Notes on Inference: Misconceptions about Bayesian and Frequentist Inference

Andrew C. Eggers

April 8, 2005

1 Introduction

The superficial similarities between Bayesian and frequentist methods of parameter estimation, and the actual substantial overlap among their methods of generating models and summarizing results, can make it hard to get a firm grip on what the real differences are. As I've studied these methods this year (first MLE and then Bayesian data analysis), I've struggled to define the distinction between the two styles of analysis. In the process, I drew what I now realize are a number of incorrect comparisons between the likelihood and Bayesian methods of inference. Here I try to straighten myself out, with the hope that it will help a few other people solidify their understanding.

2 Likelihood and probability

Misconception 1: The likelihood function is basically the same as a posterior distribution.

I think I was led astray here by two separate lines of thinking, neither one entirely wrong.

- **Misconception 1a: If you choose a "flat" prior,[1] the posterior distribution is exactly the likelihood function (or sampling density), up to a constant of proportionality.**
- **Misconception 1b: Frequentists sample from the likelihood function just like Bayesians sample from the posterior.**

The first point is technically right but fails to recognize how subjective and particular the "flat" prior really is; the second point conflates the concepts of probability and likelihood. In this section it should become clear why both points are mistaken. In the last section I'll talk in more detail about the second point, dealing with sampling and the likelihood function.

2.1 Inverse probability

Bayes' rule tells us:

$$p(\theta|y) \propto p(y|\theta)p(\theta) \tag{1}$$

In words, the probability that parameters take on a particular value, given the data, is proportional to the probability of observing the actual data given that parameter value multiplied by the *a priori* probability that the parameter takes on that value. Therefore if you choose a noninformative prior such that $p(\theta) = k$, the sampling density ($p(y|\theta)$) is exactly proportional to the full posterior. You can then normalize

[1] A particular flat prior . . . read on.

the sampling density and treat it as a probability density of the parameters θ given the data and model.

The question is how generally this neat trick can be used, ie when it's appropriate to use the flat prior $p(\theta) = k$. It is a question that appears to have caused quite a ruckus in the history of statistics. The idea that this flat (constant) prior can be generally applied, rendering the posterior proportional to the sampling density, is called "inverse probability"; in our notation, it can be expressed as

$$p(\theta|y) \propto p(y|\theta). \tag{2}$$

Certain early statisticians apparently believed that in the general case, when the researcher did not wish to impose prior constraints on model parameters, it was possible to use this flat prior and rely on inverse probability for inference: normalize the sampling density (i.e. $p(y|\theta)$) and treat it as a density expressing the distribution of parameter values, given the model and data. This was essentially my misconception, as well. While it's true that there exists a prior belief that makes the posterior distribution proportional to the sampling density (or likelihood function), it is important to realize that there is only one such prior belief, and it does not convey total ignorance about the parameter's prior distribution.

This was exactly the point made in one of Ronald Fisher's influential early papers (Fisher, 1922), which helped put the idea of inverse probability to rest.[2] If one of the parameters being estimated is σ^2, for instance, the "flat" prior would be to

[2]He also seems to be making the point (offhandedly) that complete prior ignorance about the value of a parameter is unreasonable, given that the researcher is claiming enough knowledge to be able to determine the sampling density. But I'm not sure about this point.

say that $p(\sigma^2) = k$, which asserts that the true value of the variance is as likely to be in one range $d\sigma^2$ as any other. But this prior does *not* express ignorance about the standard deviation: σ is as likely to be between 0 and 1 as it is to be between 1 and $\sqrt{2}$. In Fisher (1922)'s example, if the value p in a binomial model is transformed to be a function of a parameter θ (with a flat prior on θ), the new posterior does not match the old (although the maximum likelihood estimate is unchanged).[3] Essentially restating Fisher's points, King (1998) discusses inverse probability as a "failed model of inference" (pp. 16-21).

To these points about the singularity of a "flat" prior could be added another: a "flat" prior, although referred to as "noninformative," is not really noninformative at all, in the usual sense of the word. If you assert that $p(\theta) = k$, you are not claiming total ignorance about the values θ could take on. Rather, you are asserting that you believe it to be equally likely that θ is between 0 and 1 and between 1×10^{23} and $1 \times 10^{23} + 1$. It is hard to imagine any applied situation in which you would really have a "flat" prior of this kind.

