Econ. 556b Spring 1996 C. Sims

# Bayesian and Classical Inference

Probability statements made in Bayesian and classical approaches to inference often look similar, but they carry different meanings. Because in so many practical circumstances the statements look the same, econometricians are often not careful about the different meanings, or even not too sure what the differences are. In some circumstances the statements themselves start to diverge sharply, and since such circumstances (strongly nonlinear models, non-stationary time series) have become increasingly prominent in econometrics recently, it is important to gain a complete understanding of the differences in meaning. We will start with a very simple example.

Suppose we are helping implement the use of a testing device on an assembly line. As product items (toasters) move down the line, they pass under the device, which flashes blue, green or red. Some of the toasters are defective in a way that the device is designed to detect. If the toaster is defective, is properly placed on the conveyer belt, and is not obscured by the operator's hand or dust on the device's lens, the device is supposed to flash blue, and it does so under these conditions 99% of the time, otherwise flashing green. If the toaster is OK, the device is supposed to flash green, and again if the toaster is properly placed and there is no interference from dust or the operator's hand, it flashes green for good toasters 99% of the time, otherwise erroneously flashing blue. Five per cent of the time, independent of whether the toaster is OK or not, the toaster is misplaced on the belt or something blocks the device from getting a proper reading, and in that case it flashes red.

Classical probability statements in this context are meant to describe the characteristics of procedures based on readings from the device. For example we might construct an hypothesis test as a basis for a procedure to decide whether to retrieve toasters from the conveyor belt and toss them in a reject bin. We might consider, say, rejecting any toaster that produced a blue or red flash. We could describe this as applying an hypothesis test for the null hypothesis that the toaster is OK, with a rejection region consisting of the blue and red flashes and an acceptance region consisting of a green flash. The test has a probability of type I error, and therefore a significance level, of $.05+(.95\times.01)=.0595$, because that is the probability of getting either a red or a blue flash when the toaster is actually OK. It has a probability of type II error of $.95\times.01=.0095$, because that is the probability of getting a green light when in fact the toaster is defective. It has power $.05+.95\times.99=.9905$, because that is the probability of getting a blue or red light when the toaster is defective. Note that we might also consider a different test -- one that rejects only when a blue light is observed. This would have a significance level of .0095, a probability of type II error of .0595, and power .9405..

We might want to construct a confidence region for the status of the toaster. A confidence region maps observations on the data (here the color of a light flash) into subsets of the space of unknown parameters (here whether or not the toaster is OK). There are just two exact confidence regions available here (plus some trivial ones). A .9405 confidence region consists of "OK" when the light is green, "not OK" when the light is blue, and "neither" when the light is red. It is easy to check that this region contains the truth with probability .9405 regardless of whether the toaster is in fact OK or not OK. A .9905 confidence region consists of "OK" when the light is green, "not OK" when the light is blue, and "both" (or "either") when the light is red. This region contains the truth with probability .9905 regardless of whether the toaster is OK or not. Of course there are also the trivial 100% confidence region that contains both "OK" and "not OK" regardless of the light color and the trivial 0% region that contains neither regardless of the light color. There are also the two "perverse" confidence regions that are the complements of the .9905 and .9405 regions described above.

The behavior of the .9905 and .9405 regions when the light is red may seem puzzling. But since the red light gives us no information about whether the toaster is OK or not, the region estimator naturally must either exclude or include both possibilities at the same time in response to the red light.[1] This leads to assertions of the form "with 99.05% confidence the toaster is either OK or not OK", or "with 94.05% confidence the toaster is neither OK nor not OK". There is nothing paradoxical about these statements if we keep firmly in mind that the "99.05% confidence" assertion is a probability statement only about the procedure we are using, averaging its behavior over many potential applications, not a statement about the probability of "OK" or "not OK" given the particular data (a red light) at hand in this instance. Of course any reasonable statement about the probability of the "OK" or "not OK" status given observed data would have to give 100% probability (not 99.05%) to the status being either OK or not OK, and would have to give zero probability (not 94.05%) to the status being neither OK nor not OK. But it is perfectly possible, and sometimes as here unavoidable and optimal, that a classical confidence region sometimes contain the whole space of possible true values or be empty.

While all these classical procedures are associated with probability statements about how the procedures behave across repeated measurements, independent of the true state being measured, Bayesian inference aims instead at making probability statements about the true state of the world given a particular measurement or set of measurements. A Bayesian might, for example, assert when the light is green that she is 99% sure the toaster is OK, when the light is blue that she is 99% sure that the toaster is not OK, and when the light is red that the toaster could be OK or not OK, with each possibility having probability .5. These probability statements follow from taking the probabilities of "OK" and "not OK" as equal before the instrument reading, accepting the stated description of the behavior of the instrument, and applying the Bayes rule. For example, when the light is blue the calculation is

$$P\left[\text{OK}\,\middle|\,\text{blue}\right]=\frac{P\left[\text{blue}\,\middle|\,\text{OK}\right]\cdot P\left[\text{OK}\right]}{P\left[\text{blue}\,\middle|\,\text{OK}\right]\cdot P\left[\text{OK}\right]+P\left[\text{blue}\,\middle|\,\text{not OK}\right]\cdot P\left[\text{not OK}\right]}=\frac{.01\times.5}{.01\times.5+.99\times.5}=\frac{.005}{.5}=.01\,.$$

