AN EXAMINATION OF CREDIBILITY CONCEPTS

STEPHEN W. PHILBRICK

Abstract

Credibility is one of the more important concepts in actuarial theory. However, it is one of the more complex concepts and is not as well understood as it should be. This paper takes a fresh look at some of the fundamentals of credibility theory in order to clarify and tie together various concepts.

Several loosely related approaches are taken. A new model is introduced to explain credibility concepts, an old model is discussed in more detail, and several potential ambiguities in the existing literature are directly addressed. This paper relies heavily on existing papers, particularly those on the *Syllabus*, and is intended to be read in conjunction with the various papers.

INTRODUCTION

The casual reader of articles on credibility is unlikely to come away with a lucid understanding of the true meaning of credibility. Consider the following observations.

Longley-Cook states: "While credibility and statistical variance are related, the former is meaningful only against a stated or implied background of the purpose for which the data are to be used and a consideration of the value of the prior knowledge available." He then goes on to establish a formula for full credibility based only on the properties of the observations, i.e., independent of the purpose of the data and the value of prior knowledge. When discussing partial credibility, he uses the formula $Z = n/(n + k)$ and notes that this never gives a value of 1.0, so he increases his partial credibilities by 50% to meet the full credibility standard. He then discusses an alternative (and inconsistent) approach, the so-called square root rule. Finally, he refers to Arthur Bailey's two types of credibility, "limited fluctuation credibility" and "greatest accuracy credibility," without fully explaining the differences.[1,2]

[1] Laurence H. Longley-Cook, "An Introduction to Credibility Theory," *PCAS* XLIX, 1962, p. 194.

[2] In his defense, it should be pointed out that he was not putting forth original theories; he was merely summarizing current practice.

Hewitt summarizes Mayerson (Lange quotes this summary) by stating that "credibility may under certain circumstances be a function of:

(1) sample size,
(2) underlying hazard (mean of prior distribution), and
(3) underlying dispersion (variance of prior distribution)."[3,4,5]

In Hewitt's review of Mayerson, Jones, and Bowers he states that: "There are, then, three variables which can affect credibility:

(i) number of observations,
(ii) variation in results (estimator for process variance), and
(iii) variation of hypotheses (variance of hypothetical means)."[6,7]

Mayerson *et al.* point out that existing standards are based only on numbers of claims and set out to establish a distribution-free standard for full credibility of the pure premium.[8] They define a standard of full credibility which is based upon the familiar P and K found in Longley-Cook. Hewitt's review claims that their standard is not distribution-free.[9] In his article with a similiar title, "Credibility for Severity," Hewitt never talks about P, discusses K (but this is not the same k as in Mayerson *et al.*) and never seems to talk about the number of claims or dollars needed for full credibility.[10]

It is not surprising that actuaries are not of a single mind when it comes to discussing credibility, since the various references are apparently inconsistent.

---

[3] Charles C. Hewitt, Jr., Discussion of "A Bayesian View of Credibility," *PCAS* LII, 1965.

[4] Allen L. Mayerson, "A Bayesian View of Credibility," *PCAS* LI, 1964.

[5] Jeffrey T. Lange, "Application of a Mathematical Concept of Risk," *The Journal of Risk and Insurance*, Volume XXXVI, No. 4, 1969, p. 385.

[6] Charles C. Hewitt, Jr., Discussion of "On the Credibility of the Pure Premium," *PCAS* LVI, 1969, p. 79.

[7] Allen L. Mayerson, Donald A. Jones, Newton L. Bowers, Jr., "On the Credibility of the Pure Premium," *PCAS* LV, 1968.

[8] Ibid., p. 175.

[9] Hewitt op. cit., p. 81.

[10] Charles C. Hewitt, Jr. "Credibility for Severity," *PCAS* LVII, 1970.

In this paper I would like to accomplish the following goal:

Explain credibility, via examples, so that the reader will have an understanding of the true nature of credibility.

Being realistic, I will be satisfied if this article provides enough of a focal point so that by re-reading the various articles, you can bring together the various concepts. Each of the authors is essentially talking about the same thing, but they each make certain simplifying assumptions (some explicit, some implicit) which, in some cases, tend to oversimplify the concept; that is, some of the essence of credibility gets simplified into thin air.

The format of this paper will be as follows:

- Discuss the concept of credibility using a target-shooting example. This example is easy to follow and reasonably analogous to insurance situations.
- Expand the discussion using an example similar to Hewitt's die-spinner model. The model is slightly changed and the discussion, emphasizing a different look at essentially the same example, may be enlightening.
- Explain how ratemaking and experience rating credibility concepts differ and the impact this has on credibility formulas.
- Correct the misconception that large values of credibility are always desirable.
- Summarize some of the credibility articles which are required reading for the actuarial exams.
- Discuss some of the simplifying assumptions made by various authors that can lead to the apparent confusion pointed out in the beginning of this introduction.

CREDIBILITY AND MARKSMANSHIP

"And now for something completely different."
-Monty Python

In this section an example will be presented, somewhat removed from the world of insurance, but one that I hope will give an insight into credibility. Consider the following situation. One of four people—*A*, *B*, *C* and *D*—will be chosen at random. The person chosen, whose identity will be unknown to you, will fire a gun at a target some distance away. Your task is to provide the best estimate of the location on the target which will be hit by his *next* shot after observing the location of the shot.

