SOCIETY OF ACTUARIES

Article from

Risk Management

August 2017
Issue 39

Is Credibility Still Credible?

By Mark Griffin

With the advent of Principle Based analysis, many life actuaries who have not actively used Credibility Theory are beginning to dust off their notes. In fact, the scope of Actuarial Standard of Practice 25 on Credibility has recently been extended from P&C to include life insurance and pensions. In addition to Principles Based analysis; IFRS, Solvency II, and Embedded Value are cited as reasons the scope was extended. At the same time, insurance companies are seeking a consistent, transparent, documented approach to assumption governance as part of a "Control Environment" for financial reporting. Therefore, it is timely to ask if Credibility Theory, first developed more than 100 years ago, is still "fit for purpose," and to explore if computing power now enables a better approach.

Fundamentally, situations where we actuaries have applied Credibility Theory are no different than the hypothesis test question we all encountered in Statistics 101; "if x of our y coin flips are heads, should we continue to assume that the coin is fair?" Or put another way, when do we have enough data to change or confirm our assumption?

Consider the following situation where a hypothesis test approach could be used in place of Credibility Theory. You are the valuation actuary for a small but growing pension risk transfer business within an insurance company. The mortality experience of your block in the past year has been higher than the industry table you have been using. The natural question is, should you continue to use the industry table in light of the results? This example will translate directly to any mortality application. The same approach can be applied to any non-capital markets assumption.

Let us consider the null hypothesis to be that the industry table represents the true rates of mortality for our block. We will therefore test whether the mortality the block has experienced is plausibly a result of random fluctuation within our block's "sample."

We can test our hypothesis by calculating the degree of random fluctuation we should expect for our block. The expected level of random fluctuation will be a function of the size of the population, the period of time over which fluctuation is measured, and the assumed mortality rates. If the actual deviation from the expected is sufficiently high, we can reject the null hypothesis that the variation is random.

The standard deviation of the number of deaths in the past year is simply the square root of the sum of q_x times $(1-q_x)$ for all lives. In this situation we use the q_x from the industry table, as the null hypothesis is that these rates are the correct rates. Because we are evaluating the level of random variation, the Central Limit Theorem allows us to use the normal distribution to evaluate statistical significance without loss of generality.

The normal distribution tells us there is roughly a 5 percent probability that random fluctuation alone will give a result beyond plus or minus two standard deviations. In our example, we choose two standard deviations as our decision threshold at which to reject the null hypothesis and select a new assumption. With a threshold of two standard deviations, in just one time out of forty will we inappropriately move to a more aggressive assumption? Similarly, we will move to a more conservative assumption improperly one time in forty.

The block we are studying has just less than 4,200 exposures in the past year, with expected deaths of 168. The standard deviation of the number of deaths over the past year was 12.4 deaths. Actual deaths were 186, 16 higher than expected, representing 1.3 standard deviations. In this situation, there is some evidence that the block's experience is not representative of the industry,

Table 1
Credibility Analysis of Mortality Experience

Exposure	(1) Expected Deaths	(2) Standard Deviation	(3) Actual Variation	(3) ÷ (2)	Statistical Significance?
Current Year	168.1	12.4	15.9	1.3	No
Past 5 Years	881.3	28.4	61.7	2.2	Yes
5 Years					
Males	584.3	23.2	39.7	1.7	No
Females	297.0	16.4	22.0	1.3	No
Low $	532.9	21.9	32.1	1.5	No
High $	348.4	18.0	30.0	1.7	No
Younger 1/2	136.7	11.6	12.3	1.1	No
Older 1/2	744.4	25.9	49.6	1.9	No
Oldest 1/4	549.8	22.0	45.2	2.1	Yes

but this result is not statistically conclusive given our decision threshold of two standard deviations.

If we look at the past five years of data, exposures were 21,400, with expected deaths of 881.3. The standard deviation was 28.4 deaths. Actual deaths were 943, 61.7 higher than expected, representing 2.2 standard deviations. (Note that when the exposure base is stable, the standard deviation of n years of data is approximately the square root of n times the annual standard deviation). In this case, actual deaths are higher than expected by more than two standard deviations. Therefore, we should reject the null hypothesis and conclude that the industry table is not representative of the block's mortality experience.

The same techniques can be used to search for definitive sources of mortality variation. In our case, testing by gender or by amount of benefit does not give a conclusive result. However, testing by age band does give a conclusive result. The youngest half of our exposures are adverse by slightly more than one standard deviation, while the older half is adverse by 1.9 standard deviations. In fact, the oldest quadrant, which includes 550 expected deaths, is adverse by 2.1 standard deviations. This provides strong evidence that the industry table's rates are not only too low to represent our block, but also that the "tilt" of the rates we should use for our block are different than the industry table's "tilt."

Table 1 summarizes the analysis.

When one encounters an adverse statistically significant result, it may be tempting to do one of two things:

1. Look at a progressively longer period of data until the result seems more reasonable.
2. Set the decision threshold higher, (for example three standard deviations) to minimize the probability of drawing a false conclusion.

The downside of each action is:

- Ignoring statistically significant trends.
- Increasing the severity of the "reckoning" when incorrect assumptions are eventually updated.
- Violating the consistency and objectivity of the approach, whether it is the formal assumption governance process or not.

