like it hot," they might have had me in mind. My goal here is to convince the Society of Actuaries that they'll get more people to come if they have it places like Orlando and San Diego, not Boston or New York. At any rate, we'll see how well I do at this goal. It looks like it might work. I've drawn a regression line that I actually solved for first. I didn't just paint that line in. You can see, over time, the symposia held in northern locations. I think Orlando is not quite the furthest south, but it's in that string of three in a row looking from top to bottom. The top one had about 800 in attendance. (I forget which year that was.)

What's the summary? I've got an X-coefficient of –19. The T statistic of –2.18, which you all might remember is pretty good. T is like normal except not quite, but if you have enough degrees of freedom it is. My probability value is 4.42% and that's good. That's better than 5%. At a 95% confidence level, I would assume that that's a good significant, solid parameter for my regression. I have an R^2 statistic of 23%. Does anyone remember what that means? It's basically a measure of how well you're explaining the error in the data. If all I had were these data points of attendance and didn't know the latitude, I would have some squared error, which is the variance. It is the variance of the sample. By using my independent variable of latitude, I'm able to explain 23% of the squared error by the regression and only another 77% is not explainable by this regression. Seventy seven percent is a lot, but the point is I'm doing 23% better with my regression than I am just by knowing the mean of the sample, which is all you would have to go on otherwise. If you do a regression and it turns out that the single variable that you're regressing is not significant, all you're left with is your coefficient—the A of your A+BX=Y. Then the A is actually the mean and you're just saying the mean +/- some standard deviations is the best you can do. When you throw in your independent variable, you can do 23% better. So that's not too terrible a result.

Chart 2 is regression of attendance at Valuation Actuary Symposia versus the calendar year. There you see a very, very strong upward trend. I've drawn another regression line. This is a single regression, if I just look at calendar year. It looks quite a lot better than the other one obviously. What happens if instead of doing the single regression I do a two variable x1 and x2. In other words it's regression so I add this as a second independent variable. I get a coefficient for calendar year. The important thing is the T statistic. Once you're talking about the standard errors or whatever, you get huge significance. In fact, it's like 10^{-7}. The coefficient on latitude drops very badly and the T statistic is just -.35, which is no longer even significant or useful. So my R^2 value went up to 86.5, which is a big jump.

Any time you add a new variable, the R^2 value is going to go up. Any new variable helps explain the data that are there and where there are no new variables. If I would have done calendar year by itself, which I did do, that already has 86.4%. If I then added my latitude, that just adds 0.1 to my R^2. That's what is going to happen if you add a variable that turns out to be a poor variable; it's going to add, but not very much.

What mathematical pitfalls are actually within the regression example that I've done? First of all, it's not that issue that I've already talked about, which is selection of an inferior predictive data set. That's an error, but it's not a mathematical error; it's just possibly a judgment error. Sometimes that's all we have to go on. This is what we've thought of, and we don't know the magic bullet that would have made the regression perfect. Like I said, sometimes some data are better than no data. That at least got us 23% predictiveness. But there's two important mathematical pitfalls that many of you will probably recall from back when you were learning this stuff. The first is nonconstant variance or what I'm going to call the outlier effect. The second is extrapolation beyond the range of the data.

Outlier effect. Recall the latest data point in the most recent chart. I mean the latest chronologically in 2001. It doesn't look so good compared to most of the others. In fact, that data point does have the largest squared error. The year 1994 was close, but no others were even close as far as squared error. Is that point really indicative? What happened in 2001? That was last year. You all probably remember what happened. They cancelled the symposium because of September 11 and then held it two months later when a lot of people didn't go and frankly a lot of people weren't keen on traveling yet. I'm not saying throw the data out. No statistician would ever say that. You might need to consider outliers and what might have caused them when you are drawing conclusions. Sometimes it is appropriate to throw some data points out if they're obviously caused by data that isn't going to be in place when you're trying to predict new data points. Interestingly, that data point was right on the line in my first regression of latitude versus attendance. So maybe Orlando was a really good choice for last year after what happened.

The second error is extrapolation beyond the range of the data. Say someone's told you a linear formula can be derived between 1984 and 2001 to come up with attendance, and you want to just have the formula and guess at how many people attended between those years. You'd clearly want to use the calendar year rather than the latitude formula. It's just a much better predictor. Interestingly, despite how things look on that chart, you might or might not find that for predicting attendance in 2010 or so, that regression would estimate 1,214 attendees, and it could end up being less useful, not more useful than latitude. Why is that? Within the continental U.S., latitude is going to be what it is now. It's going to be hotter in the south, and there are going to be cool places like Orlando and San Diego to go to. Unless there's a big demographic change of some kind that I don't know about, it's going to be similar. The symposium in 1984 was brand new like the product actuary symposium is now. It grew and grew and grew. It could be to the point now where it's like the spring meeting which has been fairly flat for a while. We might or might not have maxed out what the attendance is expected to be for a meeting that's really important (except the SOA annual meeting). If the data do flatten out, they are still going to be around 800, even way out in 2010. Latitude could be useful. On the other hand, if we use latitude to predict that if they hold it in Quito, Ecuador in 2003, attendance will be 1,215 people, that would probably be even worse than extrapolating on the calendar year. The point is that a regression is very useful if used internally. You're looking at the data as a continuum, and that's why it's a regression, not a nova or something. Usually, within the range of your data, it's fairly useful to predict future data points. Outside of the range you could just lose the linear in any other shape.