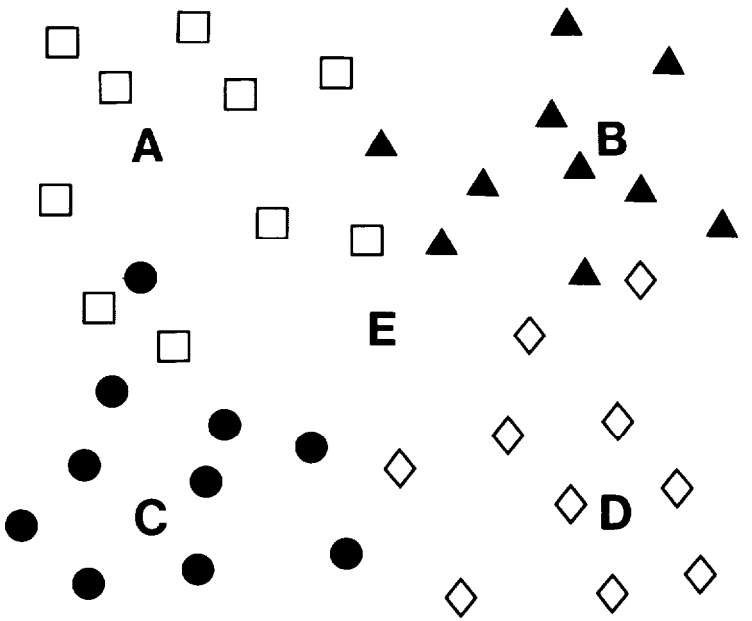
You also have some additional information. You have the results shown on Figure 1 of each of the four people firing a number of shots at an identical target. The squares represent the shots fired by person A, and the position marked A represents the center, or mean, of each of these points. Similarly, B is the center of the points marked by the triangles, C corresponds to the circles and D corresponds to the diamonds. The point E corresponds to the mean of all the points, or equivalently, the means of A, B, C, and D. Inspection will reveal that there is a clustering of the various symbols about their mean, although the clusters overlap. It can be presumed that each of the people is aiming for his respective mean, and the scattering is the result of random disturbances.

Prior to the observation of the shot, the best estimate must be based solely on the prior information; hence the best choice is E. Now we will consider the problem of making the best estimate after observation of a single shot, based on the current observation and the prior information.

If a strict Bayesian analysis procedure were followed, the next step would be to calculate the new probability that the shot was fired by A, B, C, or D, and then calculate an E based on these revised weights (see Hewitt for a discussion along this line). A Bayesian credibility approach would proceed as follows. Draw the straight line between the observed point and E. Determine the credibility Z of a single observation, and locate the point $100Z\%$ of the way from E to the observed point. The crucial point is the calculation of the credibility. In this example, the intent is not to do the explicit calculations, but to justify, on intuitive grounds, the calculations to be done in the next example.

Assume that the observed shot lies somewhere between A and E. Although a revised estimate would lie along the line connecting the observed point and E, it would probably not be far from E. Why? Although points in that region are more likely to have been produced by A, values corresponding to B and C are in the region, and D cannot be ruled out entirely. A is more likely; hence the revised estimate should be closer to A, but not much closer because the evidence for A is minimal.

FIGURE 1



Consider Figure 2. This figure was produced by four different people, A', B', C', and D'. Their "mean" shots were identical; hence E' coincides with E. However, these four are much better shots. Their shots cluster more closely around their mean. Mathematically, the process variance (the mean squared distances between the actual points and the mean points for each person) for each is reduced; therefore the expected value of the process variance is reduced.

If a shot is observed in the same place as before (somewhere between A' and E'), it is much more likely that the shot was fired by A', and the next predicted point will lie much closer to A' and much farther from E' than our previous prediction lay relative to A and E. (In a strict Bayesian analysis approach, the predicted point would probably lie even farther from E' than the observed point.)[11] Hence, the credibility attached to a single observation is increased when the process variance is decreased. Note that the variance of the hypothetical means, which is equal to the mean of the squares of the distances between E and A, B, C, and D, is unchanged between Figure 1 and Figure 2.

Now let us consider an example where the process variance is identical to that in Figure 1, but the variance of the hypothetical means is changed. Figure 3 shows such an example. In this case we can assume that our original persons A, B, C, and D are again shooting, but they are aiming for different points. We will call them (and their means) A'', B'', C'' and D'', to distinguish this example from the others.

In Figure 3, the clustering of shots around each of the means is similar to that in Figure 1, but the means are much farther apart, hence much farther removed from the population mean, E''. The variance of the hypothetical means will be much larger than in Figure 1. If a shot is observed somewhere between A'' and E'', it is more likely to have been fired by A'', so the predicted point will lie relatively closer to A'' than the predicted point in Figure 1 was to A. In other words, the credibility of the single observation is increased.

To this point, we have only examined the results of a single fired shot. If a number of shots were fired, the credibility attached to the mean of the observed shots would be greater than that for the single observation.

[11] This agrees with Hewitt's observation ("Credibility for Severity," p. 150) that the Bayesian resultant does not necessarily lie between the hypothetical mean and the observed result.

