ELSEVIER

Journal of Clinical Epidemiology 57 (2004) 4–13

Journal of Clinical Epidemiology

Bayes without priors

Mikel Aickin*

Center for Health Research, 3800 North Interstate Avenue, Portland, OR 97227, USA

Accepted 4 July 2003

Abstract

Background and Objectives: Classical statistical inference has attained a dominant position in the expression and interpretation of empirical results in biomedicine. Although there have been critics of the methods of hypothesis testing, significance testing (*P*-values), and confidence intervals, these methods are used to the exclusion of all others.

Methods: An alternative metaphor and inferential computation based on credibility is offered here.

Results: It is illustrated in three datasets involving incidence rates, and its advantages over both classical frequentist inference and Bayesian inference, are detailed.

Conclusion: The message is that for those who are unsatisfied with classical methods but cannot make the transition to Bayesianism, there is an alternative path.

Keywords: Statistical foundations; Theory of inference; Hypothesis tests; Likelihood function

1. Introduction

The foundations of classical statistical inference were laid down in the first third of the 20th century by Pearson (the elder), Gosset, Fisher, Neyman, and Pearson (the younger), among others. The methods of hypothesis testing, significance testing (*P*-values), and confidence intervals, were gradually and with some resistance accepted into social science fields, and then later, also with some resistance, adopted by biomedical scientists. At the end of the 20th century, every biomedical discipline that had any use for inference employed these methods to the exclusion of all others.

Although this portrays the larger historic trend, there have been at least two smaller countertrends. One consists of the persistent and occasionally discordant criticism directed toward the classical methods themselves (for some examples, see [1–9]). The resulting publications have ranged widely, from critiques of some of the technical infelicities of classical methods through dismay at their consistent and apparently undiminished misuse, on to incoherent diatribes by authors who clearly did not understand the ideas they deplored.

The other countertrend has been less visible, but ultimately more important. It has manifested itself as a slow but steady dribble, mostly in the statistics profession, away from the classical frequentist ideas and toward Bayesian inference. The trend has not spread very far, being restricted mostly to medical decision-making, mathematical statistics, and certain areas of systems and computer science. Disciplinary journals in the core biomedical sciences (exclusive of biostatistics) have scarcely allowed a Bayesian inference to pass their publication portals.

These countertrends leave a substantial number of inferentialists (to transcend disciplinary boundaries for a moment) in a quandary. The classical methods do, indeed, have some rather annoying deficiencies, and the tendency of noninferentialists to try to enshrine these deficiencies in "guidelines" for doing statistics degrades the art of inferential argument, at the same time that it trivializes the actual scientific results. But to abandon classical inference at the present time is virtually impossible without a Bayesian epiphany, the sudden realization that there is nothing really wrong with investing parameters with probability distributions, and then mixing these with the classical frequentist probability distributions. Standing directly in the way of the Bayesian epiphany, however, is the stark realization that one must find some way to believe in a prior probability distribution for the parameter, in the absence of any evidence, or alternatively, somehow subjectively cumulating all previous experience. As I will detail below, Bayesians have labored mightily to remove this barrier, but outside of statistics and a few other fields, they have not succeeded. The necessity of the prior distribution is both the philosophic and practical sticking point that has kept the dribble of Bayesian converts from becoming a torrent.

* Corresponding author. Tel.: 503-335-6667; fax: 503-335-2428.
E-mail address: mikel.aickin@kpchr.org

0895-4356/04/$ – see front matter © 2004 Elsevier Inc. All rights reserved.
doi: 10.1016/S0895-4356(03)00251-8

For epidemiologists in the quandary, the most obvious question is whether there is some way to try to do what the Bayesians want, without having to pretend that parameters have probability distributions, and in particular, is there some way to do without priors? The purpose of this article is to argue that the answer is "yes." The more detailed parts of the answer are both simple and complex. The simple answer is "use the relative likelihood," but the complex answer has to do with how it should be interpreted. Classical inference has greatly benefited from its underlying metaphor, making a decision, and Bayesianism likewise gains plausibility because it uses an increasingly familiar probability metaphor. Despite very good books on likelihood methods [10–12], the occasional article [13], and even a textbook [14], the inferential philosophy of direct use of the relative likelihood subsists without a metaphoric support structure.

The metaphor that I offer for likelihood-based inference is *credibility*. It is important to be clear that credibility does not mean probability, nor does it mean belief. It means, according to standard dictionaries, the *capacity for belief*. In what follows I will try to show that for those who understand the concept of a probability density, the foundations of credibility inference are trivial. Even the advanced theory is easy. The details are, as one might imagine, somewhat more involved, and so I illustrate their use in three real data examples, and relegate the mathematics to a technical appendix. I will then try to show that arguments that can be lodged against classical inference cannot be lodged against credibility inference, and finally, that credibility inference truly represents the intent of Bayesianism, without allowing parameters to have probability distributions, and in particular without allowing them to have prior distributions.

2. Inferential statistics in 2 min

Let $pr(x : \theta)$ denote a parametric probability density. The notation is general, so that x could be an observation, a sample of observations, several samples, or the values of a set of sufficient statistics. Likewise, θ is interpreted as parameter or family of parameters, in whatever generality is necessary. The *credibility function* is defined as $pr(x : \theta)$ evaluated at the actual observation x, divided by its maximum over θ, and it is denoted $cr(\theta : x)$. The value $cr(\theta : x)$ is interpreted as the credibility that θ is the true parameter, after having observed x. In the case of $cr(\theta : x) = 0$, then we say that θ is *completely incredible*, whereas if $cr(\theta : x) = 1$, then we say that θ is a *most credible value* (*mcv*), and if it is unique, we denote it by θ.

If T is a subset of the possible values of θ, then we define the credibility of T, $cr(T : x)$, to be the largest value of $cr(\theta : x)$ for θ in T. We can think of credibility applying to statements about θ, because for any statement S involving θ, there is a corresponding subset T of values of θ that make S true, and then we say that the credibility of statement S is the credibility of subset T. For example, $cr(\theta > 0 : x)$ is the credibility that "θ is positive."

