Journal of Scientific Exploration, Vol. 20, No. 1, pp. 23–38, 2006 0892-3310/06

Pure Inference with Credibility Functions

MIKEL AICKIN

Program in Integrative Medicine and
Department of Family & Community Medicine
University of Arizona
4840 N. Valley View Rd.
Tucson, AZ 85718
e-mail: maickin@comcast.net

Abstract—The vast majority of scientific inference is carried out using frequentist methods, and the only apparent challenger to the dominance of these methods is Bayesianism, which has yet to achieve a foothold of acceptance. The approach to inference based on direct use of the likelihood function, which I call "pure inference", is far less well known, and virtually unused in practical work. This is unfortunate, because pure inference has some considerable advantages over both frequentist and Bayesian approaches. The purpose of this article is to introduce the concept of credibility as a way of doing pure inference, and through the simplest realistic examples, to show how credibility functions can be computed, manipulated, and interpreted.

Keywords: statistics—likelihood function—binomial—Bayesianism—frequentism

Introduction

Scientists do not generally want to get into issues of inference, because the issues are messy, depend on a lot of math, and are frankly rather boring. As a consequence, although inference belongs to all science, it has been taken over by professional statisticians, as if it were their exclusive domain. The tools that they have fashioned have developed into a web of conventional belief and behavior, to the point that after learning how to manipulate the symbols of statistical inference, many scientists forget the principles on which those manipulations rest. With the passage of time, the ability of science professions to carry out excellent inference slowly degrades, as their understanding of statistical methods becomes more mythic and less based on rational analysis.

In my opinion this has been the history of frequentist inferential methods in the previous century. In the last half of that century, the Bayesians have mounted any number of assaults on the frequentist bastions, and continue their endless, premature announcements of victory. Whatever benefits they might bring remain unrealized, and probably will remain so. Again in my opinion, there is some hope. A very small number of statisticians have gently, and unsuccessfully, pushed forward an idea that I would call "pure inference". In statistical jargon, it is

called inference based on the norrned likelihood function, and in some avatars ‘‘profile likelihood'' inference. The situation has become somewhat confusing, in that some uses of the likelihood function are actually completely conventional frequentist approaches. These are not the developments I am discussing and are not what I will present below.

One of the first general books on this topic was Edwards (1972), who simply asserted that direct use of the likelihood function is sensible. The likelihood function is simply the probability density of the statistical model, evaluated at the observations, and therefore is a function of the parameters alone. An important epidemiology text (Clayton & Hills, 1993) also took this basic point of view, but it would be an exaggeration to say that it has had any impact on the practice of epidemiology. In a reprise of Edwards, Royall (1997) argued for direct use of the normed likelihood function, but he did not extend his clarifying arguments as far into the multivariate cases as he could have. More recently, Sprott (2000) started a book making the argument that pure likelihood inference is more scientific than either frequentist or Bayesian approaches, but then in a disappointing turnaround devoted the second half of his book to a relapse into frequentist-based thinking. I do not claim that this is anything like an exhaustive list of recent works in this area, but instead I want to set the stage by saying that fundamentally the ideas I present below are not completely new; they are simply underappreciated, and therefore massively underused.

Although I believe that pure inference based on credibility offers considerable advantages over both frequentist and Bayesian thinking, I will not complete this argument here. Instead, my purpose is limited to an explanation in the simplest possible cases of what credibility means, and how credibility functions can be computed and interpreted in the case of simple inferences. I will start by showing the relationship between probability and credibility in a hypothetical clinical situation. I will then develop credibility inference for the ''binomial'' experiment, while trying to explain why the credibility function makes sense. The case of comparing two binomial experiments is more realistic, but somewhat more complex, and so I will try to give a careful explanation of this situation. I will briefly remark on a two further topics, the introduction of evidence external to an experiment and large sample properties of credibility functions, and I will finish with an opinionated summary.

The Concept of Credibility

We start with a hypothetical example, since real examples always bring in extraneous points that make it hard to focus on the issue at hand. Suppose that on the basis of an enormous amount of data, Table 1 gives the probability of a patient being satisfied with a clinical visit, in terms of the gender of the patient and the physician. Now consider a case where the patient expressed

TABLE 1
Probability of Satisfaction

	Physician male	Physician female
Patient male	.53	.15
Patient female	.31	.87

satisfaction: The task is to infer the gender of either the patient, physician, or both, based solely on this information.

From the standpoint of pure inference, we rest on one axiom: Probable things happen more often than improbable things. In terms of Table 1, this tells us that satisfaction will happen more often in female-female cases than in any other case. It also ranks the other situations in the order of their probabilities, so that we would least expect to see satisfaction with male patients visiting female physicians. This reasoning goes from the conditions (genders) to the outcome (satisfaction).

