What is Bayesian statistics and why everything else is wrong

Michael Lavine
ISDS, Duke University, Durham, North Carolina

Abstract

We use a single example to explain (1), the Likelihood Principle, (2) Bayesian statistics, and (3) why classical statistics cannot be used to compare hypotheses.

1. The Slater School

The example and quotes used in this paper come from *Annals of Radiation: The Cancer at Slater School* by Paul Brodeur in **The New Yorker** of Dec. 7, 1992. We use the example only to make a point, not a serious analysis.

The Slater school is an elementary school in Fresno, California where teachers and staff were "concerned about the presence of two high-voltage transmission lines that ran past the school ...." Their concern centered on the "high incidence of cancer at Slater...." To address their concern, Dr. Raymond Neutra of the California Department of Health Services' Special Epidemiological Studies Program conducted a statistical analysis on the

> "eight cases of invasive cancer, ..., the total years of employment of the hundred and forty-five teachers, teachers' aides, and staff members, ..., [and] the number of person-years in terms of National Cancer Institute statistics showing the annual rate of invasive cancer in American women between the ages of forty and forty-four — the age group encompassing the average age of the teachers and staff at Slater — [which] enabled him to calculate that 4.2 cases of cancer could have been expected to occur among the Slater teachers and staff members ...."

For the purpose of our illustration we assume (1) that the 145 employees develop (or not) cancer independently of each other and (2) that the chance of cancer, θ, is the same for each employee. Therefore X, the number of cancers among the 145 employees, has a binomial $(145, \theta)$ distribution; we write $X \sim \text{Bin}(145, \theta)$. For any integer x between 0 and 145, $\Pr[X = x|\theta] = \binom{145}{x}\theta^x(1-\theta)^{145-x}$. The data turned out to be $x = 8$.

According to Neutra, the expected number of cancers is 4.2. Noting that $4.2/145 \approx 0.03$, we formulate a theory:

Theory A: $\theta = 0.03$,

which says that the underlying cancer rate at Slater is just like the national average. To address the concern at Slater school we want to compare *Theory A* to alternatives that would better account for the large number of cancers. To illustrate, we propose three additional theories. All together we have

Theory A: $\theta = 0.03$,
Theory B: $\theta = 0.04$,
Theory C: $\theta = 0.05$, and
Theory D: $\theta = 0.06$.

2. The Likelihood

To compare the theories we see how well each one explains the data. That is, for each value of θ, we use elementary results about binomial distributions to calculate

$$\Pr[X = 8|\theta] = \binom{145}{8}\theta^8(1-\theta)^{137}$$

which says how well each value of θ explains the observed data $X = 8$. The results are

$$\begin{aligned}\Pr[X = 8|\theta = .03] &\approx 0.036\\ \Pr[X = 8|\theta = .04] &\approx 0.096\\ \Pr[X = 8|\theta = .05] &\approx 0.134\\ \Pr[X = 8|\theta = .06] &\approx 0.136,\end{aligned}$$

or roughly in the ratio of 1:3:4:4. Thus we can make statements such as "*Theory B* explains the data about 3 times as well as *Theory A*"; *Theory C* explains the data slightly better than *Theory B*"; and "*Theories C* and *D* explain the data about equally well."

One point to notice is that $\Pr[X|\theta]$ is a function of two variables: X and θ. Once $X = 8$ has been observed, then $\Pr[X = 8|\theta]$ describes how well each theory, or value of θ, explains the data. It is a function only of θ; no value of X other than 8 is relevant. For instance, $\Pr[X = 9|\theta = .03]$ is irrelevant because it does not describe how well any theory explains the data. This principle is central to Bayesian thinking and is called the Likelihood Principal. The Likelihood Principal says that once X has been observed, say $X = x$, then no other value of X matters and we should treat $\Pr[X|\theta]$ simply as $\Pr[X = x|\theta]$, a function only of θ. A more complete explanation and many thought provoking examples can be found in Berger and Wolpert, 1988. The function $\ell(\theta) = \Pr[X = 8|\theta]$ is called the likelihood function.

The likelihood function says how well each theory explains the data and therefore contains all the information for distinguishing among theories based on the data. For some purposes computing the likelihood function is all that is necessary. But for other purposes it might be useful to combine the information in the Slater data with information from other sources. That is the subject of the next section.