There's one often missed mathematical pitfall concerning multiple regression. It is not missed by textbooks or theoreticians, but by people doing Excel stuff on data. A mathematical pitfall concerning multiple regression is co-linearity of data.

Remember I said that adding any variable to a regression will provide more explanatory powers measured by R^2. That means there's this data and the more data points that I put in, the better R^2 gets and the more explanation on providing that data. I have a good example to make clear what I'm saying. Let's look at attendance in 2001 at the Valuation Actuary Symposium. I regress that with the size of the red spot on Jupiter. I'm going to get a perfect linear regression with 100% R^2. Why? Because I've got two data points and one line and it's going to fit. Some version of a regression on the size of that red spot is going to fit those two data points.

But adding variables to a regression, I'm criss-crossing hyperspace—4, 5 or 6 n-dimensions. If I cross it one less time than the number of data points, I'm going to produce a perfect fit, and it's going to explain all that data. Like the red spot next to your n Valuation Actuary Symposium in 2003, it's going to be absolutely useless for predicting new data because you've just kind of bollixed the whole thing. The statistics of it are such that because the degrees of freedom become so low, there's no statistic that would be significant for any of the predictors or any of the coefficients that you're trying to estimate. If none of your coefficients are significant, and if they can all be anywhere, then none of them are going to be useful for predicting an actual new data point (one that you don't know about before you do the study).

What it turns out in this case is that calendar year is actually a fairly good predictor of latitude. A regression between those two would find a relationship at the 5% significance level also. Because calendar year predicts latitude, adding latitude when you already have calendar year does not add significantly to your regression because of the co-linearity. There is a linear relationship there. Basically anytime n times one data set + a constant gives you something close to the other data set, it's going to be no good. Even though adding the new variable will increase your R^2, it's going to foul up your predictive power. It's going to make both variables worse than they would be on their own. Actually what this says is that the SOA has already been going my way on this so I don't need to do my study. They have already been having more symposiums in the south than they used to have.

What's the crux of the matter? I've kind of already alluded to this. Despite increasing R^2, adding nonuseful variables (nonuseful because they're co-linear) can reduce its predictive power or it could be adding nonuseful variable just because they're nonuseful and because they're not predictive. That would also have the same effect. So extraneous variables might muddy the water, and that's really the conclusion.

There are a few variables that could be useful to you and are never co-linear with your first variable. One is higher powers of the independent variable. Calendar year is a good predictor. What if I use calendar year squared? It's not co-linear; it will never be co-linear, and if there is some bowing and you want to do a quadratic fit instead of a linear fit, it can produce a really good two-variable regression. The regression is still linear in the co-efficient; that's why it's still called linear regression. It doesn't have to be linear in the independent variable; it can be quadratic, tridratic, or whatever polynomial you want. Other transformations that could be useful are natural log exponential, trigonometric functions, and if you haven't totally burned out your computer, there are probably others you could come up with as well.

Miscellaneous Issues and Summary

I guess I really should have said issue, because I've only got one. I added this in the miscellaneous section because it's in the syllabus. I said I was going to do a talk on stats, and I was going to talk about speeding up run time on Monte Carlo simulations for XXX. You're all valuation actuaries, so this is definitely a concern if you have term and UL with secondary guarantees out there.

A lot of the work in testing your XXX lies in doing your Monte Carlo simulations to try to estimate the distribution of your total losses in face amount given a distribution of mortality, which is your anticipated mortality. If your losses are bigger than expected, how much bigger do they need to be to be rejected? One main way people figure that out is to do Monte Carlo and to do enough scenarios so that they have a nice, shapely estimate of the pattern of possible losses.

I'm going to credit Ed Robbins as being an expert on this. He definitely gave me the idea of how this would work, and he would, in turn, credit a *Transactions* article 20 or more years ago. He can't remember which book it was in, and I was never able to find it either, but I'll just give the credit to that author for putting the idea in his head. At some point, when you're doing this, you could have decided you're going to reject any X-factor class where you come out at the 95th percentile or higher. You could have one that's right on the cusp at the 95th percentile when you do 1,000 scenarios. You might decide you need 10 or 100,000 scenarios on this Monte Carlo simulation. This could give you run time problems. I don't know about you, but whenever I do a big run and set it up to go over the weekend, I always come in on Monday morning and find it's perfect. It had no flaws in it, and I don't have to rerun. Your laughter gave that away. Whenever I do that I'm totally fouled up for the rest of the week and have to work overtime. I would like this to be able to run overnight so I can check it Saturday morning and fix what I screwed up, and then reset it and have it run it again.