Inference goes the other way, from the outcome (satisfaction) to the conditions (genders). Pure inference continues the above point of view here with respect to probabilities, that a condition is a better explanation for what we observe to the extent that it gives our observation a relatively high probability. For technical reasons, the absolute values of probabilities are not important for inference, only relative values are. Thus, it does no harm to divide all the probabilities by their maximum (Table 2).

The entries in Table 2 are not probabilities, they are credibilities. They pertain to the genders, after we know the patient has expressed satisfaction. We say it is fully credible that both the patient and physician were female, because that condition has the largest credibility. This does not mean that female-female is certain, only that it provides the best explanation for our observation. Thus, on the credibility scale, 1.00 means fully credible, not true or certain. Conversely, we say it is not very credible that the patient was male and the physician was female. Note that we are not saying that this condition is impossible, just that it is not very credible, in the sense that there is another explanation (female-female) that gives about six times as much probability to what we observed.

Credibility extends to other statements about gender. The principle is that the credibility of a statement is the largest credibility of any individual condition that makes it true. For example, the statement "the genders are the same" is

TABLE 2
Credibility of Genders

	Physician male	Physician female
Patient male	.61	.17
Patient female	.36	1.00

made true by the male-male case and the female-female case, and has credibility 1.00, the maximum of the two. The statement that their genders differ is constituted from the other two cases, and its credibility is 0.36, again the maximum of the two. The statement that the patient is female is constituted of the bottom row and has credibility 1.00, while the credibility of a male patient is 0.61. Similarly the credibility that the physician is female is 1.00, and the credibility of a male physician is 0.61.

Pure inference thus makes statements about explanations based only on the observations that were made. In the above example, we made no use of how frequently patients are female, nor how frequently physicians are female, in the clinic where we made our observation. We can take this additional information into account, as I will explain below, but we do not have to, and we did not in this example.

The key idea is, therefore, that probability and credibility are dual concepts. Probability expresses the propensity for an event to happen, as it depends on the conditions that prevail. Credibility grades the conditions that might have prevailed in terms of how well they explain an observation that was made.

The Simplest Example of a Credibility Function

Imagine an experiment in which there are "n" opportunities for some target event to occur. We are concerned with "p", the probability that target event occurs. In this case, p is the ***parameter*** in a model for how the event (target or its opposite) occurs. The experiment itself consists of recording that "t" target events occurred in the n opportunities.

It is not possible to make any progress on inference about p without some assumptions. The conventional assumptions in this case are (1) the probability p does not change from one opportunity to the next, and (2) the different opportunities are independent of each other. These assumptions imply that the probability of observing t target events in n opportunities, in the order they occurred, is equal to

$$p^t(1-p)^{n-t}.$$

In a general notation, this quantity would be denoted pr(t:n,p), and read as "the probability of observing t target events in n opportunities when p is the probability of a target event". The notation is intended to show the results (t) to the left of the colon, and the conditions (n, p) to the right. The independence assumption is critical here, and probably should be included in the notation, but it is conventionally just assumed from the context.

Our general understanding of how probability works says that highly probable things happen more frequently than improbable things. The pure inference approach turns this around by saying that values of p that give high probability to what we actually observed are better explanations for our observations than are values of p that give low probabilities.

One might think, therefore, that $\mathrm{pr}(t : n, p)$ by itself can be used for pure inference about the true value of p. The problem is that as the number of opportunities grows (n gets larger) the probability gets smaller and smaller. Eventually we find ourselves in the position that no matter what happens, the probability that it would have happened is tiny. In order to find our way out of this problem, let us remember that inference amounts to saying something about which values of p are reasonable explanations for our observations. Our approach would then be to say that if one value of p gave much higher probability to our actual observations than another value of p did, then the former would be preferred over the latter as an explanation. This would be true regardless of the magnitudes of the individual probabilities.

One way of accomplishing this is to identify the value of p at which $\mathrm{pr}(t : n, p)$ is a maximum, and then divide any other value by that maximum. This leads to the definition of the *credibility function*

$$\mathrm{cr}(p : n, t) = \frac{\mathrm{pr}(t : n, p)}{\max \mathrm{pr}(t : n, p)},$$

Unlike probability, the credibility function puts the parameter of interest (p) to the left of the colon, and the conditions and observations to the right (n, t). This notational shift reflects the shift in thinking. Probability pertains to outcomes occurring under conditions, one of which is the value of a key parameter. Credibility pertains to the key parameter, in a context of conditions and observations. This reversal is the fundamental operation of inference.