3. A Bayesian Analysis

In fact, there are other sources of information about whether cancer can be induced by proximity to high-voltage transmission lines. Here we consider just two. First, there have been epidemiological studies, some showing a positive correlation between cancer rates and proximity and others failing to show such correlations. And second, there have been statements from physicists and biologists that the energy in magnetic fields associated with high-voltage transmission lines (purported to be the cause of increased cancer rates) is too small to have an appreciable biological effect.

The information is inconclusive. For the sake of illustration, suppose we judge the arguments on the two sides of the issue to be equally strong. We can summarize that judgement with a statement such as "*Theory A* is just as likely to be true as false" and express it mathematically as $\Pr[A] \approx 1/2 \approx \Pr[B] + \Pr[C] + \Pr[D]$. And suppose further that we have no information to suggest that any of B, C or D is more likely than any other. We can express that as $\Pr[B] \approx \Pr[C] \approx \Pr[D] \approx 1/6$. All together we have

$$\Pr[A] \approx 1/2 \qquad \Pr[B] \approx \Pr[C] \approx \Pr[D] \approx 1/6.$$

These probabilities are called the *prior* distribution. Their interpretation is that they summarize our knowledge about θ prior to incorporating the information from Slater.

An application of Bayes' Theorem, or simply the definition of conditional probability yields

$$\begin{aligned}\Pr[A|X = 8] &= \frac{\Pr[A \text{ and } X = 8]}{\Pr[X = 8]} \qquad (1)\\ &= \frac{\Pr[A \text{ and } X = 8]}{\Pr[A \text{ and } X = 8] + \Pr[B \text{ and } X = 8] + \Pr[C \text{ and } X = 8] + \Pr[D \text{ and } X = 8]}\\ &= \frac{\Pr[A]\Pr[X = 8|A]}{\Pr[A]\Pr[X = 8|A] + \Pr[B]\Pr[X = 8|B] + \Pr[C]\Pr[X = 8|C] + \Pr[D]\Pr[X = 8|D]}\\ &\approx \frac{(1/2)(.036)}{(1/2)(.036) + (1/6)(.096) + (1/6)(.134) + (1/6)(.136)}\\ &\approx 0.23.\end{aligned}$$

Similar calculations yield

$$\Pr[A|X=8] \approx 0.23 \quad \Pr[B|X=8] \approx 0.21 \quad \Pr[C|X=8] \approx \Pr[D|X=8] \approx 0.28.$$

These probabilities are called the *posterior* distribution. Their interpretation is that they summarize our knowledge about θ after incorporating the information from Slater. We can make such statements as "The four theories seem about equally likely." and "The odds are about 3 to 1 that the underlying cancer rate at Slater is higher than 0.03."

A Bayesian analysis uses the posterior distribution to summarize the state of our knowledge. The posterior distribution combines information from the data at hand expressed through the likelihood function, with other information expressed through the prior distribution.

4. A non-Bayesian Analysis

Classical statisticians, to test the hypothesis $H_0 : \theta = .03$ against the alternative hypothesis $H_1 : \theta > .03$, calculate the P-value, defined as the probability under H_0 of observing an outcome at least as extreme as the outcome actually observed. In other words,

$$\text{P-value} = \Pr[X=8|\theta=.03] + \Pr[X=9|\theta=.03] + \Pr[X=10|\theta=.03] + \cdots + \Pr[X=145|\theta=.03]$$

which, for the Slater problem, turns out to be P-value ≈ 0.07.

We claim the P-value should not be used to compare hypotheses because

1. hypotheses should be compared by how well they explain the data,
2. the P-value does not account for how well the alternative hypotheses explain the data, and
3. the summands $\Pr[X=9|\theta=.03], \ldots, \Pr[X=145|\theta=.03]$ are irrelevant because they do not describe how well any hypothesis explains any observed data.

The P-value does not obey the Likelihood Principle because it uses $\Pr[X=x|\theta]$ for values of x other than the observed value of $x=8$. The same is true for all classical hypothesis tests and confidence intervals. They do not obey the Likelihood Principle and cannot be used to compare scientific theories or hypotheses. See Berger and Wolpert, 1988 for a full explanation.

5. Discussion

Using the Slater school as an example we have illustrated the Likelihood Principle, a Bayesian analysis and a non-Bayesian analysis. In the interest of directness we have so far ignored several points which we now treat more fully.