A Bayesian can quote different probabilities given different data; classical probability statements concern the behavior of a given procedure across all possible data. Classical inference

[1] In a setup like this where data take on discrete values it is not generally possible to construct exact confidence regions -- regions that have the same probability of containing the truth regardless of what the truth is. We have set up this example so that exact confidence regions exist, but any random region that, say, included "OK" but excluded "not OK" when the light is red would necessarily fail to be exact.

eschews probability statements about the true state of the world (the parameter value -- here "not OK" vs. "OK") and treats only data (here the light color) as random.

As in many applications, the Bayesian we have cited here reaches conclusions that are usually not much at variance with those of the classical statistician. The Bayesian is 99% sure the toaster is OK when she sees green, while the classical statistician is either 99.05% "confident" or 94.05% "confident" of the same assertion when the light is green, depending on what he promises he would do if he observed a red light. The big discrepancy between the two types of inference occurs only in relatively rare samples -- when the light is red. But now suppose that we know from a previously conducted study with more accurate instruments that 99.99% of the toasters coming down the conveyor belt are in fact OK. This means that, when the light is not red, we will get a false blue light from an OK toaster .01*99.99%=.9999% of the time and a true blue light from a bad toaster .99*.01%=.0099% of the time. Thus blue lights occur as mistakes approximately 100 times more often than they occur as correct indications of a defective toaster. The Bayesian cited above who is 99% sure the toaster is defective when she sees the blue light arrives at that conclusion by taking the probability that the toaster was defective as .5 before seeing the instrument reading. If she instead has read the earlier study of the population of toasters and believes a particular toaster currently being tested to be a random draw from that earlier study's population, she would place 99.99% probability on the toaster's being OK before she saw the meter reading, and a blue light would reduce that probability, but only to about 99%.

What effect does knowledge of the probability of defects in the population of tested toasters have on classical probability statements, tests, and confidence regions? None. Those statements are constructed so that they are valid regardless of whether all the toasters are OK, none are OK, or they are OK and not OK in any mixture generated by any method, random or non-random. This can lead to apparently paradoxical results. If the toasters are in fact 99.99% OK, then the classical 99.05% confidence region will fail to contain the true parameter value about 99% of the time when the light is blue. This does not affect its validity as a 99.05% confidence region, because blue lights occur rarely. Even though the classical region is giving the "wrong signal" when the light is blue, its still gives the "right signal" 99.05% of the time, averaged across all possible types of data -- blue, red and green.

# I. Subjectivity vs. Objectivity

The sensitivity of Bayesian conclusions in this example to knowledge or beliefs about the probability of defects available before the instrument reading is observed is both a strength of the Bayesian point of view and the main reason many statisticians dislike it. The classical statistician is in the paradoxical position of asserting with "99.05% confidence" something that turns out to be wrong most of the time in the crucial case where it might lead to discarding a toaster. But the Bayesian's conclusions can be extremely different according to what she believes about the distribution of defects in the population before she sees the instrument reading. Whatever their disadvantages, the classical procedures and statements are "objective" -- they depend only on the known stochastic properties of the measuring instrument, not on a priori beliefs about the thing being tested or estimated.

In fact, we have skewed the example in favor of the Bayesian approach by suggesting that the "prior information" comes from a reliable previous study. If there were such a study, a classical statistician might be willing to treat it as objective data and thereby open up the possibility of different and more reasonable confidence intervals, combining the results from the study with the results from the current instrument readings. So even classical statisticians might agree that the Bayesian procedures are reasonable in this case. But suppose there is no such study. The Bayesian approach asserts that the only truly objective part of the process of inference is the rules by which new information modifies old beliefs. There can be no probability statement that does not depend on the old beliefs. Sometimes there may be a great deal of previous "objective" evidence on which to found these old beliefs, as in this example where we postulated the previous study. But even where there is not, conclusions still depend on the old, or "prior" beliefs, regardless of what those beliefs are founded on.

Some Bayesians and some of the seminal Bayesian statistical literature assert aggressively that all probabilities are subjective. While this is a defensible position, and perhaps worth arguing vigorously for rhetorical effect, it makes the Bayesian methodological stance less persuasive by ignoring a distinction that most scientists see as patently useful. The probability of heads on a coin flip, even if it is not exactly .5, is no doubt close to .5. Nearly everyone will agree on this, and we also agree how to check the assertion if there is any doubt. This is quite different from, say, my own beliefs about the chance that I will like escargots before I have ever eaten them. Even if I decide the chance is about .5, outside observers will not generally agree with me, and there is no way to check who is right -- even if I do eventually eat them, all we find out is that I did or didn't like them. We can't repeat the experiment to see "how often" I liked them on the first try. The coin-flip probability is objective, the escargot-liking probability is subjective. This is a distinction worth making.