In practice, we compute the credibility function by finding the value of p that produces the maximum in the denominator, denoted $\hat{p}$, so that

$$\mathrm{cr}(p : n, t) = \frac{\mathrm{pr}(t : n, p)}{\mathrm{pr}(t : n, \hat{p})} = \frac{p^t(1-p)^{n-t}}{\hat{p}^t(1-\hat{p})^{n-t}}.$$

A little bit of calculus shows that the maximizing value is $p = t/n$. Note that this is a satisfying result: The best explanation for the observations is that the probability of the target event is t/n, the fraction of target events that actually happened.

In a hypothetical example, suppose that we present a person with a series of sealed envelopes, each of which contains a card with a symbol on it. The person knows what symbols are possible, and says what symbol he or she thinks is in the envelope that is presented. This is done n times, and t is the number of times the person identifies the symbol correctly (the target event). Let us suppose that $n = 50$ and $t = 27$. We want to perform a pure inference about p, the probability that the person would correctly identify the hidden symbol. We want to do this because p is a measure of the ability of the person to detect the symbol by some other means than simply seeing it. To the extent that p is large we have evidence for this other way of seeing.

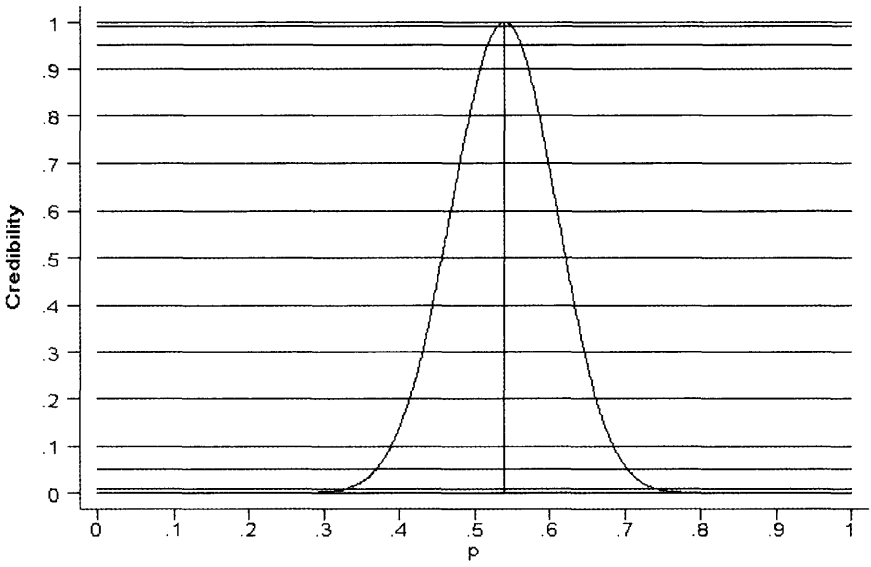

Fig. 1. Credibility function for target probability p, based on 27 target events in 50 opportunities.

Figure 1 show the credibility function for p, based on 27 target events out of 50 opportunities, cr(p : n = 50, t = 27). Obviously the peak value of 1 is attained only at p = 27/50. The credibilities of values near this are high, but the further one departs from this value, the credibilities fall off sharply. Note that the width of the credibility function is an expression of how much evidence there is. As the sample size increases, the credibility function becomes narrower and narrower, making the range of parameter values with high credibility smaller and smaller. When the sample size is small, the credibility function is wide, meaning that one cannot draw any sharp distinctions among parameter values based on little evidence.

Although computations are best performed by computer, it is perhaps useful to point out how that goes. Due to the sizes of the numbers involved, it is far better to use natural logarithms in the computation. Thus, the first step would be to compute

$$L = 27^* \ln(p) + 23^* \ln(1 - p) - 27^* \ln(27/50) - 23^* \ln(23/50)$$

and then the second step is

$$cr = \exp(L).$$

This would be done for a grid of values of p, which then could be plotted, used to look up specific credibilities, or interpolated to get credibilities for values of p not on the grid.

It is important to understand that in pure inference, everything we can say

inferentially about p is contained in the credibility function, or its graph in Figure 1. We identify the maximizing probability 27/50 as the best explanation for our observations, but we recognize that there are other explanations, values of p near 27/50, which also provide good explanations. We can also identify ranges of values that seem implausible, because their credibilities are so low. For example, the parameter values with credibilities below 0.01 are about 0.33 and below and 0.74 and above. We might say that these are incredible ranges, and behave as if they did not contain the true value of p.

We have seen that pure inference in this case is quite simple. It uses a model for the observations (independent opportunities with the same probability for the target event), and everything else follows directly from the model, with no additional structures or decorations. This is important, because the two competitors for providing inference in this case both add something else to the inference problem.