- Our analysis used four discrete values of θ. A better approach is to treat θ as continuous with values between 0 and 1. The likelihood function is still $\ell(\theta) = \binom{145}{8}\theta^8(1-\theta)^{137}$. It is plotted in Figure 1.

Figure 1

Figure 1 about here

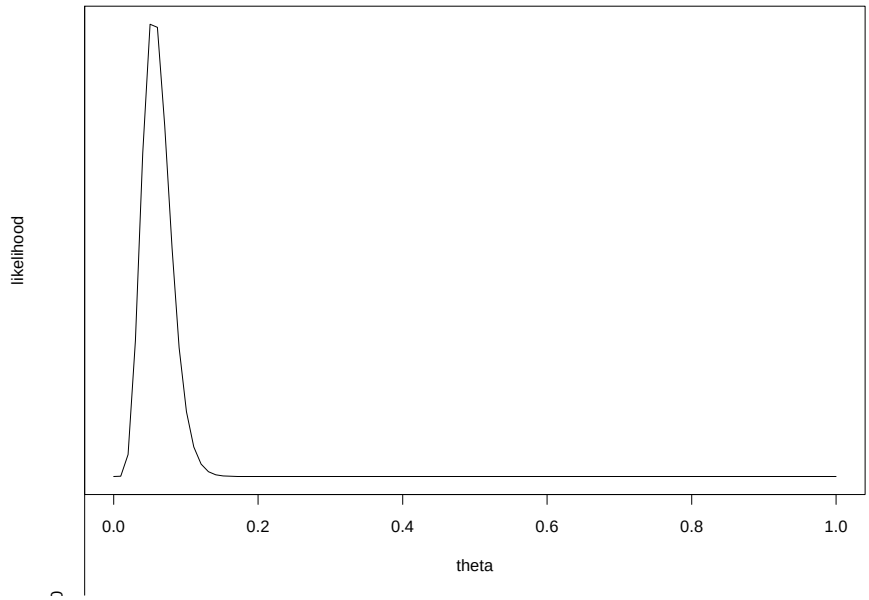

Figure 1. The likelihood function $\ell(\theta) = \Pr[X = 8|\theta] \propto \theta^8(1-\theta)^{137}$

A continuous treatment would entail the use of prior and posterior probability density functions $f(\theta)$ and $f(\theta|X = 8)$ rather than prior and posterior probabilities. The posterior density for any value of θ would be calculated as

$$f(\theta|X = 8) = \frac{f(\theta, 8)}{\Pr[X = 8]} = \frac{f(\theta)\Pr[X = 8|\theta]}{\int f(\theta)\Pr[X = 8|\theta]\, d\theta}.$$

A continuous treatment with a reasonable prior density would yield an answer quantitatively similar to that from our discrete treatment. In either case, the Likelihood Principle is correct and the P-value should not be used to compare values of θ.

- A Bayesian analysis divides information into two types. One type is the data being analyzed — 8 cancers in 145 employees in our example — which is encoded in the likelihood function. The other type is everything else and is encoded in the prior distribution. The prior distribution may be based on earlier studies of similar phenomena, as in the Slater example, on our best understanding of the phenomenon being investigated, as in the Slater example, on previous data from directly relevant studies, or on anything else we deem relevant.

  The Bayesian paradigm says that the investigator should use a prior distribution but does not say what that prior distribution should be. The investigator is free to choose any prior he or she desires. In principle the choice should accurately reflect the investigator's knowledge about the phenomenon under study. But however it is chosen, the choice of prior leaves an analysis open to charges of subjectivity.

  Bayesians reply to the charge in various ways. Some (See Berger and Berry, 1988, for example.) say subjectivity is good, that different scientists reach different conclusions because they have different priors, and that making the priors explicit (as opposed to the *apparent* objectivity of classical statistics) is a good thing. Others heed the call for objective analysis by proposing prior distributions that satisfy some criterion of objectivity and argue that such priors ought to be adopted and accepted as default in cases where true prior information is weak or where there is strong disagreement about what the prior distribution should be. See Kass and Wasserman, 1996 for a review.

  Yet another approach is to examine sensitivity of the posterior distribution to changes in the prior. Following this approach, a statistician may propose an entire class of prior distributions, or a set of deviations from an

initial prior, meant to approximate the set of all prior distributions that would be used by reasonable people. Each prior distribution in the class can be updated to create a class of posterior distributions which can be examined for sensitivity. As an example, one could compute upper and lower bounds on $\Pr[A|X = 8]$ over the set of all prior probabilities belonging to some reasonable class.