3. Advanced inferential statistics in 5 min

The values of $pr(x : \theta)$ for different θs provide probabilities for the observation x that we actually made, computed under different assumptions about which θ is the true parameter value. Because events of higher probability occur more frequently than those of lower probability, the higher $pr(x : \theta)$ is, the more evidence it supplies that the value θ is a good explanation for the observed x. The credibility function simply rescales $pr(x : \theta)$ to lie between 0 and 1. Explanations that are sufficiently poor, for small enough $cr(\theta : x)$, can effectively be ruled out.

The values of $cr(\theta : x)$ as θ varies in a set T represent credibilities of explanations in T. The largest value of $cr(\theta : x)$ for θ in T is the credibility of T, according to the dictionary definition of credibility as *the capacity to believe*. High values of $cr(\theta : x)$ or $cr(T : x)$ do not represent measures of belief, but of the capacity to believe. We *could* believe that T contains the true θ if there are values of θ in T that have high credibility. We *could not* believe that T contains the true θ if all values of θ have low credibility.

If $pr_i(x_i : \theta)$ represent independent observations for $i = 1,2,\ldots,n$, then by definition their joint probability is $pr(x_1,x_2,\ldots,x_n : \theta) = pr_1(x_1 : \theta)\ pr_2(x_2 : \theta)\cdots pr_n(x_n : \theta)$. According to the definition $cr(\theta : x_1,x_2,\ldots,x_n)$ is defined by dividing $pr(x_1,x_2,\ldots,x_n : \theta)$ by its maximum (over θ). But if we divided $cr(\theta : x_1)cr(\theta : x_2)\cdots cr(\theta : x_n)$ by its maximum, we also would get $cr(\theta : x_1,x_2,\ldots,x_n)$. This shows that the rule for combining credibility functions from independent sources is to multiply them together and then divide by the maximum. Note especially that this would be true if $\theta = (\theta_1,\theta_2,\ldots,\theta_n)$ and $pr_i(x_i : \theta) = pr_i(x_i : \theta_i)$, so that information on the ith component of θ came only from the ith sample.

In general, if one has multiple parameters θ_i $(i = 1,2,\ldots,n)$, then one can make multiple statements, such as $T =$ "θ_1 is in T_1 & θ_2 is in T_2 &…& θ_n is in T_n." Here, it is in the nature of credibility and belief that $cr(T : x)$ will go down as n goes up. In this case, we may want a value that summarizes the credibilities of the individual T_is, and an appropriate choice is

$$\overline{cr}(T : x) = cr(T_1 \;\&\; T_2\ldots\&\; T_n : x)^{1/n}$$

which we call the *summary credibility*. When the credibilities of the T_i's are based on independent sources, then the summary credibility is the harmonic mean of the individual credibilities.

Credibility functions form a special case of plausibility functions [15] in the theory of belief functions [16,17]. We now turn to three examples in which credibility inference is applied to rates.

4. Malignant mesothelioma mortality rates

The dataset in Table 1 consists of the survival times, in months from diagnosis, of 38 malignant mesothelioma

Table 1
Survival and censoring(*) times in months for 38 malignant mesothelioma patients [30]

2 2 3 3 3 4 4 4 5 6 6 6 6 7 8 11* 11
12 13 13 14 14 14 14 17 18 18 19 19*
20 20 25 26 27 31 37* 45* 49

patients, including four censored observations. The total time-on-study was $T = 556$, and because there were 34 deaths, the mcv is $\hat{r} = (12)(34)/556 = 0.7338$, in terms of events per person-year. The graph of the credibility function for r is shown in Fig. 1, and demonstrates the skewness to the right that is typical for rate estimates. This graph is useful for summarizing inference about r, because the credibility of any subset of r values can be obtained by approximating the maximum of the credibility function over that set. Conversely, one can also easily find subsets of r that are incredible (say, $cr < 0.10$) and their complements. For example, as shown in Fig. 1, one can visually approximate that the set where $cr > 0.10$ is [0.50,1.04]. Exact computation gives [0.4958,1.0381]. Because the probability of survival 1 year past diagnosis is $p = 1 - e^{-r}$, we immediately and easily compute that the set of probabilities p having $cr > 0.10$ is [0.3909, 0.6459] (because $0.3909 = 1 - e^{-0.4958}$ and $0.6459 = 1 - e^{-1.0381}$), which indicates the level of precision of the results on a familiar probability scale.

In this example, it could be questioned whether an analysis based on a single incidence rate is adequate. Credibility inference is brought to bear on this issue by considering more elaborate models, perhaps with time-varying incidence rates, and by computing the credibility of the statement that a single, constant rate applies throughout the study period, relative to the point of view expressed by the more complex model. For example, one might divide the study period into two or four subintervals with approximately equal values of time-on-study, and compute the credibility of the statement that the separate rates for each of the subintervals are, in fact, equal. This gives the results shown in Table 2.

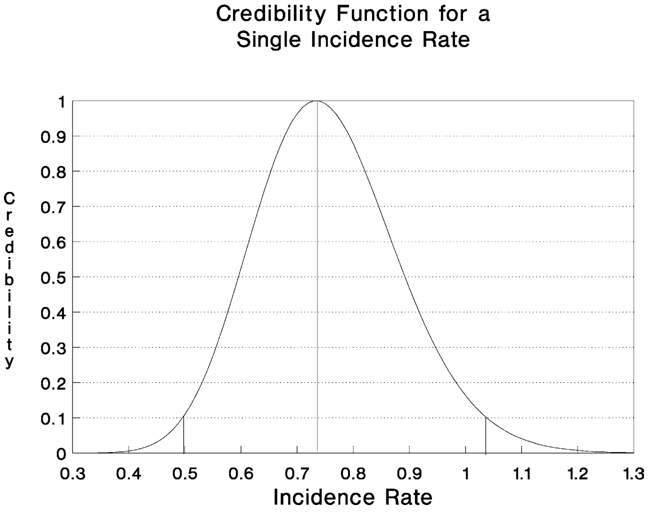

Fig. 1. Credibility function for the rate (per person-year) of death among malignant mesothelioma patients.

In the case of four subintervals, one might want not only the credibility that all four rates are equal, but also the summary credibility, which is obtained by taking the 1/3 power. (The power is 1/3 because the statement that four rates are equal is equivalent to three logically independent statements; viewed differently, we are going from a model with four rate parameters to a submodel with only one, a difference of three parameters.)