Conventional frequentist inference proceeds as follows. First, some value of p needs to be identified as the "null hypothesis" value. This means that a theory, usually opposed to the theory we would like to be true, says that some distinguished value p_0 would be the true value. For instance, in our symbol identification experiment, if there were three symbols, then the natural value of p_0 would be 1/3, because this would say that the person had no extra-sensory way of perceiving the hidden symbols, the opposite of what we would like to conclude. In this paradigm, statistical inference consists of a decision, whether the observations are sufficiently extreme for us to decide that 1/3 is not the true value (we "reject the null hypothesis"), or whether they are not so extreme (so we "confirm the null hypothesis"). Thus, the null hypothesis plays the role of an inferential attractor. The data must escape from the attractor by a sufficient degree for us to conclude anything. If they do not escape, we conclude nothing (in null hypothesis testing, confirming the null is actually a non-result, not a positive statement that the null is true). Placing an inferential attractor into the parameter space like this is the hallmark of conventional frequentist inference. It can, however, be considered a pre-observational bias, since it sets up a distinction among parameter values in the absence of any evidence whatsoever. Frequentists do not look on their approach this way, but from the standpoint of pure inference, it is precisely what they are doing.

There is, however, another frequentist approach, based on "confidence intervals". Basically, this is an interval computed from the data that has a certain high probability of containing the true value of the parameter p. Although this method does not put an inferential attractor into the parameter space, it does have several points of contact with the null hypothesis testing approach. First, it divides the parameter space (values of p) into two sets, those inside the confidence interval and those outside. Thus, it leads to an "inside/outside" distinction, just as null hypothesis testing leadings to a "reject/confirm" distinction. Second, the connection runs deeper. Those values inside the confidence interval are precisely the ones, if they had been null hypothesis values, that

would have been confirmed, and those outside would have been rejected. In the end, therefore, confidence intervals are a thinly-disguised reworking of null hypothesis testing.

The other approach, much less popular among practicing scientists, is Bayesianism. In a way, this approach is misnamed, because it is not actually based on the theorem of probability that bears Bayes's name. Instead, its novelty comes from the idea that parameters can have probability distributions. In frequentist inference, observations can have probability distributions, because probability is a technical way of fashioning a story about how chance observations can come about. It is often said that probability therefore measures uncertainty, because we are uncertain about which event will be observed. Bayesians take this metaphor literally. Because they are uncertain about the true value of the parameter p, they conclude that probability is the way to express that uncertainty. They see no distinction between uncertainty about parameters and uncertainty about observations. Frequentists do not share this view, and thus are unwilling to make probability statements about parameters, as a basis for inference.

In the Bayesian approach to our example, they would imagine a prior probability distribution over the possible values of the parameter p. This would somehow capture all of their beliefs about p just before they saw the observations (t target events in n opportunities). They would then use conventional probability calculations, treating the parameter p as if it were a chance observation, to come up with the "posterior" probability distribution for p, given the data and all previous information. How this probability distribution is subsequently used depends on the purpose of the inference, but this will not concern us here.

Thus, Bayesians are willing to use what I called $pr(t : n, p)$, but they inject the new element $pr(p)$, the probability distribution of p before the current experiment. In this sense, they introduce into the inference problem "something extra", just as the frequentists do. They claim that their approach is more virtuous than the frequentist, because they make their pre-experimental bias known, whereas the frequentists hide it in their null hypothesis attractor. From the standpoint of pure inference this is a quibble without consequence. The point is not who is more forthcoming about the unnecessary additions they bring to the inferential problem, but the fact that they bring those additions at all.

It is perhaps worth emphasizing that the pure inference approach is so simple and natural that it is hard to see that it accomplishes anything. Both frequentist and Bayesian inference approaches are much harder, largely due to the unnecessary additions they bring to the inference problem. Because they can pose hard questions and then solve them (for example, how do you find the shortest confidence interval that covers the true parameter with a fixed probability?), they seem to be accomplishing something. But again from the pure inference perspective, the mathematical pyrotechnics that are necessary to solve unnecessary problems do not add to the argument for the validity of the method. The fact that pure inference does not pose or solve such problems is a strength, not a weakness.

Credibilities for Sets of Parameters

I have suggested that we should look only at the credibility function for pure inferences. But this function appears to provide information only about individual values of the parameter. Thus, saying $cr(p = 113 : t = 27, n = 50) = 0.0096$ only makes a statement about $p = 113$. As we will see especially in the next section, we often want to make credibility statements about sets or ranges of parameter values. An example would be $cr(p \leq 215 : t = 27, n = 50)$, the credibility that the true value of p lies below 2/5, based on the observations.

The pure inference solution is to assign to a set of parameter values the maximum value of the credibility of any element of that set. Thus, $cr(p \leq 2/5 : t = 27, n = 50)$ would be computed by finding the largest value of $cr(p : t = 27, n = 50)$ for $p \leq 2/5$, which is 0.1365. This means the statement that a given set contains the true parameter value is equal to the largest credibility of any parameter value in that set. In other words, we find the credibility of a range of parameter values by finding an example in that range that maximizes credibility. (For purists, the credibility of a set is the supremum of the credibilities of its members, but I want to avoid this level of technicality here.)