FIGURE 2

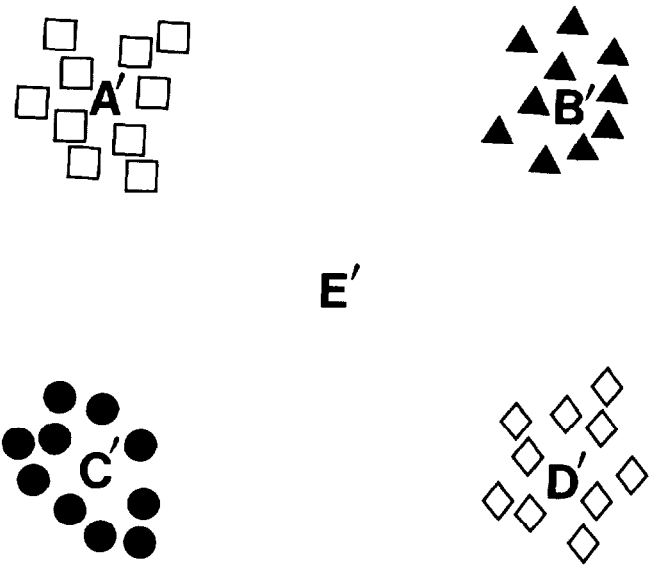
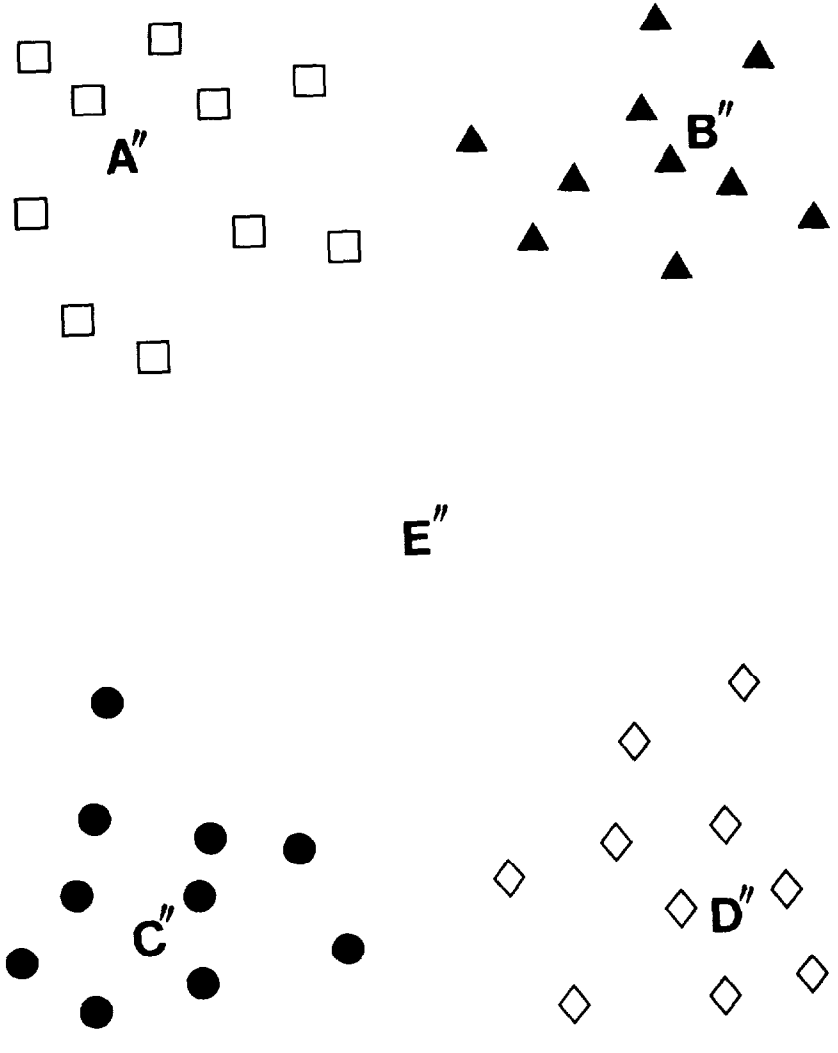


FIGURE 3



To summarize, it has been shown that, when projecting the location of a future shot based on current information and prior knowledge, the credibility attached to the current observations will increase with:

- Increasing number of observations,
- Decreasing process variance, and
- Increasing variance of the hypothetical means.

Finally, it will be helpful to analyze what happens at the extremes—as the three basic elements approach either zero or infinity. If no current observations are made, we have to rely totally on the prior information. This is equivalent to stating that $1 - Z = 1$, which implies that $Z = 0$. As the number of current observations goes to infinity, the pattern of shots will begin to resemble one of the four clusters, and the mean will tend toward the mean of the cluster. At the limit, the weight associated with the observed pattern will become one, hence, $Z = 1$ as n goes to infinity.

As the process variance goes to zero, the clusters will tend to shrink to single points. In terms of our example, we say that the marksmen are becoming better shots. At the limit, each of the four marksmen can hit the exact center of the target with every shot. The observation of a single shot will be sufficient to identify the marksman, and the next shot can be predicted with certainty. Hence, the credibility associated with the current observation goes to one as the process variance goes to zero. Conversely, as the process variance increases without bound, the clusters of shots tend to spread apart, and overlap one another. The observation of a single shot provides little information as to the identity of the marksman firing the shot, and the best estimate of the next shot will remain E. As a consequence, as the process variance goes to infinity, the credibility goes to zero.

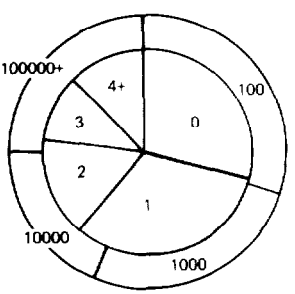
As the variance of the hypothetical means goes to zero, the clusters tend to move closer together. At the limit, each of the respective means coincides. When this happens, the observation of a shot will add nothing to our knowledge; the weight, or credibility, given to this observation will be zero. Increasing variance of the hypothetical means has the effect of moving the clusters apart. When they are sufficiently distant, a shot fired by one of the marksmen can be uniquely associated with one of the clusters. This situation also points out the difference between a pure Bayesian approach and Bayesian credibility. Using Bayesian credibility, the best estimate of the location will be the observed shot, because it has been given credibility equal to one. A pure Bayesian approach would select the mean of the cluster to which the observed shot is closest, rather than the position of the shot itself.