When I do 10,000 Bernoulli trials, and each trial is a death or a nondeath for each of my policies. I do 10,000 trials across my programming matrix for each of 10,000 policies, I have a hundred million independent zeroes and ones. If I did 100,000 trials, I would have a billion independent zeros and ones, so that could take a lot of computing time where one is a death and zero is no death. The typical methodology that I see people use in their programming is to arrive at each entry one at a time with a random number. However, if you have an anticipated 1% for the year that the study was done, and if you get 0.01 or less on your random number, that's a death. If you get anything higher, that's no death. That's how you fill your grid in, but that's not the only correct way.

Anytime I have a sequence of independent trials with equal chance p of success, where p is a death, there's a probability distribution for the first trial where I get a success or a death. That cdf is called the geometric distribution. For instance, if my chance of success on any trial is 1% because my q is 1%, the chance of my first success being in the very first square that I fill in is 1%. The chance that it comes on my second trial for that policy is 1% times 99%, because, to get to that second box, I've got to first pass the first one, which has a 99% likelihood. The third trial is 1% times $0.99\%^2$ and so on. If p is really q, my expected mortality rate for a cell, I can create a cumulative distribution for that cell using the geometric probabilities.

Then all I have to do next is run the test across a row for that policy or that cell, drawing a random number, and making a comparison to the cdf to get a numerical result in. I draw 0.4 and that turns out to be the cumulative distribution function random number for like 50 in my geometric distribution of 0.01s—there is a 1% chance of death. I would then go across to the 50th cell and put my first "one" there for a death. The first 49 cells would be at zero. Stochastically, that's the same thing as drawing each of the 50 random numbers. It gives me a 0.01 chance that I was going to get a one in my first cell; a 0.01 x 0.99 chance that I would live on my first cell and die on my second one. Stochastically, it's identical. But I only had to draw one random number, and it got me all the way out to 50. In fact, I could do that all the way out until that row of 10,000 trials is complete. Clearly I must start the test over for each new row since the mortality rate will probably change from policy to policy (it's not going to be exactly the same). However, the end result is an expected reduction in run time approximately in proportion to your average expected mortality rate (i.e., if your average mortality rate across 10,000 policies or 100,000 policies is 0.01, then you should reduce your run time by a factor of 100 or it should take 0.01 times as long as it did). I'll point out a subtle hint there. This may be offset a bit by some added programming time or consulting fees on the front end.

Speeding up Monte Carlo. Once the grid is filled in, you still have your 100 million ones and zeros. Then you just tally your results down the columns as you would have done if you filled them in one at a time. So the stochastic implications, as I said, are identical.

In summary, normal is somewhat normal, but you have to be careful about approximating a distribution that way as opposed to a mean. The mean of a sample from most distributions that you'll be working with becomes normal if the sample is big. When a distribution is not normal, statistics from the sample (a sample being a bunch of scenarios) can still be used to make inferences on tail percentiles where you make those inferences using tail scenarios.

When generating stochastic scenarios, it is good to be really clear about what the scenario statistics really mean. There's the mean, the volatility, and anything you get from diversification. Regression is a very useful tool, but it is important to realize that increasing R^2 is not the only or even the most important goal. Testing of x factors might be able to be sped up, depending on what you're doing.

CHART 1
Regression of Attendance at Valuation Actuary Symposium
vs. Latitude

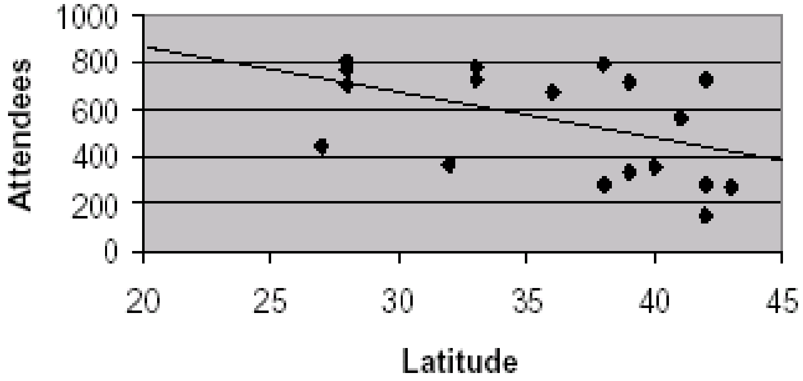
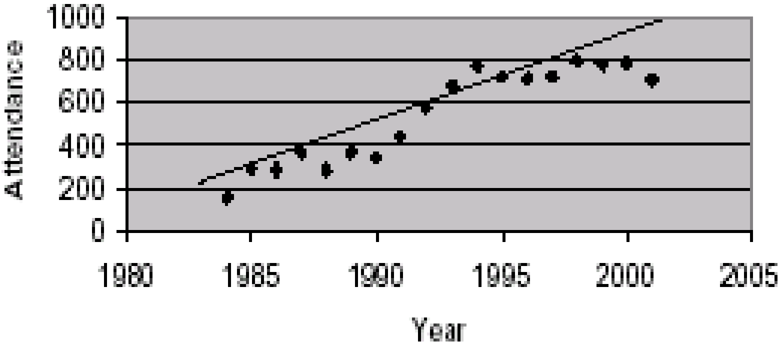


CHART 2
Regression of Attendance at Valuation Actuary Symposium
vs. Calendar Year


Attendance
Year