The Meaning of Credibility

Both frequentist and Bayesian inferentialist have good metaphors for their versions of inference. The frequentist uses the decision metaphor, which is practical and understandable, but as I have argued above, it necessarily creates a bias. The Bayesian uses probability, which is widely if somewhat mistakenly held to be the only way to express uncertainty, and which again has the benefit of understandability. Through the prior probability distribution, the Bayesians insert their own version of bias into the inference problem.

An appropriate metaphor for credibility is the definition that some dictionaries give to the word "credibility". It is the "capacity for belief'. To say that a statement has high credibility does not mean that it should be believed. Instead it means that if one were to interpret the ambiguous or uncertain evidence to favor the statement to the maximum degree that is reasonable, then under this circumstance it would be believable.

In a sense, credibility can be used most forcefully for the reverse kind of inference. If the outside of a range of parameter values has rather low credibility, then even if we were to interpret the evidential ambiguities in its favor, it would still not be believable. In this sense, then, belief is created that the range we are considering does contain the true value provided the set outside that range has low credibility. So high credibility means that it is plausible or possible that a range contains the true parameter, but low credibility makes a stronger statement, that it is implausible that the range contains the true value. What credibility does not do is force yes-or-no dichotomies as in frequentist inference. It is much closer to Bayesianism in that it provides a richer collection of inferential statements. Unlike Bayesianism, however, it does not require

probabilities for parameters, and in particular it does not require probabilities that are not based on evidence.

The Second Simplest Example of a Credibility Function

Although one does occasionally see one-parameter examples, like the one we have been considering, it is far more frequent to see two or more parameters used in a model for comparing groups of observations. We can extend our previous example into this realm by imagining that one group of observations has n opportunities with t target events happening and target probability p, and then a second, independent group with "m" opportunities, "s" target events, and target probability "q". With no further assumptions, the probability of the outcomes (in the order observed) is

$$pr(t, s : n, p, m, q) = p^t(1-p)^{n-t}q^s(1-q)^{m-s}.$$

This is the probability of observing t target events in the first group and s in the second, based on their respective numbers of opportunities and target probabilities.

The credibility function in this case follows from the same principles as in the previous case. Credibility is the probability of our observations, divided by the maximum such probability. The maximizing values of the parameters are $p = t/n$, $q = s/m$, as one might expect. Thus

$$cr(p, q : n, t, m, s) = \frac{p^t(1-p)^{n-t}q^s(1-q)^{m-s}}{\hat{p}^t(1-\hat{p})^{n-t}\hat{q}^s(1-\hat{q})^{m-s}}.$$

Now we see that each pair of parameter values (p, q) has its own credibility. This *joint credibility function* is shown in Figure 2, where t and n are as before, and $s = 42$, $m = 60$. The interpretation is the same as before; the credibility of a given pair (p, q) is the probability of our observations if they were the true parameter values, relative to the largest such probability over all parameter values. The peak of the graph in Figure 2 occurs over the point (27150, 42/60).

Now consider the following problem: Given the above credibility function for pairs (p, q), what is the credibility for p alone? The answer comes from the way any credibility function extends from points to sets. The set here is, for a fixed value of p, all of the pairs (p, q) as q ranges from 0 to 1. Just looking at the above joint credibility, it is obvious that the maximum credibility happens when $q = q$. But then we simply get back the credibility that we had for p before. And this makes perfect sense, because ignoring the independent experiment with s, m, and q should put us right back into the situation we were in before, since the second experiment is truly irrelevant to p. This shows that the extension of credibility from points to sets by taking maxima corresponds to our common-sense idea of how inference should work.

Let us turn to a more interesting question, whether p and q are equal. From the pure inference standpoint, the only thing to compute is $cr(p = q : n, t, m, s)$, the