The analysis as presented thus appears justified, because it is credible that separate rates over either two or four subintervals are equal to each other. There is, of course, nothing to prevent the consideration of more complex time-varying models, such as $r(t) = \alpha + \beta t$, but we do not pursue this method here.

The classic approach to this example would be to compute a confidence interval for r. If the question of heterogeneity of rates were to be raised, however, this would lead to some sort of homogeneity hypothesis test, and this gives rise to two difficulties. The first is that a simple subdivision of the study time, as used above, would be on shaky ground, because the choice of time intervals would depend on the accrued person-time. But accrued person time is itself regarded in classic inference as part of the outcome, and so it is not clear that the hypothesis chosen for testing is independent of the data on which it is to be tested [18]. The second difficulty is that there is no technology for combining the significance level of a homogeneity test with the confidence coefficient of a subsequent confidence interval to produce a single, overall assessment of the results.

In contrast to the classic approach, the credibility approach attains simplicity by placing all assessments on the same footing, as expressed by credibilities. Moreover, the credibility approach does not place roadblocks in the way of investigating other possible interpretations. The choice of subintervals of person-time, or of more complex models that are suggested by the data, is legitimate. This is true because the credibilities that are produced are objective facts about

Table 2
Analyses of the data from Table 1, showing the computation of the credibility of equal rates over time subintervals

End of subinterval	Time on study (months)	Number of deaths	MCV of r (per year)	Credibility of equal rates	Summary credibility
Two subintervals					
9 months	276	15	0.6522		
49 months	280	19	0.8143	0.8110	0.8110
Four subintervals					
4 months	145	8	0.6621		
9 months	131	7	0.6412		
17 months	147	9	0.7347		
49 months	133	10	0.9023	0.7325	0.9014

the observations, given that the model generated the observations.

5. Childhood leukemia cluster in Illinois

Methods for comparing rates are illustrated by the data set in Table 3, from a classic study on childhood leukemia incidence in Illinois. Person-years of experience over the 5-year period 1956–1960 were computed by multiplying the 1960 census figures in the 0–15 year age group by 5. Leukemia cases were ascertained by physician reports or death certificates.

The point at issue in this study was whether there was an elevated incidence of childhood leukemia in the Niles area, compared to the incidence in the four nearby communities. The mcv for Niles was $\hat{r}_2 = (100000)(8)/35380 = 22.61$ (per 100,000 person-years), whereas the mcv for the pooled sample of the other communities was $\hat{r}_1 = (100000)(13)/220375 = 5.90$. Consequently, the mcv of the rate ratio was $\hat{u} = 22.61/5.90 = 3.83$, and for the rate difference, $\hat{d} = 22.61 - 5.90 = 16.71$.

The credibility function for the rate ratio appears in Fig. 2, and shows the considerable skewness one expects for this measure. The credibility of the statement that Niles' rate equals that of its neighbors is obtained by computing the combined rate for all five areas, $\hat{r} = 8.21$, and then $cr(r_1 = r_2 : x) = 0.0222$. It is, therefore, not credible that the Niles' rate equals that of its neighbors. The credibility function for the rate difference appears in Fig. 3, and suggests that the interval with $cr > 0.10$ for the excess number of cases (per 100,000) in Niles runs from about 3 to 37. It is perhaps worth noting that the credibility function for d is not symmetric about the mcv, wheareas in classical inference, symmetric confidence intervals are often used for rate differences.

One might take the position in this analysis that the decision to pool the four nearby areas requires some justification. The credibility that these four rates are equal is 0.5429, and so it is credible that they are in fact equal. In pooling, however, we are, in effect, selecting a submodel to use for the analysis, and so it may be more appropriate to use the summary credibility, the one-third power of 0.5429, which is 0.8158, indicating that the submodel is quite credible. By these arguments, the analysis as presented is justified.

Table 3
Person-years experience and leukemia cases among residents aged 0–15 in five Illinois regions [31]

Area	Person-years	Cases
Niles	35380	8
Des Plaines	59760	3
Lincolnwood	18270	1
Morton Grove	39975	4
Skokie	102370	5

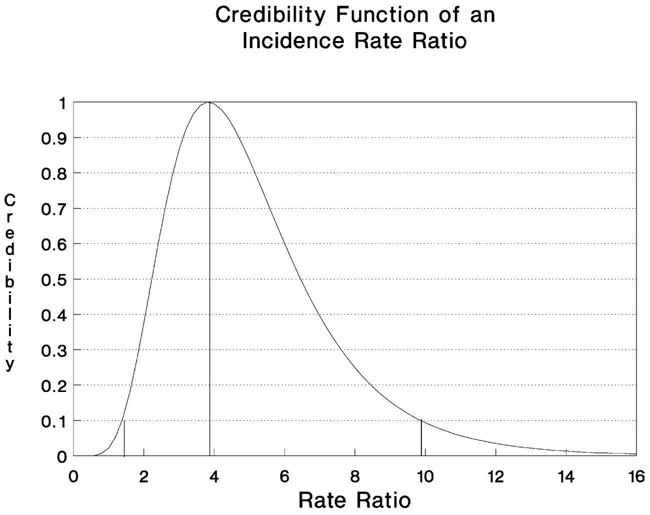

Fig. 2. Credibility function for the ratio of childhood leukemia incidence in Niles, IL, relative to surrounding areas.

There are, of course, other inferential paths one might choose for this analysis. One could compute, for example, the credibility that all five rates are equal, or one could make all four pairwise comparisons of Niles with each other community, and report both the individual and summary credibilities. Any and all of these approaches are justified. The reason for selecting the approach we did was not because it was dictated by credibility inference, but because it was the simplest, most direct, and most revealing analysis that answered the question, whether Niles differed from its neighbors.

