

BAYESIAN METHODS: AN INTRODUCTION FOR PHYSICAL OCEANOGRAPHERS

Joseph B. Kadane

Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213

“You could not step twice into the same river, for new waters are ever flowing on to you,” Heraclitus, as quoted in Bartlett (1980).

ABSTRACT

The Bayesian approach to statistics is a conceptually simple method of treating uncertainty. It involves modeling uncertainty with probability, and conditioning on such data as become available. Because of its flexibility, there are many styles of application. Using the same examples as George Casella's paper in this volume, I discuss how this Bayesian method approaches such problems.

1. A GENERAL INTRODUCTION TO BAYESIAN IDEAS

Most statistical analyses begin with some data, denoted x , and a parameter, denoted θ . These may be discrete or continuous, and may have vector, matrix, or more complex structures. For the purposes of this section, the nature of x and θ do not matter, but in application they are very important.

The mechanism generating the data is called the likelihood function, and is written $f(x|\theta)$. Here f may be a probability mass function, if x is discrete, or a probability density, if x is continuous. In both cases it describes the probabilistic behavior of the data, x , for a fixed value of the parameter θ . The second part of a statistical model is a prior distribution $\pi(\theta)$, which again might be a probability mass function if θ is discrete, or a probability density if θ is continuous.

These two ingredients determine the joint distribution of x and θ as follows:

$$h(x, \theta) = f(x|\theta) \pi(\theta) \tag{1}$$

Once the data x are observed, the laws of probability prescribe how the conditional distribution of θ given x is to be calculated:

$$g(\theta|x) = \frac{h(x, \theta)}{p(x)} = \frac{h(x, \theta)}{\int_{\Omega} h(x, \theta) d\theta} = \frac{f(x|\theta) \pi(\theta)}{\int_{\Omega} f(x, \theta) \pi(\theta) d\theta} \quad (2)$$

The distribution is called the posterior distribution of θ , because it is the distribution of θ after having observed x . Thus the import of the data is to change the distribution of θ from the prior, $\pi(\theta)$, to the posterior, $g(\theta|x)$. Everything in this paper is a discussion or an application of these ideas.

The essential idea here is the use of probability to express uncertainty. Having decided to do that, formula (2) follows from formula (1) by very simple and non-controversial steps.

One important matter is the interpretation given to probability here; whose probabilities are these? Although there are some Bayesians who would give other answers, the dominant answer now, (and the one to which I subscribe) is that these probabilities are subjective, and reflect the opinion of the writer, or opinions the writer believes others hold. Bayesians do not come to this view gladly. We wish there were a way to guarantee that the equations written capture the objective truth, but such guarantees do not seem possible. We observe in science disagreements in which none of the sides has made a provable, mathematical error. The progress of a science might then be thought of as the development of informed opinion on a subject.

The name "Bayesian" incidentally, is in honor of Rev. Thomas Bayes, an eighteenth century English minister and "natural philosopher." He found the principle now embedded in (2), and hence this way of thinking about and doing statistics is named for him.

Finally, note that the quantities x and θ are simply random variables with some joint distribution, although one is written with a Roman letter and one with a Greek letter. If one began with the joint distribution $h(x, \theta)$ and learned θ , the posterior on x given θ would be $f(x|\theta)$, and would represent what was known about x after θ had been learned. Thus the model is symmetric in x and θ , although to encourage intuition it is customary to think of the former as data and the latter as a parameter.

In the remainder of this paper, I discuss Casella's iceberg example in section 2, breaking waves in section 3 and bubble data in section 4. In section 5, I give my views on frequentism and the possibility of compromises between Bayesian and frequentist ideas. Finally in section 6 I give my conclusions.

2. THE ICEBERG DATA

Before discussing the elements of a model, I think it is most useful to get the question straight, which corresponds most closely to Casella's steps 5 and 6.

Everyone with a modicum of liberal arts training knows about “compare and contrast” questions. The point is that there are always similarities and always differences. Either can be celebrated.

Looking at the graph of relative frequencies of icebergs, it is clear that most of the story here is in the similarity of the patterns. But one could also look for differences, for the “contrast.” If you ask me to believe that the frequencies, month-by-month, of icebergs are exactly the same to an arbitrary number of decimal places, I must reply that I cannot. Thus I regard Casella's null hypothesis as foolishness. I put zero prior, and hence zero posterior, on its truth. So I need a better question.