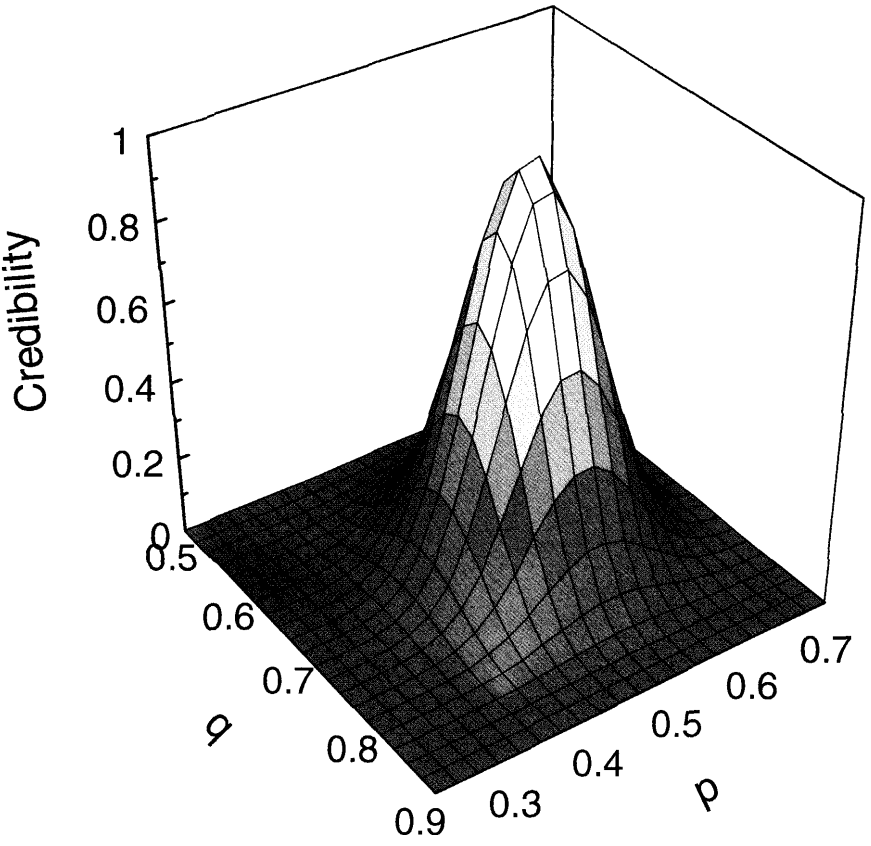

Fig. 2. Credibility function for two target event probabilities, p and q, the first based on 27 target events out of 50 opportunities, and the second based on 42 target events out of 60 opportunities.

credibility of $p = q$ based on the observations. Now the statement $p = q$ corresponds to the set of parameter points of the form (p, p) as p runs from 0 to 1. Thus, we are asking for the credibility of this set. Make the almost trivial observation that the credibility at (p, p) is the same as the credibility for a single parameter p, based on $n + m$ opportunities and $t + s$ target events. Thus, the maximum credibility occurs at $\hat{p}_e = (t + s)/(n + m)$. The subscript e here stands for "equal" and is to remind us that this is the credibility maximizer over the set of values (p, p). So we have the credibility that the parameters are equal as

$$cr(p = q : n, t, m, s) = \frac{\hat{p}_e^{t+s}(1 - \hat{p}_e)^{n+m-t-s}}{\hat{p}^t(1 - \hat{p})^{n-t}\hat{q}^s(1 - \hat{q})^{m-s}}.$$

This is the maximum probability for our actual observations under the assumption that the parameters p and q are equal, divided by the maximum probability

for our observations with no assumptions about p and q. It is the only thing that pure inference can tell us about the statement that the parameters are equal to each other. For our example,

$$\mathrm{cr}(p = q : n = 50, t = 27, m = 60, s = 42) = 0.2244.$$

This would pertain to an experiment that extended the one just described by including the second set of observations with m, s, and q. Perhaps the first person would be someone who might or might not have a gift for "seeing" the hidden symbols, whereas in the second case we had a confirmed seer who could regularly do better than random guessing. The statement $p = q$ is just that the first person is as good at seeing as the second person. Our result shows that this possibility is not incredible, but it also does not have very high credibility. Thus, our experiment is not yet very decisive on the point we would like to know, but there is an indication that the two people do not see with the same probability.

It is important to follow several things here. First, using the rule that credibility extends to sets by maximization, one can move from simple models with a few parameters to more complex models with multiple parameters. Second and in particular, it is easy to compute credibilities of statements that assert relationships between parameters. Third, the frequentist solution to the problem of $p = q$ is substantially more complex. The typical textbook approach uses a normal-distribution approximation that considerably simplifies the problem. This does not work for small sample sizes, however. There are multiple competing small sample solutions that do not agree with each other, and it is difficult to decide among them. Fourth, in some ways the Bayesian solution is even worse. Most Bayesians would put a continuous probability distribution over the parameter pairs (p, q). It then follows that the after-data (posterior) probability of them being equal is zero (this is essentially because the diagonal of values (p, p) for p running from 0 to 1 has area equal to zero). What is relatively simple and straightforward for pure inference is fraught with difficulties in the two major conventional approaches to inference—and so far we have considered only the second simplest possible example.

More on the Second Simplest Example

The approach I took to p and q above was much like null hypothesis testing. This was just to show how pure inference deals with this kind of question. A more informative analysis would be to consider some measure of how different p and q are, with equality as a special case, and then compute the credibility function for that measure of difference.