Whatever credibility approach one chooses for this data set, it seems obvious that hypothesis tests and confidence intervals are problematic. To apply either of these devices, it is necessary that one specify the procedure (whether a test or confidence interval, and if a test, the null value)

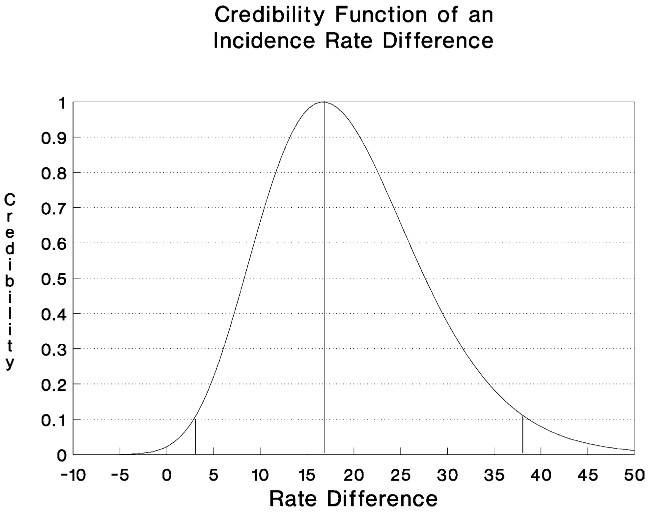

Fig. 3. Credibility function for the difference between the childhood leukemia incidence rate (per 100,000 person-years) in Niles, IL, relative to surrounding areas.

before looking at the data. But certainly the decision to compare Niles with its neighbors was suggested by the high-observed rate in Niles, and would not have been contemplated before the data became available. Credibility inference, on the other hand, is intended to assess statements about parameters after the data are in hand, and it is perfectly legitimate to compute the credibilities of statements that are suggested by the data. Thus, in this example, the use of the credibility approach must acknowledge that the results for Niles and its neighbors might have been selected as the worst possible such comparison that could have been chosen in the state of Illinois during the given time period, and that if true, this circumstance has material bearing on the conclusions. In contrast, the hypothesis testing approach is more easily and convincingly carried out if this caveat is ignored.

6. Childhood leukemia cluster in Phoenix

Table 4 contains data from a childhood leukemia cluster investigation in Maricopa County, Arizona, which coincides with the Phoenix SMSA. All incident cases occurring to resident children aged 0–19 were ascertained for 1965–1986. The purpose of the study was to determine whether a 1.95 standardized mortality ratio [19] for leukemia in West Central Phoenix (compared to the remainder of Maricopa County) would be confirmed in an incidence study.

Table 4
Counts of incident cases and person-years experience (in 1,00,000s) of childhood leukemia in Maricopa County (Time is coded 1 = 1965–1969, 2 = 1970–1981, 3 = 1982–1986).

			West Central Phoenix		Remaining county	
Sex	Time	Age	Person-years	Cases	Person-years	Cases
M	1	<5	0.280	3	1.326	3
		5–9	0.350	1	1.574	2
		10–14	0.322	0	1.530	0
		15–19	0.223	0	1.202	3
	2	<5	0.269	1	1.277	3
		5–9	0.330	1	1.518	4
		10–14	0.314	1	1.500	0
		15–19	0.231	0	1.200	2
	3	<5	0.691	5	4.037	22
		5–9	0.782	7	4.443	16
		10–14	0.811	3	4.829	15
		15–19	0.745	2	4.933	7
F	1	<5	0.667	4	3.843	19
		5–9	0.747	2	4.301	7
		10–14	0.776	0	4.677	6
		15–19	0.733	1	4.798	5
	2	<5	0.363	5	2.714	19
		5–9	0.349	0	2.578	7
		10–14	0.312	2	2.528	6
		15–19	0.343	1	2.826	8
	3	<5	0.351	4	2.609	11
		5–9	0.326	3	2.458	10
		10–14	0.297	0	2.430	9
		15–19	0.304	3	2.680	3

Although the sex and age distributions did not differ much between the two subpopulations, it was felt that these two factors should be used for adjustment before making a comparison, and this is the chief methodologic distinction between this example and the preceding one. There was more concern about potential differences resulting from population growth, so that the time period was divided into three segments, for the purpose of adjustment. The coefficients used to define the incidence rate ratio were the person-years figures from West Central Phoenix, so that the comparison was an (indirectly) standardized incidence ratio. The mcv was $\hat{u} = 1.67$, and the credibility function is shown in Fig. 4.

The credibility that the incidence rate ratio was at or below 1 was only 0.0102, effectively ruling out this eventuality. Conversely, the credibility of 1.95 (the previous mortality ratio) was 0.6182, indicating support for a value this elevated. The interval with $cr > 0.10$ for u was 1.17–2.33, indicating that values above 2 could not be ruled out.

Classic methods for approaching this example might be based on the large-sample approximation to the distribution of a standardized rate ratio [20], or on Poisson regression [21,22], which also uses large-sample approximations. The use of large-sample approximations raises the issue of their accuracy, which must be developed through additional statistical research, while the credibility approach is exact and requires no further work. Poisson regression is particularly inappropriate for this example, because it proceeds by fitting models, which include parameters for age, sex, and period-specific effects that are not of interest [23]. Analyses that siphon off information by estimating parameters that are not related to the question of interest only risk weakening the precision of the analysis of the question that is of interest.

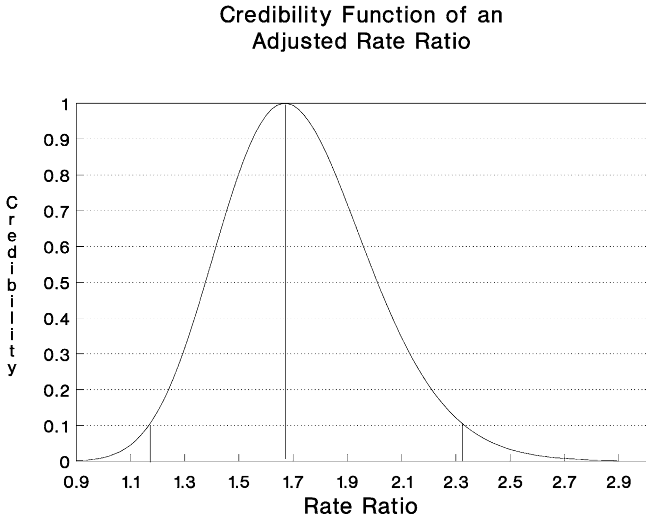

Fig. 4. Credibility function for the adjusted incidence rate ratio for childhood leukemia in West Central Phoenix relative to the remaining Phoenix SMSA.