Suppose instead I ask what I consider to be a better question: how far apart are $\theta^N = (\theta_1^N, \dots, \theta_{12}^N)$, frequencies of icebergs south of 48°N, and $\theta^S = (\theta_1^S, \dots, \theta_{12}^S)$ frequencies of icebergs south of Grand Banks? I could measure this in a variety of ways, such as

$$d_1 = \sum_{i=1}^{12} (\theta_i^S - \theta_i^N)^2$$

and

$$d_2 = \sum_{i=1}^{12} |\theta_i^S - \theta_i^N|.$$

Now a prior on (θ^S, θ^N) and a likelihood on counts given (θ^S, θ^N) will yield a posterior, and I can compare what I thought about a distance measure d before I saw the data with what I think after I see the data.

This is a measure of what I have learned from the data about how different θ^S and θ^N are. So this is how I think a modern Bayesian would structure the problem.

What are the data? If they are a complete census of all icebergs from 1900 to 1926, then we know that the hypothesis H_0 is false. So suppose that these are a random sample of a larger population of icebergs. How do these particular icebergs come to be in the data set? Because someone observed them, presumably. Is it reasonable to assume that icebergs have the same chance of being observed, regardless of month? I should think that the summer months are easier to observe than the winter months, because more observers will be around and weather conditions are better. The critical issue is whether the observation bias, I should believe, is the same for the two areas. Thus if θ_i^N is the probability of a random iceberg in region N being there in month i , and η_i^N is the probability of its being observed, then $\eta_i^N \theta_i^N$ is the probability of an iceberg being there and being observed, and the frequencies observed have probabilities

$$\psi_i^N = \frac{\eta_i^N \theta_i^N}{\sum_{i=1}^{12} \eta_i^N \theta_i^N} \quad (3)$$

Note that if I believe that iceberg generation is constant by month ($\theta_1^N = \dots = \theta_{12}^N = 1/12$), then ψ_i^N gives information about η_i 's, the observation intensities. Which interpretation to give to the data depends on what you believe. The Bayesian method can't say which is right or wrong, but it does provide for (and insist on having) a full, probabilistic statement of what the investigator believes. Reasonable people need not agree on these matters. This allows readers to judge those beliefs, and possibly approximate the calculations the reader might do with his own beliefs. The argument affects the likelihood as well as the prior; both are subjective. Note that I now have more parameters than I have data points. Hence a frequentist treatment of such a model is impossible. Frequentist analysis thus encourages you not to delve too deeply, not to ask such questions.

Even the above formulation is too simplistic, since it assumes that the probabilities θ and η are constant over years. Since during the period of the data collection both the sinking of the Titanic and World War I occurred, it is hard to believe that η , the observation probabilities, were constant. A careful modeling of the data would have to take this into account and would treat skeptically claims of a vast increase in icebergs in the latter half of the period.

Priors on θ are important for the inference in question. The first tool a statistician would think of in this regard is a Dirichlet distribution (a multivariate Beta distribution) on the vector $(\theta_1, \dots, \theta_{12})$. However the Dirichlet has some unattractive features for this purpose, principally that it treats all the months symmetrically, without making use of the adjacency of them. I would prefer to think of a continuous model in which the critical parameter is an angle, which could be given a Fisher/von Mises distribution, which is like a normal (or Gaussian) distribution for angles and has as hyperparameters a central tendency, ν , and a measure of spread, τ^2 . Thus ν would indicate the direction of greatest iceberg intensity, thinking of time through the year as circular. Looking at the data, perhaps a good estimate would be $\hat{\nu} = \text{May } 10$. The measure of spread, τ^2 , would indicate how peaked the distribution is. To complete the model, a prior on both ν 's (North and South), and both τ 's would be necessary.

In these terms, I think that the quantity $d_3 = \nu^S - \nu^N$ would be useful as an alternative to d_1 and d_2 . The main advantage of d_3 is that its units are days, which is natural and might have a physical interpretation.

It is now time to turn attention to inference, in Casella's sense. The frequentist statement, applied to this situation, is that if the null hypothesis of no difference were true, and the experiment were repeated an infinite number of times with the same parameter values, the

data would be as or more extreme in only $1 - 0.977 = 0.023$ proportion of the cases. Thus the conclusion is that either the null hypothesis is false or something unusual has happened. But frequentists cannot say which, or even give a probability on which. Note that 0.023 is **NOT** the probability that the null hypothesis is false. Not believing H_0 , I don't find this frequentist probability calculation useful.

Casella's version of a Bayesian treatment of this problem is not recognizably Bayesian to me. All it does is condition on both margins in the table (total icebergs observed by month, and total icebergs observed N and S), and then calculates a frequentist p -value. The only warranted statement from his calculation is again that either something unusual happened (with probability less than 0.006), or the null hypothesis is false. But again he cannot say which, nor give a probability for it. I see no justification for Casella's statement that "the probability of the null hypothesis is 0.994."