2.2 Likelihood

Having dismissed the idea of a generally valid inverse probability, Fisher basically renamed the left side of the equation and forged ahead. The sampling density $p(y|\theta)$ is not proportional to $p(\theta|y)$, but it is, he declared, proportional to $L(\theta|y)$, which he called the "likelihood." He made clear that he used the term likelihood not "as a loose synonym of probability," but rather quite the opposite: he used the term

[3] Actually, I think the posterior before and after transformation would actually be quite close if not identical in his example. I must be doing something wrong.

to distinguish $p(y|\theta)$, which is not a probability statement about θ, from $p(y|\theta)$, which is. (The fact that the terms likelihood and probability are commonly used as loose synonyms added to my confusion here.) He defined likelihood as the relative frequency with which you would observe a particular outcome given one set of parameters compared to another set. In other words, $L(\theta|y) \propto p(y|\theta)$. By Bayes' Law, we could convert this into a probability statement if we knew the probability of observing any particular θ (i.e., if we had a prior distribution). In Fisher's view we usually wouldn't. But we can find the θ with the highest likelihood ($\hat{\theta}$, the maximum likelihood estimate), which is the parameter value that maximizes the probability of observing the data we gathered. It turns out that this estimator has some good properties, one of which I'll discuss below. Most importantly, it is a consistent and unbiased estimator of the true parameter value.

3 So how do you analyze a maximum likelihood estimate?

Since the likelihood function is not a density, you can't draw conclusions about the probability that θ will take on any particular value, nor can you sample values of θ from it. There are two common ways to analyze it, though, both of which help clarify the differences between analyzing a Bayesian posterior and likelihood function, and both of which I misunderstood until recently.

3.1 Likelihood ratio

Here there's another misconception I had. **Misconception 2: I thought a likelihood ratio was the ratio of the probabilities of two outcomes**, but it isn't. What is it? As the name indicates, it is the ratio of two *likelihoods*:

$$L(\theta_1|y) \propto k(y)p(y|\theta_1) \tag{3}$$

$$L(\theta_2|y) \propto k(y)p(y|\theta_2) \tag{4}$$

$$\frac{L(\theta_1|y)}{L(\theta_2|y)} = \frac{k(y)}{k(y)}\frac{p(y|\theta_1)}{p(y|\theta_2)} \tag{5}$$

$$= \frac{p(y|\theta_1)}{P(y|\theta_2)} \tag{6}$$

The last line is the ratio of two probabilities, but in terms of θ_1 and θ_2 these expressions are likelihoods. A natural use of the likelihood ratio would be to set θ_2 equal to $\hat{\theta}$ and choose a θ_1 that is constrained in some way (e.g., one of the sub-parameters is set to zero). The question you would like to pose is whether you were much more likely to observe the data you saw with the full model ($\hat{\theta}$) than with the restricted model. If not, then the restrictions didn't make much difference.

The test statistic is R, defined as follows:

$$R = -2ln\frac{{L_R}^*}{L^*} \sim f_{\chi^2}(r|m) \tag{7}$$

where r is the observed value of R and m is the number of restricted parameters. Essentially the test is whether the difference between the true log-likelihoods of the restricted and unrestricted models could be zero. If that possibility cannot be

rejected, then the restricted model cannot be rejected.

Another way of thinking about what's going on: The χ^2 distribution with m degrees of freedom is the distribution of observed likelihood ratios under the null hypothesis that the likelihood is the same at the two points. We are asking how likely it would be to get the observed value of r (or more extreme values) if the likelihoods were in fact the same. Thus while it is not possible to say how much more likely (in an absolute sense) one spot on the likelihood function is than another, we settle for saying how much confidence we can have in rejecting the null hypothesis that their likelihood is the same.

If the likelihood ratio did establish a ratio of probabilities, as I at first thought, it would be possible to convert a likelihood function into a density: once all points have been converted into a fractional probability of the mode, we can normalize the probability of the mode such that all the probabilities add to one. Meanwhile Ronald Fisher would turn in his musty grave.