7. Advantages of credibility inference over classic inference

7.1. Metaphor

Hypothesis testing rests completely on a decision metaphor; one either confirms the null hypothesis or rejects it in favor of the alternative hypothesis. Significance testing (and therefore the *P*-value) retains this metaphor, changed only in what it means by the “alternative.” Confidence intervals are sometimes portrayed as having a different base, but in the end a $1 - \alpha$ confidence interval is nothing more than those values of the parameter which, if they had been null hypothesis values, would have been confirmed at level α.

There are, to be sure, instances in epidemiology where a decision is warranted. They are usually cases in which there has already been substantial study, the methods have been well developed, the pitfalls and problems have been identified, the measurement issues have been resolved, and it is time to find out whether Nature works *this way* or *that way.* These situations are, however, extremely rare. It would be preposterous to imagine that, among the routine kinds of studies that are carried out and reported across the broad front of biomedicine, more than a tiny fraction have actually reached the decision-making point. Moreover, in those cases, if a decision is really warranted, then a careful evaluation of the benefits of being right and costs of being wrong should be undertaken, prior evidence should be assessed, and a formal decision-making procedure crafted. Because routine hypothesis testing leaves out these other elements, it is in a sense “half-baked” decision-making.

Metaphors are sometimes said to be unimportant, useful only for teaching the basic ideas to students, and not to be taken seriously in actual practice. This is almost entirely untrue in the case of hypothesis testing, as demonstrated by the powerful meaning invested in the term *significant*. When the results are *significant*, evidently we are warranted to say that something has been demonstrated, a rejection of a null hypothesis has been decided, work has been accomplished. One would be hard pressed to find a biomedical scientist who would say that the foundation for advancement in their field was determined by a tutorial *metaphor,* and yet that it precisely what has happened in classic inference.

Credibility analysis does not ask scientists to do anything other than to see which parameter values provide good explanations for the observations they actually made, and which provide poor explanations, conditional on the appropriateness of the statistical model.

7.2. False dichotomies

Making a binary decision when it is not warranted is, of course, the most important false dichotomy of classic inference, but there are others. In fact, there seems to be a powerful drive to make everything about inference binary. Theoretically, *P*-values could be reported instead of being dichtomized to $P < .05$ and $P > .05$. In fact, a *P*-value could be reported for every value in the parameter space. Likewise, there is no need to state a single confidence interval (dichotomizing the parameter space into the confirmed and rejected values). The tendency to make continuous presentations of evidence into binary reports is as unjustified as it is ubiquitous.

The credibility function for one or two parameters, or multiple credibility functions for the same parameter, can easily be displayed in their entirety, allowing the reader to examine the explanatory power of individual points or certain intervals. For higher dimensional problems, it becomes part of the art of presenting and defending inferences to devise a display that informs, and does not dichotomize. Of course, one can cram a credibility function into a single result (like the interval with $cr > 0.10$, as was done above), but one can also resist this temptation. The methods of presenting classical inference have, unfortunately, gone too far to change.

7.3. Illogical results

If a statement about parameters is contradictory, then its credibility is zero by definition. This cannot be said of hypothesis testing. Although it may be unfair to single out any particular instance, one very important case involved the aspirin arm of the Physicians’ Health Study [24]. By a legitimate use of hypothesis testing the investigators concluded that aspirin prevented death from myocardial infarction, had no effect on other cardiovascular causes of death, and had no effect on the total rate of cardiovascular death. Because the total mortality rates in the aspirin and nonaspirin groups were the sums of the myocardial infarct and nonmyocardial infarct death rates, the primary finding of the trial was logically contradictory. A credibility analysis if these data says that it is credible that aspirin protects against myocardial infarction death, it is less credible but still credible that aspirin raises the rate of nonmyocardial infarction death (especially hemorrhagic stroke), and it is nearly perfectly credible that these two opposite forces cancel each other out in the total cardiovascular death rates. The credibility of the published trial report is zero. The reader may review the original article to see whether this represents the data as fairly as I believe it does, but only those with a high tolerance for contradiction could accept the original report at face value. Thus, although this was a large trial in which a decision was highly indicated, a credibility analysis makes more sense of the observations.

7.4. Rigid conventions

Because classical inference fails to foster the skill of inferential presentation, it invites the efforts of those who would like to reduce the process of inference to a series of rigid rules, or “guidelines,” as they are often misleadingly called. This is the source of the 90% or 95% confidence interval, the 0.05 significance level, and the pernicious “80%

or 90%" power rule, unwritten but almost universally adhered to in NIH scientific merit reviews of grant applications. It is also the source of mindless rules, such as the emerging notion that in reporting on clinical trials, a sample size justification must be provided *after the fact*, *even if the results are significant!* Credibility avoids some of these problems because it is not connected to a decision metaphor, but it enjoys perhaps only temporary immunity until the fabricators of guidelines catch up. It is no longer possible to retrieve classical inference from guidelinization, but perhaps with some vigilance credibility analysis could achieve a different fate.

7.5. Hard to teach

I showed above that a complete course in the fundamentals of credibility inference takes 7 min. Of course, there are plenty of details to be fleshed out, and certainly enough for one, or perhaps two statistics courses. But there is no reason to imagine that anyone will trip over the fundamentals. No one who has tried to teach statistics, and is honest about the experience, will say the same of classical inference. At least in the United States, the vast majority of students leave their statistics training with virtually no concept of what a sampling distribution means, nor why it is central to classical inference. Moreover, if the biomedical literature is any guide, they are perfectly capable of going forth into the world and doing genuine harm by putting their fundamental misconceptions into practice. They are, it is sad to say, even further abetted in this by the small but continual dribble of articles by nonstatisticians attempting to explain statistics for once and for all to their colleagues. Ph.D-level statisticians shake their heads at this bizarre literature, but they are evidently not called upon to referee it.