One interesting way to think about these statistical procedures is to ask what happens as the sample size grows large. In frequentist statistics, no sharp null hypothesis (such as $\theta^N = \theta^S$) is significant if the sample size is small. However as the sample size grows large, every such hypothesis will turn out to be significant. Thus significance measures sample size more powerfully than it does the extent to which the "straw-man" null hypothesis is false. Since better measures of sample size are generally available, significance testing is, in my judgment, not very useful.

By contrast, in the Bayesian analyses I have been discussing, as the sample size grows, the posterior distribution of whichever d you like will converge to a point. You will then effectively know how far from true the hypothesis of equality is, by your chosen measure. What to make of it then depends on what you are doing scientifically, whether you want to emphasize the "compare" or the "contrast" side.

I have written at some length about the iceberg data because it gives me an opportunity to illustrate how Bayesian thinking helps me to model a process. The important points, in my view, are

- The frequentist hypothetical infinite sequence of identical circumstances is a figment of their imaginations.
- Priors and likelihoods are important because they correspond to something real: what you believe about the data.
- Frequentist ideas can get in the way of good modeling because you can easily get too many parameters.
- Testing sharp null hypotheses is generally a foolish undertaking, because they are each, to a greater or lesser degree, wrong.

3. BREAKING WAVES

The principal difference between this example and the previous one is that the null hypothesis is no longer sharp. That is, inferential attention focuses on a single parameter b , and whether $b \leq 4$ or $b \leq 3$.

Unlike Casella, I would not center the prior at the hypothesized value, but would instead have it represent my honest opinion, or my view of what some other scientific opinion might honestly be. My summary would be the posterior distribution on the parameter b , from which one could calculate $P(b \leq 4 | \text{data})$, $P(b \leq 3 | \text{data})$, and any other probabilities that might be of interest.

4. BUBBLE DATA

This is similar to the breaking wave data, except that there are several regressions instead of a single one. Such a model is called hierarchical. These have proven useful in a very wide variety of domains.

At the first level, the log bubble population is modeled as

$$N(Z) = A_u + b_u Z + \epsilon$$

where ϵ is Gaussian with mean 0 and variance σ^2 , $N(Z)$ and Z are observed, and a_u , b_u , and σ^2 are parameters. At the second level, there might be a bivariate Gaussian distribution on (a_u, b_u) with some mean (a, b) and some covariance matrix Σ . Finally, a third level would specify a prior in (a, b, Σ, σ^2) . Such a model is complete if each quantity mentioned has a distribution. A complete model permits a Bayesian analysis, conditioning on the observed data, as a Bayesian should. Interest may focus on the parameters at any level: (a_u, b_u) might be of interest, or (a, b) , or any of the others.

5. ON COMPROMISES

As explained just above, a complete hierarchical model is fully Bayesian, and not a compromise. "Empirical Bayesian models" are incomplete; they forget the upper levels of a hierarchy and treat the remaining parameters frequentistically. There is no advantage to a Bayesian in doing so. If the posterior distribution is peaked in the parameters taken to be fixed, there may not be too much loss in this method as an approximation. However in great generality estimates of uncertainty derived from the "empirical Bayesian method" will be underestimates of the same measure derived from a fully Bayesian approach, because parameters are taken as known with certainty that are not known with certainty.

To be successful, a compromise must offer something to each party. Empirical Bayes methods do represent a compromise on the frequentist side, because some (but not all) parameters are treated as random variables with distributions. But to a Bayesian, this “compromise” offers no advantages over a straight Bayesian analysis.

6. PRAGMATIC CONCLUSIONS

In principle, I am convinced that Bayesian ideas are the right way to structure thinking about inference. We are still learning how to use this powerful tool in an effective way. If the problem you have can't be done now in a Bayesian way, then you have to work your problem as best you can, approximating a fully Bayesian analysis.

Even the pre-Socratic philosopher Heraclitus understood that frequentism does not apply to oceanographic problems.

Research supported by NSF SES-8900025 and DMS-9005858, and ONR N00014-89-J-1851.

REFERENCES

- Bartlett, John (1980). *Familiar Quotations: A Collection of Passages, Phrases and Proverbs, traced to their sources in Ancient and Modern Literature*, 15th edition, Little-Brown & Co., Boston, p. 70.
- Casella, G. (1993). *Illustrating Frequentist and Bayesian Statistics in Oceanography*, this volume.