Therefore, it is important to look at shorter data periods for statistical significance first. If there is no statistical significance, then one should look to longer data.

The goal is to draw an accurate conclusion as quickly as possible. By expressing the decision threshold in terms of standard deviation, the conclusion is based on the degree of variation, the level of exposure and the underlying probabilities. The threshold may be reached over any period of time.

Let's address the same question using Credibility Theory. While there are different versions of Credibility Theory, the Limited Fluctuation Method is commonly used. Many actuaries will turn to a matrix of claims levels shown in the seminal 1962 paper "An Introduction to Credibility Theory" by L. H. Longley-Cook.

The paper shows a table of the number of claims required in order that a data set be deemed fully credible. One dimension of the table is the probability (P), which is similar to the decision threshold in the hypothesis test example. The other dimension is the "maximum departure from expected" (k).

The paper gives no explanation of how to select values for P or k, or how to incorporate the values into subsequent conclusions or analysis. The paper merely explains that the choice of P and k are arbitrary. Traditionally, the actuary has chosen values for P and k (presumably arbitrarily) and compared the required number of claims in the table to their volume of experience to determine if they have enough data to validate using their own experience as the assumption.

The VM20 manual directs that P should be 95 percent or higher, and k should be less than 5 percent. Once, again there is no guidance on how the choice should be made. In the Longley-Cook table, the value corresponding to both of those limits is 1,537 claims. The value for P=99% and k=2.5% is 10,623 claims, almost seven times higher, which shows how sensitive the application can be to these parameters. In the VM20 application, the credibility level is used to determine prescribed margins and how quickly company experience must be graded into the applicable industry table.

Returning to our example, if we chose the lowest value cited in the VM20 range, 1,537 claims, we would need almost 10 years of data to draw any conclusion, regardless of how good or bad our block's experience was relative to the industry table. As a consequence, Credibility Theory would direct us to stay with the industry table and revisit the analysis in another five years.

We should ask the following questions with respect to Credibility Theory:

1. Does it make sense that credibility analysis is not based on the degree of variation between the two sets of data being compared? In the coin toss example, if 20 of the first 20 outcomes are identical, we know the coin is biased.
2. Shouldn't the analysis incorporate the probabilities involved in some way? Going back to the coin toss example (one last time), if the null hypothesis were that heads would come up one time in 10, then 20 tails out of 20 flips would not disprove the null hypothesis. In terms of mortality analysis, rules of thumb that apply at younger ages should not be expected to be useful at older ages.
3. How can credibility theory be applied to other assumptions?

The hypothesis test approach has the following advantages relative to Credibility Theory:

1. The math is straightforward. It is easy to identify statistically significant results.
2. The methodology generalizes to other assumptions. A company could apply the same technique and decision threshold to all of its experience relative to its non-capital market assumptions, giving a cohesive, consistent, transparent approach company-wide. The decision threshold could be agreed on as part of a company's risk appetite setting process.
3. The hypothesis test is transparent and is easily understood by many outside the actuarial profession.

In our era of computing power, the hypothesis test is clearly a more accurate approach to assumption governance and should be used in place of credibility theory going forward.

Actuaries who have worked with predictive analytics will relate to the examples of testing the significance of gender, amount of pensions and age. Predictive analytics relies on hypothesis testing and incorporates the same probability measure to determine statistically significant relationships within the data. In fact, the "machine learning" version of predictive analytics can be thought of as a hypothesis test where the "machine" develops the hypothesis.

Use of the hypothesis test approach also lays the groundwork for the following best practices:

1. In our example, we are only interested in the assumed rate of mortality. Insurance company stakeholders are very interested in the financial results which are driven by assumptions such as mortality, withdrawals and premium payments. In our example, by simply multiplying the q_x times $(1 - q_x)$ calculation by the net amount at risk, we can calculate the expected dollar variation in net claims. In the same way, we can measure the expected variation in the impact of withdrawals, or premium payments, by multiplying by the financial severity of the event. It is the author's experience that calculating these metrics and proactively communicating across the actuarial, finance and risk functions helps to build a common understanding of the expected level of variation, and of the process for resetting assumptions. This understanding is critical in the financial close process, as it clearly delineates between plausible variation and areas where more focus is warranted.
2. Many insurance company's assumptions governance process now includes an annual review of all major assumptions by

an assumptions committee. While this is a better process than waiting for an assumption's owner to identify the need to change their assumption, it is still difficult for the committee to be objective about an assumption they have previously approved, perhaps multiple times. In any case, if expected variation is calculated for all major assumptions (as per the first best practice above) the assumptions committee's work can be transformed to focusing on only the assumptions where the results are outside a specific range (for example one standard deviation). There is no need to spend significant time on other assumptions. Such a consistent escalation protocol should also resonate with management, auditors, rating agencies, regulators, etc.

3. The hypothesis test approach can be used for experience studies, as shown at a high level in our PRT example for gender, amount and age.

Credibility Theory was presumably developed as a short cut to hypothesis testing and was well suited to the days of very limited computing power. In today's environment, computing power allows us to apply hypothesis testing directly, precisely and consistently across a wide range of assumptions. The hypothesis test is a simple but powerful tool and its adoption will enable actuaries to navigate numerous evolving analytical and process requirements. ■

Mark Griffin can be reached at *markgriffinct@aol.com.*