HEWITT REVISITED

In the following examples, the assumption is made that the process which creates losses can be modeled as a collection of spinners with inner and outer sections.[12] The universe is represented by the total collection of spinners and each individual risk corresponds to a single spinner. The inner portion of the spinner will be used to simulate the frequency of the risk and the outer section will be used to simulate the severity. An accident year consists of the selection of one (or more) of the risks (spinners), possibly at random, spinning once, observing the inner value (frequency), and spinning that many additional times, with each observation of the outer ring constituting a loss.

Figure 4 is a typical risk in the universe. In this risk, the probability of having exactly 0 claims is approximately 1/3, the probability of having exactly 1 claim is approximately 1/3, the probability of having exactly 2 claims is approximately 1/6, etc. For each claim, the probability of each of the possible severities equals the area corresponding to each value. (It would be trivial to extend to a continuous severity, but for simplicity, we will stick to the discrete case.)

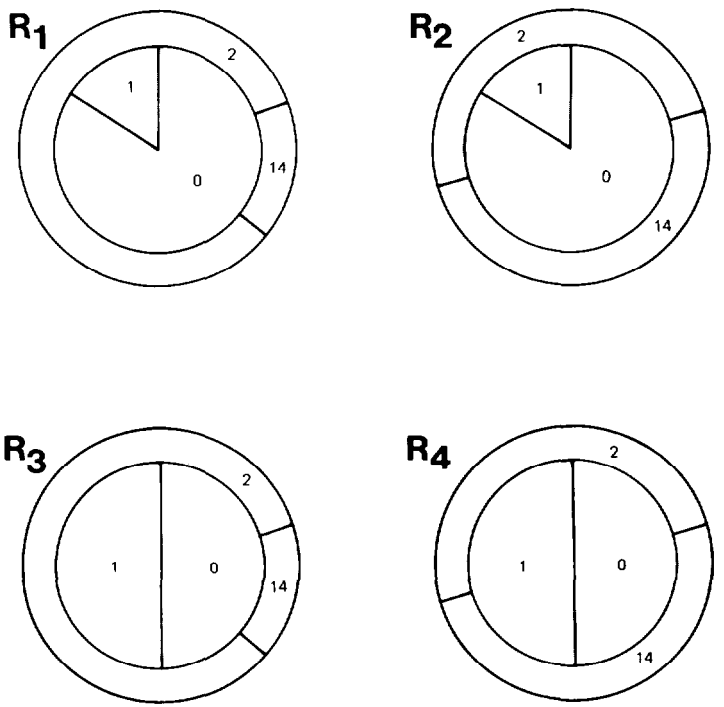
FIGURE 4



[12] The reader will notice the similarity between this and the examples used by Hewitt in "Credibility for Severity." This is intentional; his example is an excellent tool for explaining concepts. This paper is intended to expand on those ideas and provide additional insight into credibility concepts

In Figure 5 four spinners represent a universe of risks. The values associated with the areas in Figure 5 are either 1/6, 1/2 or 5/6. The universe has exactly one of each of the R_i and each has an equal probability of being chosen in a random sample.

FIGURE 5



This paper will take as assumptions the definitions and assumptions in Hewitt's "Credibility for Severity."[13] Briefly, a compromise estimate is chosen as a function of prior information (hypotheses) and current observations, according to the formula

$$C = ZR + (1 - Z)H$$

where R equals the mean of the observations,
H equals the mean of the hypotheses,
C equals the value of the compromise, and
Z is obtained from the formula $Z = n/(n + K)$.

The volume of observations is measured by n (number of trials or exposure units) and K is the K defined by Bühlmann:

$$K = \frac{\text{Expected value of process variance}}{\text{Variance of the hypothetical means}}.$$[14]

Although the terms "prior information" and "current observations" will generally be used, it should not be assumed that the two sets of data are necessarily different in time. For example, when calculating a class rate, the prior information might be the entire state (or countrywide) pure premium indication, and the current observations could be the specific class indications.

Further, it is important to note that the assumption is made that all parameters concerning the universe are known, although the identity of a particular risk chosen at random is not necessarily known. This implies that no parameter risk is involved in the example, only process risk.

The problem to be solved can be stated as follows. The universe has been described and all its parameters are known. From this information, hypotheses can be made regarding the correct premiums to be charged. A risk or risks are selected at random and observations of their experience are made. How can the prior knowledge (represented by the hypotheses) and the posterior knowledge (represented by the observations) be combined? From the point of view of an insurance company, assume that the universe, as defined, represents the universe of all possible insurable risks. Although there are four distinct types of risks (in terms of their loss process parameters), we will assume that these risks cannot be separately identified by any *a priori* characteristics other than historical experience. Hence, we can assume that there is only a single classification for

[13] Hewitt, op. cit., p. 149.

[14] Hans Bühlmann, "Experience Rating and Credibility," *The Astin Bulletin*, Volume IV, 1967.

the insurance company. The company must determine a rate and select a risk at random. The identity of the risk (in terms of its parameters) will remain unknown, although the actual experience will be observed. Based on the prior knowledge (of the universe) and the actual observations of this risk's experience, we wish to determine what is the best choice of a rate for *the same risk* in the subsequent year. (For convenience we ignore any timing problems related to calculating a new rate *after* observing the experience but *before* the new year commences.) It is vitally important to understand that the derived rate is applicable to the risk creating the experience, not to a new risk chosen at random.

Although it has been stated that the identity of the risk is not determinable, it will be instructive to first examine what action would be taken if the identity of the risk could be determined.