3.2 The distribution of $\hat{\theta}$

Misconception 3: I thought the multivariate normal density around $\hat{\theta}$ from which we sample parameter values was an approximation of the likelihood function. As should be clear from what we've seen above, approximating the likelihood function and sampling from it would not make any sense, since it is not a probability density.

When we sample around the likelihood function, we're sampling from the asymptotic distribution of $\hat{\theta}$ itself. In other words, we are observing the potential observed

values of $\hat{\theta}$, given that the actual observed value $\hat{\theta}$ was the true value. Contrast this with sampling from a Bayesian posterior: in that case, the posterior tells us where the actual value is likely to be, given the data. In the MLE setting we are looking at what values for $\hat{\theta}$ we would expect to observe, conditional on the value we actually observed being the true value.

It turns out that the (negative) inverse of the Hessian of the log-likelihood evaluated at $\hat{\theta}$ describes the asymptotic distribution of the $\hat{\theta}$ about the actual parameter value (θ). In particular, this matrix is the variance covariance matrix of the multivariate normal that describes the sampling distribution of $\hat{\theta}$, which we center on the observed value (since $\hat{\theta}$ is an unbiased estimator of θ.

Why does the Hessian of the log-likelihood function have this desirable property? From quickly reading a few sources, I only understand the outline of the reasoning. Kennedy (427-428) contains a very cursory treatment of the subject; Johnson [4] goes into more depth. The contention seems to be that

- there is a lower bound, called the Cramer-Rao Lower Bound, for the variance-covariance matrix of any asymptotically unbiased estimator;
- one asymptotically unbiased estimator, the maximum likelihood estimator, will have the smallest variance-covariance matrix; therefore
- the variance-covariance matrix of $\hat{\theta}$ asymptotically approaches the Cramer-Rao Lower Bound.

The next point is that the Cramer-Rao Lower Bound can be estimated by the method

[4] http://cnx.rice.edu/content/m11266/latest/

mentioned above (calculating the negative inverse of the Hessian matrix) but I don't yet understand why this is possible.

Gary King's explanation of this in his Gov 2001 class[5] takes a different approach, which I think was what led me to think we were approximating the likelihood function in the first place. The end result is still that the sampling distribution of $\hat{\theta}$ is normal and has the variance covariance matrix obtained by inverting the Hessian. Along the way he develops the point that the multivariate normal obtained by inverting the Hessian approximates the log-likelihood function at its maximum. What is not yet clear to me is why we would want to sample from a distribution that approximates the log-likelihood function, given that the latter is not a density. Of course the answer is that this distribution is also the sampling distribution of $\hat{\theta}$, but this is precisely the link that I don't understand from his notes.

4 Conclusion: the prior, again

The differences between Bayesian and frequentist analysis that I have brought up here ultimately trace back to the question of the prior. If we are willing to assert a prior distribution, it seems to me, there is no advantage to sticking with maximum likelihood methods: the Bayesian posterior has a more direct interpretation, does not rely on large sample properties of any distribution, and describes more of the parameter space. The crucial question is therefore, "In what cases are we willing to make an assertion about the prior distribution of parameters?"

One answer is that, since choosing a prior is so arbitrary, the bar should be very

[5] http://gking.harvard.edu/g2001syl/inference.pdf

high. Frequentist approaches guarantee an unbiased parameter estimate; adopting a prior may make posterior analysis more straightforward, but it invites criticism and the suspicion that prior information drives the results. Many researchers include a sensitivity analysis to show that their findings do not depend on the choice of a specific prior, which is a good step toward dispelling this criticism that one's priors are arbitrary. But it also begs the question of why the researcher chose to forego a more straightforward frequentist approach in the first place. Often the reason is that Bayesian priors add structure to a complex model that would otherwise be unidentified. The utility of Bayesian methods is clear here.

On the other hand, perhaps we should not be scared off by the apparent arbitrariness of choosing a prior. This is only one of many arbitrary elements to any statistical enterprise. Most important, choosing a sampling distribution is necessarily arbitrary. The "curse of dimensionality" means we must rule out almost all possible ways in which a set of data might have been produced, instead focusing on an extremely narrow and simplistic subset. We can make reasonable assumptions and decisions in doing this, of course, but the same is true of choosing a Bayesian prior.