In the Slater example the posterior is highly sensitive to the choice of prior and the range of posterior probabilites would be large. And that's because there is not much information in the data to distinguish between the four values of θ. The likelihood function varies only by a factor of about 4 to 1 for the values of θ we consider. In other problems the likelihood function can vary by many orders of magnitude. There, the sharpness of the likelihood function would overwhelm distinctions between prior distributions in quite a large class.

The question of sensitivity is a question of whether the likelihood function is sharply peaked relative to priors in a reasonable class. If the likelihood is sharp then the posterior is insensitive to the choice of prior. In the Slater example, the likelihood function clearly points to values of θ less than about 0.15. Had our prior information been weaker, leading us to consider values of θ in the whole unit interval, then the likelihood function would have appeared sharply peaked over the range of *a priori* reasonable values of θ. Our actual prior restricted attention to $\theta \in [.03, .06]$, a range over which the likelihood function is relatively flat.

- Bayesian analyses are sometimes criticized as philosophically unsound. Specifically, the Bayesian analysis treats θ as though it were a random variable whereas classical analysis treats θ as a fixed constant, albeit unknown. And the truth, at least the notional truth behind the binomial sampling model for X, is that θ is fixed, not random. So there is no meaning to the concept of θ as a random variable. That is, θ either *is* or *is not* equal to 0.03. There is no such thing as $\Pr[\theta = .03]$. Furthermore, probability is defined as the limit, as the number of trials gets large, of a relative frequency in a sequence of events. Since there is only one instance of θ and not an infinite sequence of θ_i's, quantities such as $\Pr[\theta = .03]$ have no meaning.

  The Bayesian reply is twofold. First, treating θ as a random does not mean that we believe θ is random. Rather, it expresses the state of our knowledge about θ. A sharply peaked distribution expresses strong knowledge about θ; a relatively flat distribution expresses weak knowledge. The distribution describes us and our knowledge, not some fundamental property of θ.

  And second, the relative frequency definition of probability is too confining. Partly because the degree of knowledge interpretation is so useful, probability can and should be interpreted broadly enough to accomodate it. Like any other mathematical construct, its definition is purely mathematical, not tied to the physical reality of whether there exists an infinite sequence of events. Any mathematical construct can be applied wherever it is useful. And the Bayesian paradigm is useful.

- The distinction between Bayesian and classical statistics would be of only philosophical interest if both approaches led to similar conclusions. So it is worthwhile investigating whether they do. We saw in the Slater example that the classical P-value is approximately 0.07, or very close to the widely accepted critical value of 0.05, below which null hypotheses are rejected. In other words, a classical analysis of the Slater data is very close to rejecting $H_0 : \theta = .03$. On the other hand, both the likelihood function and the Bayesian analysis say that the evidence against $\theta = .03$ is not very strong — only about 3 or 4 to 1. The two approaches disagree.

  The question of when Bayesian and classical analyses yield similar results, or when a classical P-value can be interpreted as approximately a Bayesian posterior probability, has been studied in some generality. A good place to begin dipping into the literature is the pair of papers Casella and Berger, 1987 and Berger and Sellke, 1987 and the accompanying discussion.

- In assessing the evidence at Slater school Dr. Neutra "went on to estimate that as many as a thousand out of a total of eight thousand schools in California might be situated near power lines, and he concluded that 'clusters like this in schools near power lines could well occur by chance' ". Neutra's reasoning went something like this: Of the thousand or so schools near power lines, it is the one or ones with the largest number of cancer cases that will call itself to our attention as a possible cancer cluster. If in fact $\theta = 0.03$, then it is

quite possible that at least 1 of the 1000 schools will have as many as 8 cancers. So the fact that one school has 8 cancer cases is not much evidence against $H_0 : \theta = 0.03$.

To see how this reasoning might be made more formal, suppose there are 1000 schools in California situated near power lines each having 145 employees at risk. Let X_i be the number of cancer cases in school i and define $Y = \max(X_i)$. Then a classical P-value would be

$$\begin{aligned} \text{P-value} &= \Pr[Y \geq 8|\theta = .03] \\ &= 1 - \Pr[Y \leq 7|\theta = .03] \\ &= 1 - \prod_i^{1000} \Pr[X_i \leq 7|\theta = .03] \\ &\approx 1 - 0.9282974^{1000} \\ &\approx 1. \end{aligned}$$

Even if there were only 10 schools instead of 1000, the P-value would be $1 - 0.9282974^{10} \approx 0.52$ and Neutra is right: there is not much evidence against H_0, at least according to the classical notion of evidence.