Assume that a risk is chosen and you know that you have selected R_1. You still could not predict with certainty the *actual* outcome of the loss process although you could calculate a mean value and a variance about the mean. If you had selected R_3, (and knew its identity) you could also calculate a mean and a variance but these would differ from the mean and variance of R_1. If you knew which risk you had, the choice of pure premium would be straightforward. You would set it equal to the mean of the risk.[15] However, suppose you chose a risk at random from the population. Before observing any loss experience of this risk, you would set the pure premium equal to the average of the means of the R_i, which is the same as equating it to the population mean.

There is an important difference between the two situations. In the first, where the identity is known, the actual experience will not exactly equal the expected in any one year, but over a long period of time, the average experience will tend toward the mean of the risk. In the second, the actual experience also will not reproduce the population mean, but, over the long run, the average experience will tend to the particular risk's mean, *not* to the universe's mean.

With enough observations of the risk experience, the cumulative mean of the observations will become arbitrarily close to the theoretical mean of that risk, and that value will be used for the pure premium. Before any observations are made, the mean of the population will be the best choice for the pure premium.

Credibility is concerned with the choice of the "best" pure premium based

[15] In order to make the example less complicated, risk loadings are being ignored. However, the extension of the example to a risk loading should be straightforward depending on one's choice of a risk measure.

upon a body of prior knowledge and a limited body of observations. It may be helpful to think of credibility in terms of the value of information. The prior knowledge has a certain amount of information about the "proper" pure premium and the actual observations also contain information. Credibility is concerned with the efficient blending of the information from the two sources.

Let us now examine our example in more detail. First calculate the pure premium that we would use for each R_i if we knew which R_i we had chosen. The mean pure premium is the product of the mean frequency and the mean severity. Thus, for R_1, the mean frequency is $1/6 \times 1 + 5/6 \times 0 = 1/6$, and the mean severity is $1/6 \times 14 + 5/6 \times 2 = 4$. Therefore, the mean pure premium is $1/6 \times 4 = 2/3$. Each of the others is calculated similarly. The details are shown in Table 1.

TABLE 1

	Frequency	Severity	Pure Premium
R_1	$1/6 \times 1 + 5/6 \times 0 = 1/6$	$1/6 \times 14 + 5/6 \times 2 = 4$	$1/6 \times 4 = 2/3$
R_2	$1/6 \times 1 + 5/6 \times 0 = 1/6$	$1/2 \times 14 + 1/2 \times 2 = 8$	$1/6 \times 8 = 4/3$
R_3	$1/2 \times 1 + 1/2 \times 0 = 1/2$	$1/6 \times 14 + 5/6 \times 2 = 4$	$1/2 \times 4 = 2$
R_4	$1/2 \times 1 + 1/2 \times 0 = 1/2$	$1/2 \times 14 + 1/2 \times 2 = 8$	$1/2 \times 8 = 4$

Assume that a risk is chosen at random and that risk is R_1 (although its identity is unknown to the observer). As observations are made, their cumulative average will tend to 2/3, although the average may vary significantly from 2/3 for the first few observations. As an example, the string of observations in Table 2 was randomly generated from R_1. Even after 10 trials, it is not clear that the choice of risk was R_1, because the cumulative average is greater than that expected for even R_2.

TABLE 2

Trial Number	1	2	3	4	5	6	7	8	9	10
Observation	0	0	0	14	0	0	0	2	0	0
Cumulative Average	0	0	0	3.50	2.80	2.33	2.00	2.00	1.78	1.60

Consider a different process as shown in Figure 6. The expected value of R'_1 is calculated in Table 3. Note that R'_1 has the same expected pure premium as R_1. Table 4 shows a string of observations from R'_1.

FIGURE 6

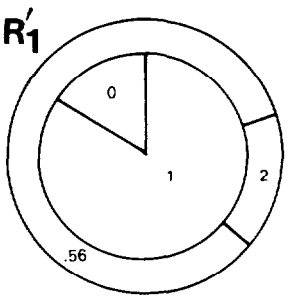


TABLE 3

	Frequency	Severity	Pure Premium
R'_1	$5/6 \times 1 + 1/6 \times 0 = 5/6$	$5/6 \times .56 + 1/6 \times 2 = .8$	$5/6 \times .8 = 2/3$

TABLE 4

Trial Number	1	2	3	4	5	6	7	8	9	10
Observation	.56	0	2	.56	.56	.56	.56	2	.56	.56
Cumulative Average	.56	.28	.85	.78	.74	.71	.69	.85	.82	.79

This time, it is much more obvious that we have chosen R'_1 rather than R_2, R_3 or R_4. Why should this be so, if R'_1 and R_1 have the same expected value? The answer lies in the variance of the process, appropriately called process

variance. Let us now calculate the variance of the process for R'_1 and R_1. The formula for the variance of a compound process is

$$\sigma^2 = \text{E(Frequency)} \times \text{Var(Severity)} + \text{Var(Frequency)} \times \text{E}^2\text{(Severity)}.$$

Each of the components is a binomial process and can be calculated easily as shown in Table 5. Substituting these values into the formula, the process variance for the two risks is as follows.

R_1 $1/6 \times 20 + 5/36 \times (4)^2 = 5.56$
R'_1 $5/6 \times (.288) + 5/36 \times (.8)^2 = .329$

The process variance for R'_1 is much less than for R_1 which coincides with our observations.

TABLE 5

	R_1	R'_1
E (Frequency)	$(5/6) \times 0 + (1/6) \times 1 = 1/6$	$(5/6) \times 1 + (1/6) \times 0 = 5/6$
Var (Frequency)	$(1/6) \times (1 - 1/6) = 5/36$	$(5/6) \times (1 - 5/6) = 5/36$
E (Severity)	$(5/6) \times 2 + (1/6) \times 14 = 4$	$(5/6) \times .56 + (1/6) \times 2 = .8$
Var (Severity)	$((5/6) \times 2^2) + (1/6) \times 14^2 - 4^2 = 20$	$((5/6) \times (.56)^2) + (1/6) \times 2^2 - (.8)^2 = .288$

Roughly speaking, the process variance tells us "how far apart" the actual results can be for each trial. The smaller the variance, the closer the actual results will be to each other and to the expected. This is just another way of saying that the confidence interval around the expected (for a given probability) will be smaller for small variances.