There are several possible difference measures. One is the simple arithmetic difference

$$d = q - p,$$

with $d = 0$ corresponding to $p = q$. Another is their ratio

$$r = q/p,$$

with $r = 1$ corresponding to $p = q$. A more elaborate one that is much admired in epidemiology is the "odds ratio"

$$w = \frac{q(1-p)}{(1-q)p},$$

again with $w = 1$ corresponding to $p = q$. Obviously this list could go on, but perhaps it is clear that the next few steps will be the same for any and all difference measures.

Let us consider "d". It is clear that (p, d) is equivalent to (p, q), in the sense that if we know one we know the other. The same is true for "r" or "w", that is (p, r), (p, w), (p, d), and (p, q) are all equivalent. Now returning to d, the credibility of a particular (p, d) is the same as the credibility of the corresponding (p, q). Since the corresponding $q = d + p$, we only have to substitute $d + p$ for q in the joint credibility $cr(p, q : n, t, m, s)$ that I computed in the preceding section. This value gives me $cr(p, d : n, t, m, s)$, the joint credibility function of p and d. The same sequence of steps will give me $cr(p, r : n, t, m, s)$ or $cr(p, w : n, t, m, s)$.

Since I want to focus on d rather than p, I now want just the credibility for d alone. But we have already seen how to compute this; just plug in the maximizing value for p. In this case (and in general) the maximizing value of p depends on the value for d, and so we have to write it as $\hat{p}(d)$. Thus

$$cr(d : n, t, m, s) = cr(\hat{p}(d), d : n, t, m, s).$$

Again, the same maneuver can give us $cr(r : n, t, m, s)$ or $cr(w : n, t, m, s)$ for the ratio or odds ratio.

To summarize, there are really only two steps. First express q in terms of p and d, then substitute into the (p, q)-credibility function to obtain the (p, d)-credibility function. Second, for any fixed value of d, maximize the (p, d)-credibility function to obtain the d-credibility function. It should be obvious that both of these steps are most reliably and quickly done by a computer program, especially since in the end we want the d-credibility function for a reasonably fine grid of values so we can plot it. As I indicated above, it is best to work with the natural logarithm of the credibility expressions, carry out substitutions and maximizations on this scale, and then use the exp function to convert back to credibility at the last step.

The results for our hypothetical example, for the d-parameter, are shown in Figure 3. The peak occurs at $d = 42/60 - 27/50$, and the curve is reasonably narrow, indicating that the range of highly credible values is acceptably small. We might rule out values of d below about -0.15 or above 0.35.

Using External Evidence

The Bayesian approach requires that there be some way of expressing uncertainty about the parameter, before making observations, with a probability

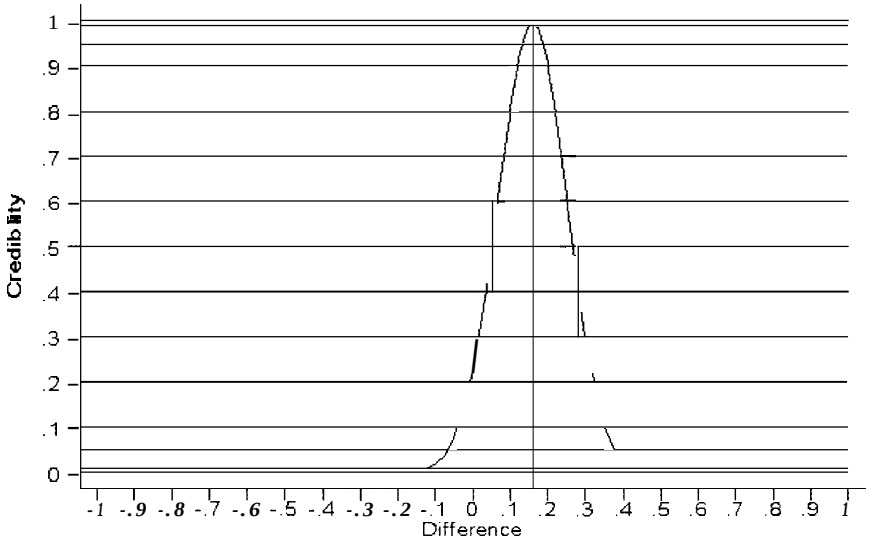

Fig. **3.** Credibility function for the difference between the two target event probabilities, $d = p - q$.

distribution. When no such distribution is apparent, the Bayesians have a problem, and they have been busy making up plausible (to them) solutions. But if there really were pre-data information about the parameter, then the Bayesian approach seems to have an advantage, because it has a way of expressing it.