7.6. False problems

One of the consequences of forcing a nondecision problem into a decision metaphor is that it raises issues that can be entirely avoided. Two of the data examples I choose above were deliberately taken from the area of cluster investigations, because these enterprises often bring out the worst in classical methods. When an apparent excess disease or death rate is observed in an area, initially *P*-values are quoted to declare that it cannot be due to chance, and then a statistical expert claims that the *P*-value is invalid because it was computed on the area that gave rise to the statistical question in the first place (the so-called "Texas sharp-shooter" phenomenon; shoot the barn then draw a bulls-eye around the hole). The political forces ignore this and mount an expensive investigation anyway, which risks not being taken seriously because it tests so many hypotheses, either through consideration of multiple regions, or identification of multiple potential causes, that statisticians can easily label it a data-dredging operation.

The credibility approach does not try to mediate the interface between the data and mind of the beholder. It reports how well any statement about differences among regions explains what was observed. It invites further scientific inquiry through the ordinary processes of science, instead of expecting that the numbers and the numbers alone will solve all of the dilemmas.

Another false issue is the multiple testing debate. Because researchers generally collect more data than they need, they are susceptible to performing more hypothesis tests than they should. When they try to report their tests in a single article, they are set upon by the guideliners to adjust their *P*-values for multiple testing. The problem is created because the researcher does not have a single, well-developed question that deserves a hypothesis test, but is still at an earlier phase of the research process in the specific area. The researcher would do better to report results in credibility terms, which need not make any adjustment, except in certain very special circumstances. In fact, the most popular evasive maneuver is for the researcher to publish multiple articles on the same dataset, each one appearing to test so few hypotheses that no multiple testing adjustment is necessary. In addition to promoting bad publishing policy, the debate on this topic has created an entirely artificial division between adjusters and nonadjusters, and the arguments have degenerated to the point that a supposedly educational article in a leading journal was recently published [25], in which almost every statement made was either false or misleading. The point here is that opponents of multiple testing adjustments do not feel they have any ground to stand on in criticizing hypothesis testing itself, so they must resort to plausible-sounding (but false) arguments to try to stop an indefensible guideline.

These topics do not come close to exhausting the subject. Although there are many articles containing other valid criticisms of classic methods across a variety of literatures, there are also some containing exceedingly ignorant or even facetious arguments against these methods. It is, perhaps, a final criticism of classic methods that they are so susceptible to being so thoroughly misunderstood and misrepresented.

8. The advantage of credibility inference over Bayesian inference

8.1. No priors

Because Bayesian inference is hardly ever used in disciplinary biomedical journals, it does not have as rich a history to criticize. Indeed, the problems with Bayesianism come almost entirely from its philosophy. The frequentist believes that an observation has a sampling distribution because of the manner in which it was procured. He/she does not believe that a parameter is generally capable of having a probability distribution, because there is no issue of its procurement (there are exceptions, but this is the leading, practical case). Bayesians believe that it is sensible to say that a parameter has a probability distribution, not in the chance-probability sense of the frequentist, but in the belief-probability sense. That is, they take the position that it is reasonable to express beliefs about a parameter in the face of uncertainty with a probability distribution. They also believe that belief-probability follows the same rules as, and can be mixed with chance-probability. Thus, Bayes theorem is a tautology for chance-probability (a simple restatement of the definition of conditional probability), but it is a philosophic assertion when it contains parameters with probability distributions.

The overwhelming stumbling block for those who would be philosophic Bayesians is that they must come up with a probability distribution for their parameter before seeing any evidence, the so called *prior distribution*. The idea that the subjective opinion of the researcher should be explicitly injected into the computations of inference must challenge some deeply felt taboo, or virtually all scientists would now be Bayesians, whereas virtually none actually are. The Bayesians counter the subjectivism argument by saying that the frequentists just want to hide all of the subjective elements of their statistical analyses, while the Bayesians are completely forth-coming in specifying their prior distribution. This argument has also evidently fallen upon deaf ears, and deservedly so. Surely the Bayesians already have all of the subjective elements inherent in frequentism, and so adding another, whether explicit or not, is hard to see as a virtue.

To counter the subjectivism charge, Bayesians have turned to so-called "noninformative" priors. These often turn out not to be probability distributions, but measures that give infinite mass to the parameter space. In another maneuver, genuine priors can be shrunk toward an imaginary limit that is argued to be noninformative. This leaves the Bayesians in the hard position of arguing for priors on philosophic grounds, and then replacing them with nonprobability distributions, or trying to make them magically disappear in actual analyses. Yet another problem created by the "noninformative" argument is that the prior that is "noninformative" for a ratio of two rates is "informative" about their difference. This kind of attitude toward consistency is rather difficult for many statisticians and most scientists to swallow.

A Bayesian would be willing to use the credibility function, because it is proportional to the likelihood function that appears in Bayes' theorem. Thus, in a sense a credibility function is Bayesian, but without a prior. Because there is no prior, the consequence is that a credibility function cannot be interpreted as a probability. It might be argued, however, that this is an advantage rather than a defect. For one thing, a Bayesian posterior distribution requires that all probability either favor or oppose any given statement about the parameter. Credibility is capable of taking the more flexible possible position, that when the argument is pressed for the statement it is credible, but when the argument is pressed for its opposite that is also credible. This comes a good deal closer to representing what uncertainty really means, than to assigning partially arbitrary and very crisp probabilities.

On the other hand, if one has prior beliefs, then credibility inference can accommodate them. The beliefs only need to be expressed as a no-data credibility function over the parameter space. The usual rule for combining credibilities holds (multiply them, divide by the maximum). In fact, one could take one's prior credibility to be identically equal to one (*the ignorance credibility*), and recover the use of the credibility function that we have seen above. What is impossible for the Bayesian (expressing prior ignorance) is almost automatic for credibility inference.

Finally, credibility inference is trivial for functions of parameters:

$$cr(f(\theta) = \varphi : x) = cr(f^{-1}(\varphi) : x)$$

which just says that the credibility of "$f(\theta) = \varphi$" is the credibility of the set of θ for which $f(\theta) = \varphi$. Solution of this problem for Bayesian probability distributions is considerably more difficult.