We are now at the watershed between "classical" credibility and Bayesian credibility. Classical credibility continues down the road of confidence interval analysis, making various assumptions about the form of the distribution, deciding whether to include the claim severity or ignore it, and calculating the appropriate number of claims necessary to ensure that, with probability P, the actual claims (numbers or dollars) will be within $100K\%$ of the expected. The analysis then continues by arbitrarily assigning 100% credibility to this resulting value and exploring various ad hoc measures to calculate partial credibility. To the extent that the observations are not "fully credible," the complement of the credibility is assigned to the prior knowledge.

When a body of information receives less than 100% credibility, the implication is that the data is not "good" enough, that the variations from expected are too large to be acceptable for ratemaking. But when the complement of the credibility is assigned to prior knowledge, there is no discussion of whether the result of the combination is "good" enough in terms of the standard. And when data is assigned 100% credibility because the standard is met, there is no discussion of the fact that the result could be further improved with *some* weight assigned to prior knowledge.

We will now explore the path leading to Bayesian credibility. As we have seen, the process variance will tell us to what extent the actual results will tend to cluster around the mean. Another measure critical to the concept of credibility is the variance of the hypothetical means.

The hypothetical means are the expected values for each of the R_i. They have already been calculated as the pure premiums in Table 1. The variance of these values can be easily calculated; the result:

Variance of hypothetical means = 14/9.

The variance of the hypothetical means is a measure of the spread of the means—how far apart the means are from each other.

When we were examining the effects of process variance, we looked at two situations in which the process variances were different but the variances of the hypothetical means were the same. We found that we were more certain of the identity of the actual risk when the process variance was smaller. Now we will examine two situations with identical process variances but different variances of the hypothetical means.

The first situation will be the same as before; that is, the universe contains R_1, R_2, R_3, and R_4, and the same string of observations is randomly generated by R_1. In the second situation, we consider a universe containing R_1, and three new risks R'_2, R'_3, and R'_4. We assume that these new risks have the same process variance as their counterparts, but the pure premiums are 10, 20 and 30 respectively. The variance of the hypothetical means is now approximately 120. We can be more certain that the string of observations are generated by R_1 in the second situation, than in the first situation.

We have now demonstrated that it is easier to discern the true identity of an R_i chosen at random when

- the process variance becomes smaller, and
- the variance of the hypothetical means becomes larger.

Additionally, it is easier to discern the identity of the R_i as the number of observations increases.

Much of the emphasis has been placed on the ability to discern the true identity of the R_i. It is not necesssary to be absolutely certain of the true identity of the R_i; it is only necessary to change the *a priori* estimates of the probabilities of each of the R_i (this establishes the link with Bayesian analysis). With relatively small process variance and/or large variance of the hypothetical means, the *a posteriori* estimates of the probabilities of each R_i can be modified significantly from their *a priori* values. Because we presume that we have knowledge of the mean of the universe, and the mean of each of the risks, and are ignorant only of the actual identity of the R_i, it is that knowledge that will lead us to a better estimate of subsequent loss experience.

To recap, credibility should increase with

- the number of observations,
- decreasing process variance, and
- increasing variance of the hypothetical means.

The derivation of the proper function relating these variables can be found elsewhere and will be stated here without proof. Considering the complexity of the concepts, it is a remarkably simple formulation:

$$Z = n/(n + K)$$

where n is the number of observations,

$$K \text{ is } \frac{\text{expected value of the process variance}}{\text{variance of hypothetical means}}, \text{ and}$$

Z is the resulting credibility.

EXPERIENCE RATING CREDIBILITIES

It is interesting to compare the development of the credibility formulas used in experience rating with those used in ratemaking. The formula used in the worker's compensation experience rating plan has been essentially of the form $E/(E + K)$ since 1918.[16]

[16] Paul Dorweiler, "A Survey of Risk Credibility in Experience Rating," *PCAS* XXI, 1934 or *PCAS* LVIII, 1971.

Although various changes have been made to the formula to reflect different treatments of normal versus excess losses, the form has remained basically unchanged. It is also noteworthy that the formula is a function of *expected* losses, rather than *actual* losses, as is generally the rule for the ratemaking formulas. Although the form is the same as that discussed by Hewitt and Bühlmann, the value of K does not coincide with Bühlmann's derivation.[17,18]

According to Dorweiler, "the members of the committee, after consulting with underwriters, chose those curves which in their opinion produced the best results for the set of risks and thus established the constants K_1 and K_2 . . ." Later, the value of K was derived from an *ad hoc* selection of the "swing" of the plan.[19]

The concept of a Q-point and S-value, between which the value of Z no longer is calculated from $E/(E + K)$ but rises smoothly from $Q/(Q + K)$ to 1 as E varies between Q and S is not justified on theoretical grounds but can be justified on pragmatic grounds: for risks sufficiently large, the difference between the modifications resulting from the "correct" versus the *ad hoc* formula is insignificant and does not justify the additional computations necessary for the theoretically preferable formula. This pragmatic approach compares closely with the concept of full credibility described in Mayerson.[20]

Ratemaking formulas started with a formula for full credibility and then made adjustments to accommodate the need for partial credibility, whereas the experience rating formulas started with a formula for partial credibility and made adjustments to accommodate the practical need for full credibility. As trivial as this distinction may sound, it turns out to explain many of the historical problems with classical credibility. Assuming one accepts the formula $Z = E/(E + K)$, which approaches but never reaches unity, any attempt to define a unique full credibility standard is doomed to failure. Moreover, because the full credibility standard will not be based on $E/(E + K)$, the derivation of partial credibilities consistent with this formula will be more difficult.