So does the credibility approach. The pre-data credibility is simply formulated as a credibility function, cr(p : pre-data). We then make our observations and compute the credibility function cr(p : data) from the probability model, as I have done in the examples above. The two credibilities are then combined by

$$\text{cr}(\text{p} : \text{data } \& \text{ pre-data}) = \frac{\text{cr}(\text{p} : \text{data})\text{cr}(\text{p} : \text{pre-data})}{\max \text{cr}(\text{p} : \text{data})\text{cr}(\text{p} : \text{pre-data})'}$$

That is, multiply the two credibilities together, then divide by their maximum. This is the general rule for combining credibilities from independent sources, and it extends to the combination of any number of credibility functions.

I have observed that frequentists put their "prior" beliefs into the parameter space by setting up a "null hypothesis" value. Credibilists can mimic this process. Simply use a pre-data credibility that is maximized at the "null" parameter value. How narrow this credibility function is measures how much "prior belief' we have in the "null hypothesis".

Large-Sample Results

Although it is not the purpose of this article to go into what happens in large samples, it is certainly true that virtually all of frequentist inference is justified

by frequentist procedures having good properties as the sample size grows. In the typical case (and the only one I mention here) we are dealing with independent repetitions of an experiment, with the data always generated from the same model (that is, the parameters stay the same as the sample grows). This is not restricted to the simple cases I considered above; it is completely general.

The first result is that with no further assumptions whatsoever, the credibility of a false parameter value tends to zero with increasing sample size, and moreover this happens at an exponential rate (in terms of sample size). In other words, we are absolutely guaranteed that with enough observations we will be able to find the wrong parameter values incredible. Moreover, under a reasonable continuity assumption (for the mathematicians; the sequence of credibility functions is locally equicontinuous at the true parameter value), the credibility of the true parameter value always tends to 1.00. Again, the interpretation is that we will not, in sufficiently large samples, make the mistake of ruling out the true parameter value, and to the contrary it will appear more and more credible in larger and larger samples. There are, as one might imagine, other asymptotic results that can be proved for credibility functions, but these simple ones demonstrate that credibility is certainly on no more shaky grounds than is frequentist inference.

Final Comments

There seems to be a general principle in inference that any reasonable procedures will tend to yield similar results, when applied to the same observations. For this reason, some people see pure inference based on credibilities as simply being a re-packaging of familiar ideas, varying only in form but not content from conventional methods.

I would argue the reverse, that the form is important. Here are my reasons: First, credibility functions are easier to understand than the frequentist and Bayesian devices. This is because the principles are few and simple. You get credibility by dividing a probability function by its maximum. You compute the credibility of a set of parameter values by maximizing the credibility over that set. You combine credibilities by multiplying them, and then dividing by the maximum. It would be very difficult to compact either the frequentist or Bayesian philosophy into anything remotely as simple.

Second, pure inference is pure. It uses only the probabilities of the outcomes that were observed. It makes no additions of cultural or normative devices to distort the process of inference. It does not demand that inference be crammed into the form of a decision (like the frequentist methods), nor does it require that we engage in the intellectual fantasy that parameters (which are, after all, unknown constants) are somehow capable of having probability distributions. We are further freed from the Bayesian albatross, the necessity of having, at some point in an inferential argument, a prior distribution for the parameter that comes out of thin air (Aickin, 2004).

My third point is, however, the most important. Somehow during the previous century scientists lost the ability to present their data in ways that lead to discussion and enlightenment. They did this primarily by reducing all inferences to null hypotheses and p-values, which foreclose discussion rather than fostering it. In the process they severely restricted the kinds of inferential questions they are permitted to ask, and drove themselves (in biomedicine, at least) to require enormous sample sizes at astronomical costs in order to claim to have found effects that are often trivial and frequently misleading. The notion that one cannot act on results that are not "statistically significant" has become the dogma of modem statistical methods, and this has done irreparable harm to the enterprise of empirical science.

Pure inference is novel, as null hypothesis testing was when it was introduced to scientists. But it is also possible for ordinary scientists to learn how to use it so that they are not dependent on statisticians, who look at the empirical enterprise fundamentally differently than they do. As I hope to demonstrate in further articles, credibility functions can be used to pose and answer questions that are simply beyond conventional methods, and this is critical to our ability to make scientific progress in an environment of restricted resources for scientific exploration.

References

Aickin, M. (2004). Bayes without priors. *Journal of Clinical Epidemiology, 57, 4–13.*
Clayton, D., & Hills, M. (1993). *Statistical Models in Epidemiology*. Oxford University Press.
Edwards, A. W. F. (1972). *Likelihood.* Cambridge University Press.
Royall, R. M. (1997). *Statistical Evidence: A Likelihood Paradigm.* Chapman & Hall.
Sprott, D. A. (2000). *Statistical Inference in Science*. Springer Verlag.