9. Discussion

There is considerable literature documenting the struggles between Bayesians and frequentists, which is not referenced here. Nevertheless, Efron [26] offers a particularly accessible insight into the intricacies and trade-offs of these two different world views, in which he emphasizes the problem created by the necessity of a prior distribution for the parameter. In a commentary following Efron's article, Carl Morris [27] clearly points out that the "empirical Bayes" framework includes (in a certain sense) both the dedicated frequentist and the dedicated Bayesian points of view as extremes, with "empirical Bayes" methods offering alternatives somewhere in between. There have also been attempts to unify Bayesian and frequentist approaches by constructing maneuvers that give the "same" answers, interpretable in either framework [28], but these have been severely criticized (see the comments following this latter article). For a dedicated Bayesian view one can certainly cite Lindley [29].

It is not the purpose of this article to resolve the Bayesian/frequentist duality, nor to offer a hybrid method of inference that satisfies both camps. The fundamental insight of Bayesians is that inference should consist of the assignment of values (for them, probabilities) to statements about parameters, both before and after informative observations have been made. If one is willing to accept the assignment of credibilities, rather than probabilities, then the necessity of priors vanishes, along with the necessity of probability distributions for parameters. This perspective says that probabilities are appropriate for observations, because observations are procured in ways that support a probability story. It further says that information about a parameter is procured through a mechanism by which particular values of the parameter have the power to explain particular values of the observations, and that this is a credibility story.

Probability and credibility are dual concepts, in the same way that observations and parameters are dual concepts. The indication here, and in works such as Sprott [12], is that Bayes without priors is a feasible method of inference. It might be argued that this approach is particularly appropriate for many epidemiologic studies. These studies generally contribute to the cumulating evidence about a risk-factor/disease association, rather than being definitive, so that speaking in terms of credibility instead of decision-making fits better with their purpose. Furthermore, due to the simple rule for combining credibility functions, one can easily combine newly reported evidence with a credibility function based on all past evidence (including discussion of homogeneity of results, if waranted) as part of the discussion. Because credibility inference can be readily understood by anyone who has studied elementary probability, it is therefore less susceptible to the arguments and misunderstandings that have separated the frequentists from the Bayesians. As perhaps its greatest virtue, however, by turning its back on ritualistic data presentations, credibility inference provides the opportunity to reintroduce the art of presentation and interpretation of scientific evidence into epidemiology.

Technical appendix

Inference for a single rate

In a study with times t_i to a target event or censoring, then assuming a constant incidence rate we have for each individual

$$pr(t_i : r) = r\begin{cases} r\exp(-rt_i) & \text{for those experiencing the event} \\ \exp(-rt_i) & \text{for those who are censored} \end{cases}$$

It follows that N = number of events and T = total of the t_i's are sufficient, and

$$pr(N,T : r) = r^N \exp(-rT)$$

The mcv is easily obtained from calculus arguments to be $\hat{r} = N/T$, and then the credibility function is

$$cr(r : N,\ T) = \frac{pr(N,\ T : r)}{pr(N,\ T : \hat{r})}$$

$$= \begin{cases} \left[\dfrac{rT}{N}\right]^N \exp(N - rT) & (N > 0) \\ \exp(-rT) & (N = 0) \end{cases}$$

This formula was applied to the malignant mesothelioma example data.

Poisson model for rates

In a quasi-cohort study, individuals can both enter and leave the cohort over time, and there is an implicit assumption that these two processes balance each other. If N events happen during a period in which T units of person-time in the cohort have accrued, then the Poisson model is

$$pr(N,T : r) = (rT)^N \exp(-rT)/N!$$

Again, calculus gives $\hat{r} = N/T$, and the resulting credibility function is identical to that given above. This shows that the formula above for cr also applies to the two cluster examples.

To compare incidence rates from two independent studies, we first compute their joint credibility

$$cr(r_1,r_2 : x_1,x_2) = cr(r_1 : x_1)cr(r_2 : x_2)$$

where we now abbreviate the data $x_i = (N_i, T_i)$. One comparative measure is the rate ratio, defined by $u = r_2/r_1$. Because $r_2 = ur_1$ as a matter of definition, it follows immediately that the joint credibility of r_1 and u is

$$cr(r,u : x_1,x_2) = cr(r_1 : x_1)cr(ur_1 : x_2)$$

The credibility of the statement that the rate ratio is some particular value u is found by maximizing the above expression over r_1, as an immediate consequence of the definition, extending credibility from points to sets. This value is

$$\hat{r}_1(u) = \frac{N_1 + N_2}{T_1 + uT_2}$$

Note how the notation indicates a mcv for one parameter (r_1) with the other (u) held fixed. This value is now substituted back into the joint credibility function, in effect eliminating r_1:

$$cr(u : x_1,x_2) = \frac{\left(\dfrac{N_1 + N_2}{T_1 + uT_2}\right)^{N_1+N_2} u^{N_2}}{\left(\dfrac{N_1}{T_1}\right)^{N_1}\left(\dfrac{N_2}{T_2}\right)^{N_2}}$$

Of course $cr(u = 1 : x_1,x_2) = cr(r_1 = r_2 : x_1,x_2)$, the credibility of equal rates.

This same basic strategy works for any other comparative measure. For example, let $d = r_2 - r_1$. Then

$$cr(r_1,d : x_1,x_2) = cr(r_1 : x_1)cr(r_1{+}d : x_2)$$

The mcv for r_1 with d held fixed is

$$\hat{r}_1(d) = \left[(\hat{r} - d) + \sqrt{(\hat{r} - d)^2 + 4N_1d/(N_1 + N_2)}\right]\Big/2$$

This is substituted into the preceding joint credibility function, in effect eliminating r_1 and yielding the credibility function for d. Of course, $cr(d = 0{:}x_1,x_2) = cr(r_1 = r_2{:}x_1,x_2)$, and this will be the same value that was obtained starting out with the rate ratio. These formulas were used to analyze the two cluster examples.

Inference for many rates

The general homogeneity problem is easily worked out. With counts of events N_i and times-on-study T_i over studies $i = 1,2,\ldots,n$,

$$cr(\text{equal } r_i s) = \frac{\hat{r}^N}{\prod_i \hat{r}_i^{N_i}}$$

where $N = N_1 + \ldots + N_n$, $T = T_1 + \ldots + T_n$,

and

$\hat{r} = N/T$.