[17] Hewitt, op. cit.

[18] Bühlmann, op. cit.

[19] Francis S. Perryman, "Experience Rating Plan Credibilities," *PCAS* XXIV, 1937 or *PCAS* LVIII, 1971.

[20] Allen L. Mayerson, op. cit.

A MISCONCEPTION

One of the misconceptions surrounding credibility is that a large value of credibility is a desirable situation. This sounds quite reasonable. After all, why would you prefer a situation where the experience has low credibility to one where the credibility is high? However, as will be shown, a situation where experience has low credibility can be preferable to the high credibility situation. I believe that this misconception rests on the confusion of the terms "credibility" and "confidence." The two terms sound similar but have different meanings. Credibility in the familiar sense (as opposed to its technical meaning) *is* almost a synonym for confidence. However, "credibility" is used in some places where the term "confidence" is meant.

Since the difference between these two terms is so important, it is appropriate to set down definitions of the terms.

CREDIBILITY—The appropriate weight to be given to a statistic[21] of the experience in question *relative* to other experience.

CONFIDENCE—The likelihood that a statistic is close to the theoretical value.

Several observations are pertinent.

- Credibility is a *relative* concept while confidence is an *absolute* concept.
- Credibility naturally produces values between 0 and 1. Confidence measures are not as well-behaved. Traditionally, confidence is measured as a probability P that the true mean is within $100K\%$ of the observed mean.
- Credibility can be thought of as relative confidence. Even though the mean of a particular set of observations has a low measure of confidence, if the prior information also has a low measure of confidence, the credibility of the current set may be high.

In the second example, the universe was assumed to consist of a single classification containing four elements. Let us redefine the universe to include a larger number of elements that have been partitioned into classifications. Assume that one of these classifications is comprised of the four elements in the

[21] Recall that a statistic is simply a function of the observed values. Generally, we will be referring to the sample mean, but the concept of credibility should generalize to other statistics (with appropriate changes in the calculation of K).

original example. The credibility of a single observation can be calculated as follows.

$$Z = n/(n + K)$$

where $n = 1$,

expected value of process variance[22] = 154/9, and

variance of hypothetical means = 14/9;

thus $Z = \dfrac{1}{1 + \dfrac{154/9}{14/9}} = \dfrac{1}{1 + 11} = \dfrac{1}{12}$.

Suppose the original class of R_1, R_2, R_3, and R_4 is replaced by R_1, R'_2, R'_3, and R'_4. The expected value of the process variance is unchanged, but the variance of the hypothetical means is now 120 and

$$Z = \frac{1}{1 + \dfrac{154/9}{120}} \approx \frac{1}{1 + .143} \approx .88.$$

With our new classifications, we have 88% credibility for a single observation compared to 8% for the old classifications. Does this indicate a preferable situation? Absolutely not. The credibility is high because the new classification is much less homogeneous than the old one; the hypothetical means are much farther apart. The confidence surrounding the classification mean is extremely low. The absolute confidence of the observations has not changed, but the relative confidence has increased. Credibility is high, not because the sample information is so "good," but because the prior information is so "bad."

Does this mean low credibility is always desirable? Of course not. To understand when high credibility is desirable and when it is not, it will be helpful to examine our universe more closely.[23] Typically, our universe is composed of a number of classifications, each of which contains a number of

[22] Although this value was not explicitly calculated in this paper, it is straightforward and can be calculated easily, or the reader may refer to Hewitt, "Credibility for Severity," p. 158.

[23] Here, as before, the term "universe" is used in the mathematical sense. It includes not all possible things, but the entire set of items relevant to the question at hand. For example, if the question concerns automobile liability ratemaking, the universe would include experience relevant to automobile liability, but not homeowners experience.

individual risks so that our structure has three levels: risk, class, and universe.[24] There is an important distinction among these three levels. While there is generally no latitude as to the definitions of the universe or the individual risks, we are free to aggregate the individual risks into classes as we wish. Once a particular class plan has been chosen, there are two major uses of credibility:

- RATEMAKING—Ratemaking consists of two major steps. First, a new mean pure premium for the universe is calculated by credibility weighting the indications of the most recent data with the pure premium presently in use.[25] Second, the indications of each class are credibility weighted with the new mean premium for the universe to derive the new pure premiums for each class.
- EXPERIENCE RATING—Experience rating, or individual risk rating, consists of credibility weighting the actual experience of the risk with the pure premium of the particular risk's class.

The ultimate goal for each individual risk is a rate which is as close to the true mean of the risk as possible. Because the experience of each of the individuals is fixed, and, equivalently, the overall experience is fixed, the only variable is the class plan. Creation of a class plan is equivalent to a stratification of the universe of risks; hence the ideas in Lange's paper, "Implications of Sampling Theory. . .," are applicable. In this paper, he discusses a desirable property of a stratification, namely, that the resulting strata should be as homogeneous as possible.[26] Although Lange's immediate goal was to improve the estimate of the overall mean, while we are interested in the pure premium for the individual risk, the goals are consistent.

If we have homogeneous strata, or classes, then the experience of the individual risks within a class will be similar to each other, hence close to the class mean. But with homogeneous classes, the means of the various classes tend to be "farther apart" from each other than if we have non-homogeneous classes.[27] When considering the use of credibility in ratemaking, a credibility

[24] The situation where two levels of classifications exist (as distinct from a two-way class system), such as in Workers' Compensation, where individual risks make up classes that are aggregated into industry groups, is slightly more complicated and will not be addressed here.