A Bayesian analysis would begin with the likelihood function

$$\begin{aligned} \ell(\theta) &= \Pr[Y = 8|\theta] \\ &= \Pr[Y \leq 8|\theta] - \Pr[Y \leq 7|\theta], \end{aligned}$$

which yields, for our four θ's,

$$\ell(.03) \approx 1.3e^{-14}, \quad \ell(.04) \approx 1.9e^{-60}, \quad \ell(.05) \approx 1.7e^{-156}, \quad \ell(.06) \approx 5.1e^{-308},$$

which favors *Theory A* over the other three theories by 46 orders of magnitude or more. The likelihood function for all values of θ is plotted in Figure 2. It has its maximum near $\theta = 0.015$ which is the value of θ most strongly supported by this likelihood function. In fact, this likelihood function strongly suggests that θ is less than about 0.02.

Figure 2

Figure 2 about here

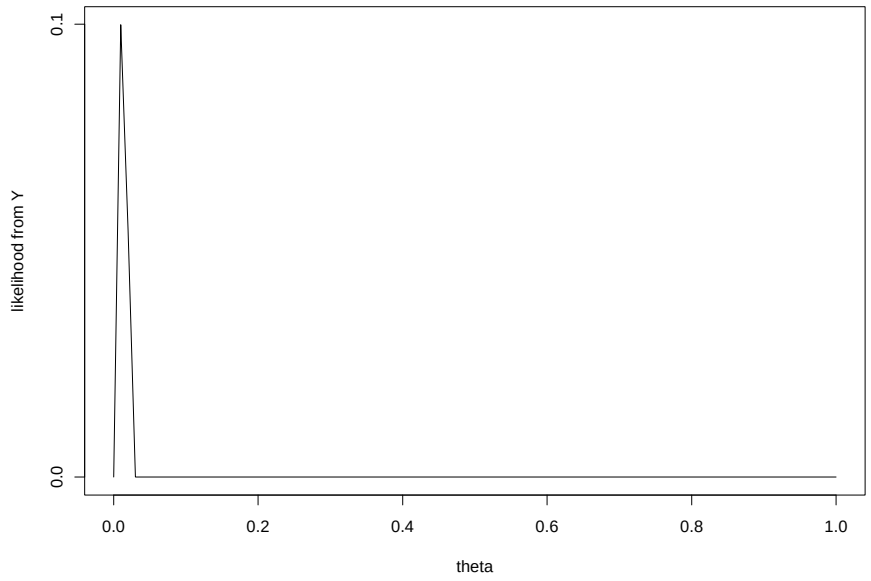

Figure 2. The likelihood function $\Pr[Y = 8|\theta]$

Of course our formalism is only approximate. There are not exactly 1000 California schools near power lines, they don't all have 145 employees and, most importantly, Slater is not necessarily the one with the largest number of cancers. A better analysis of the data would probably yield a conclusion somewhere between that reached by treating Slater alone and that reached by considering only the maximum of the X_i's.

References

Berger, James O. and Berry, Donald A., Statistical analysis and the illusion of objectivity, American Scientist, *76,* 159-165, 1988.

Berger, James O. and Sellke, Thomas, Testing a point null hypothesis: The irreconcilability of P values and evidence, Journal of the American Statistical Association, *82,* 112-122, 1987.

Berger, James O., and Wolpert, Robert L., *The Likelihood Principle, 2nd ed.,* 208 pp., Institute of Mathematical Statistics, Hayward, Calif., 1988.

Brodeur, Paul, Annals of radiation: The cancer at Slater school, The New Yorker, *Dec. 7,* 1992.

Casella, George and Berger, Roger L., Reconciling Bayesian and frequentist evidence in the one-sided testing problem, Journal of the American Statistical Association, *82,* 106-111, 1987.

Kass, Robert E. and Wasserman, Larry, The selection of prior distributions by formal rules, Journal of the American Statistical Association, *91,* 1343-1370, 1996.

M. Lavine, Institute of Statistics and Decision Sciences, Duke University, Box 90251, Durham, NC 27708-0251. (email: michael@stat.duke.edu)

This preprint was prepared with AGU's LaTeX macros v4. File paper formatted March 9, 2000.