It turns out that a number of interesting questions about rates can be investigated by computing the credibility of statements of the form

$$a_0 + \sum a_i r_i = 0$$

In the case of two rates, taking $a_1 = 1$ and $a_2 = -1$ makes a_0 equal to the difference between the rates, while taking $a_0 = 0$ and $a_2 = -1$ makes a_1 the rate ratio. In general, however, the credibility of a linear equation is most useful for adjusted rates. Here, there are two populations ($i = 1,2$) and multiple rates within corresponding groups $j = 1,2,\ldots$ nested in the populations. We use subscript ij for the jth group of the ith population. We also require person-time figures T_j* from a reference population to use for the standardization. Now define

$$a_{1j} = T_j^*/T^* \quad a_{2j} = -T_j^*/T^* \quad T^* = \sum T_j^*$$

so that a_0 is the difference between two adjusted rates, or else define

$$a_0 = 0 \quad a_{1j} = uT_j^* \quad a_{2j} = T_j^*$$

so that u is the ratio of two standardized rates.

The method of Lagrange multipliers [32] gives the following equations for the mcvs of the r_is under the linear restriction:

$$r_i = \frac{N_i}{T_i + \varphi a_i} \quad a_0 + \sum a_i r_i = 0$$

where φ is the so-called Lagrange multiplier. In the case of a rate ratio u, these equation specialize to

$$r_{1j} = \frac{N_{1j}}{T_{1j} + \varphi u T_j^*} \quad r_{2j} = \frac{N_{2j}}{T_{2j} - \varphi T_j^*} \quad u = \frac{\sum r_{2j} T_j^*}{\sum r_{1j} T_j^*}$$

The special case of a rate difference yields the same equations for the r_{ij} (but with u set to 1), and then

$$d = \sum r_{2j} T_j^* - \sum r_{1j} T_j^* / T^*$$

This is the method that was used in the Phoenix cluster data example.

References

[1] Carver RP. The case against statistical hypothesis testing. Harv Educ Rev 1978;48:378–99.

[2] Salsburg DS. The religion of statistics as practiced in medical journals. Am Stat 1985;30:220–3.

[3] Poole C. Beyond the confidence interval. Am J Public Health 1987;77:195–9.

[4] Poole C. Confidence intervals exclude nothing. Am J Public Health 1987;77:492–3.

[5] Gigerenzer G. The superego, the ego, and the id in statistical reasoning. In: Keren G, Lewis C, editors. A handbook for data analysis in the behavioral sciences: methodological issues. Hillsdale, NJ: Lawrence Erlbaum; 1993. p. 311–39.

[6] Falk R, Greenbaum CW. Significance tests die hard. Theory Psychol 1995;5:75–98.

[7] Hunter JE. Needed: a ban on the significance test. Psychol Sci 1997;8:3–7.

[8] Feinstein AR. P-values and confidence intervals: two sides of the same unsatisfactory coin. J Clin Epidemiol 1998;51:355–60.

[9] Krueger J. Null hypothesis significance testing: on the survival of a flawed method. Am Psychol 2001;56(1):16–26.

[10] Edwards AWF. Likelihood. Cambridge, UK: Cambridge University Press; 1972.

[11] Royall RM. Statistical evidence: a likelihood paradigm. London, UK: Chapman & Hall; 1997.

[12] Sprott DA. Statistical inference in science. New York: Springer Verlag; 2000.

[13] Goodman S, Royall R. Evidence and scientific research. Am J Public Health 1988;78(12):1568–74.

[14] Clayton D, Hills M. Statistical models in epidemiology. Oxford, UK: Oxford University Press; 1993.

[15] Aickin M. Connecting Dempster-Shafer belief functions with likelihood-based inference. Synthese 2000;123:347–64.

[16] Shafer G. A mathematical theory of evidence. Princeton, NJ: Princeton University Press; 1976.

[17] Shafer G. Belief functions and parametric models. J R Stat Soc Series B 1982;44:322–52.

[18] Brillinger DR. The natural variability of vital rates and associated statistics. Biometrics 1986;42:693–734.

[19] Flood TJ, Meaney FJ, Vertz D, Laubham KA, Porter RS, Aickin M. Case–referent study of childhood leukemia in Maricopa County, Arizona 1965–1990. Phoenix: Arizona Department of Health Services; 1997.

[20] Flanders WD. Approximate variance formulas for standardized rate ratios. J Chronic Dis 1984;37:449–53.

[21] Breslow NE. Cohort analysis in epidemiology. In: Atkinson AC, Fienberg SE, editors. A celebration of statistics: The ISI centenary volume. New York: Springer-Verlag; 1985. p. 109–43.

[22] Frome EL, Checkoway H. Use of Poisson regression models in estimating incidence rates and ratios. Am J Epidemiol 1985;121:309–23.

[23] Aickin M, Chapin C, Flood T, Englender S, Caldwell G. Assessment of the spatial occurrence of childhood leukemia mortality using standardized rate ratios with a simple linear Poisson model. Int J Epidemiol 1992;21:649–55.

[24] Steering Committee of the Physicians' Health Study Research Group. Final report on the aspirin component of the ongoing Physicians' Health Study. N Engl J Med 1989;321(3):129–35.

[25] Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998;3(16):1236–8.

[26] Efron B. Why isn't everyone a Bayesian? Am Stat 1986;40:1–5.

[27] Morris CN. Comment on Efron. The American Statistician 1986; 40(1):7–8.

[28] Berger JO, Boukai B, Wang Y. Unified frequentist and Bayesian testing of a precise hypothesis. Stat Sci 1997;12(3):133–60.

[29] Lindley DV. Making decisions. London, UK: John Wiley & Sons; 1985.

[30] Tiainen M, Tammilehto L, Rautonen J, Tuomi T, Mattson K, Knuutila S. Chromosomal abnormalities and their correlations with asbestos exposure and survival in patients with mesothelioma. Br J Cancer 1989;60:618–26.

[31] Heath CW, Hasterlik RJ. Leukemia among children in a suburban community. Am J Med 1963;34:796–812.

[32] Critchley F, Ford I, Rijal O. Interval estimation based on the profile likelihood: strong Lagrangian theory with applications to discrimination. Biometrika 1988;75:21–8.