[25] With appropriate adjustments for trend, development, etc.

[26] Jeffrey T. Lange, "Implications of Sampling Theory for Package Policy Ratemaking," *PCAS* LIII, 1966, p. 288.

[27] If this is not obvious, consider a class plan where individual risks are randomly assigned to classes. These classes will be quite non-homogeneous, and the experience of each class will tend to approximate that of the universe; hence the classes will tend to be "close to each other."

for each class must be calculated. Since in a "good" class plan, the experience of risks within a class is similar, the process risk (or "within" variance) will be smaller than if the classes were not homogeneous. Also, the variance of the hypothetical means (the "among" variance), which is the variance of the various classes, will be higher for a class plan with homogeneous classes. Hence, the calculation of K (to be used for credibility) for a class plan with homogeneous classes will result in a relatively small value of K, since

$$K = \frac{\text{expected value of process variance}}{\text{variance of hypothetical means}} .$$

A small value of K implies high credibility based on the formula

$$Z = n/(n + K).$$

So we see that, with a "good" class plan, the credibility of the class experience will be higher than for a poorer class plan.

The situation is different for individual risk rating. Here we are credibility weighting the individual experience with the class experience. The process variance refers to the variance of the individual risk's experience, while the variance of the hypothetical means is the variance of the means of the individual risks within the class. If a class plan is created that has a very non-homogeneous class, then the variance of the hypothetical means will be large, making K small, resulting in large credibilities for individual risk experience.

In summary, we desire a class plan with homogeneous classes, which results in classification experience that has *high* credibility, but individual risk experience with *low* credibility. The relatively low credibility assigned to the experience of a single car is not a cause for concern, but an indication that the class is doing a relatively good job.[28]

HISTORICAL PERSPECTIVE

The simple chart in Table 6 may be helpful to an understanding of the relationships among some of the articles on credibility.

[28] Robert A. Bailey and LeRoy J. Simon, "An Actuarial Note on the Credibility of Experience of A Single Private Passenger Car," *PCAS* XLVI, 1959, p. 159.

TABLE 6

	FREQUENCY ONLY	PURE PREMIUM
CLASSICAL	Longley-Cook	Mayerson, Jones, Bowers
BAYESIAN	Mayerson	Hewitt

This partitioning is reasonably accurate; some of the exceptions are:

1. Longley-Cook does suggest ways to handle the pure premium but does not go into detail.
2. Longley-Cook states the importance of the prior information but does not utilize it in any formulas.
3. Mayerson summarizes the classical view.

The introduction of this paper contains several apparently inconsistent statements regarding credibility. Much of the confusion surrounding credibility arises from two sources:

- primary focus on the properties of the current observations, and
- an attempt to tackle the full credibility standard before the partial credibility standard.

Longley-Cook stressed the importance of the value of prior information. But his statements were not motivated by the same reasons that caused us to examine the statistical properties of the prior distribution. In his example, he concluded that Oregon fire data is inappropriate for New York ratemaking, not because of the arguments discussed in this paper, but because of the lack of applicability to the existing problem.

The development of classical credibility is closely tied to the traditional concerns regarding the proper balance between responsiveness and stability. Large weights given to the more recent data, or to the specific class data, will tend to increase responsiveness and decrease stability. In addition, it was correctly perceived that it is easier to defend a rate when the data used to make the rate is "local," in terms of time, geography, or class. These considerations quite naturally led to the attempt to assign the maximum weight possible to the current observations, subject to a stability restriction. The calculations of classical credibilities outlined by Longley-Cook follow directly from these arguments. In addition, Arthur Bailey's limited fluctuation credibility and greatest

accuracy credibility[29] provide the link between the responsiveness/stability argument and Longley-Cook's statements regarding prior information.

The issue, related to the use of a standard based on frequency but applied to the pure premium, was a bit of a dilemma. Based on the foundations of classical credibility, it was difficult to refute the arguments for consideration of severity on theoretical grounds. But to include severity would cause practical problems. If the stability constraints were unchanged, the standards would be considerably increased. This would reduce responsiveness and increase the weights needed for "external" data. In addition, most actuaries felt that the existing standards were fairly reasonable. In theory, the stability constraints could be altered so that the same standards would result. But then the actuaries would be in the uneviable position of trying to justify to management and regulators a major change in approach which creates an insignificant change in results. Although the subject continued to receive theoretical attention, the ratemakers took the only practical course—they ignored the issue.

Bayesian credibility provides solutions to some of the problems associated with classical credibility, but at a cost: it is not trivial to understand, nor easy to apply in practice. Hewitt's reviews and his paper have contributed significantly to this subject. Hewitt's first list of three critical variables is mentioned in this paper because it was quoted by Lange in "Application of a Mathematical Concept of Risk." Hewitt clears up any misconception arising from this list, but this clarification is contained in a footnote of a review and might be missed.[30] The reader is urged to read this footnote carefully. The second list, which is consistent with this paper, applies to the more general case.

SUMMARY

The use of classical credibility has served the actuary well for many years. But the increased refinement of actuarial science requires that we turn to the theoretically preferable Bayesian credibility. The increased scrutiny of our methods requires that we be able to defend and explain our methods. It is hoped that this paper has contributed to the understanding and explanation of credibility concepts.

[29] Arthur L. Bailey, "Sampling Theory in Casualty Insurance," *PCAS* XXIX, 1942, p. 50 and *PCAS* XXX, 1943, p. 31.

[30] Hewitt, Discussion of "On the Credibility of the Pure Premium," p. 78.