Bayesian scientists who report statistical results are relying on the existence of objective components in readers' probability distributions. Generally they use a "model", meaning that they postulate a probability distribution for observable data conditional on some "parameters". Readers who find the model totally implausible will not be interested in the reported results. By choosing the model well, we can summarize a part of the probability distribution for what determines the data that is common across many potential readers, thereby producing an interesting paper. The "parameters" are then likely to represent aspects of the determination of the data about which there is considerable disagreement.

## II. Bayesian Reporting: The Likelihood Principle

Notice that in our formula (1) the probability for the status of the toaster given a blue light depends only on the prior unconditional probabilities P[OK] and P[not OK] and on the conditional probabilities of the observed blue light given the two possible states: P[blue|OK] and P[blue|not OK]. In reporting the results of a study that had observed a blue light to an audience whose views of the unconditional probability that the toaster was OK might vary, the statistician could summarize the results with the two numbers P[blue|OK] and P[blue|not OK]. (Actually, only one number, the ratio of these two, matters.) A Bayesian can therefore report results objectively, in the sense of doing so independently of any prior, by reporting P[blue|OK] and P[blue|not OK].

This idea -- the likelihood principle -- which may at first seem simple, helpful in freeing Bayesians from the curse of subjectivity, and obvious, actually by itself still implies a big gap between Bayesian and classical methodology. It implies, for example, that once we have observed a blue light, our conclusions are unaffected by what the probability of a red light was. This is apparently reasonable. Once we see the blue light, we know the instrument did not fail. Whether the instrument fails (shows a red light) 50% of the time or .1% of the time, it behaves the same way when it does not fail: giving the correct color of light 99% of the time. But as we have seen above, classical statements about significance level, power, etc. are strongly affected by the probability of the red light.[2]

The example that perhaps brings out most strongly the contrast between the implications of the likelihood principle and of usual classical procedures is that of stopping rules. Suppose some toaster factory claims it produces toasters with zero defects. We bring our instrument to the factory to check the claim. We test 10,000 toasters, finding 120 cases of blue lights. (For simplicity here, suppose that there are never any red lights -- the machine shows blue for not-OK toasters 99% of the time and green for OK toasters 99% of the time.) A classical statistician might then calculate the probability of 120 or more blue lights when there are actually no defective toasters, finding it to be .028. He might then reject the null hypothesis that the factory actually produces toasters with zero defects. But now suppose instead that the study was carried out differently. Instead of an objective tester who tests 10,000 toasters, then cites conclusions, the test is constructed by a spy for another company who actually simply continues testing toasters until she has obtained enough blue lights so that the probability of that many blue lights is .028 under the null of no defects, if calculated in the ordinary way assuming that the sample size was fixed in advance. Suppose further that the spy happens to have obtained her target result after testing 10,000 toasters, so that she also has obtained 120 blue lights in 10,000 tries. The classical statistician, knowing that this was how the study was done, concludes that the probability of obtaining the apparently "significant" result was 1.00, regardless of whether the factory produced any defective toasters or not.[3] Therefore the conclusion from the study is very different from (and less negative for the factory than) the study done by the objective researcher.

[2] In our example, there is what is known as an "ancillary statistic": a function of the data whose distribution does not depend on the unknown parameter (OK or not OK). This statistic is "whether or not the light was red". Its distribution is by assumption the same whether or not the baseball is OK. Most classical statisticians would agree that one should conduct analysis conditional on ancillary statistics, meaning that, e.g., significance levels of hypothesis tests should be conditional on the ancillary statistic. This would make the significance level of the test that rejects when the light is blue or red 1% in the case of the blue light and 100% in the case of the red light, removing the worst discrepancies between Bayesian and classical conclusions. However, if the probability of a red light given an OK baseball is .051 and that given a not-OK baseball is .049, we no longer have an ancillary statistic. Most classical statisticians would not want to condition on any statistic here, but the example would still behave almost exactly as our original example. See Berger and Wolpert for a more extensive discussion of classical approaches to conditioning on part of the data.

[3] An important question here is whether the spy's procedure is guaranteed to stop after some finite number of tested toasters. It is in this case. With stopping rules that are not guaranteed to quit in finite time, Bayesian conclusions can be affected by what is assumed about reporting of trials that are not complete at the date of reporting.

The likelihood principle can be shown to imply that the occurrence of 120 blue lights in 10,000 toasters has the same implications -- because it implies the same likelihood function -- regardless of which of the two procedures was followed. Many statisticians decide after reading this last sentence that they have now had enough of the likelihood principle. But careful consideration of the example shows that the problem is with classical methods, not with the likelihood principle. Conventional classical approaches would test the null hypothesis of zero defects at some standard significance level -- say .05 -- regardless of sample size. For the same reason that the spy's procedure is guaranteed to terminate in finite time, a classical statistician who tested at the .05 level repeatedly as more and more toasters were tested would with probability one eventually reject the null hypothesis of zero defects, and would continue to do this infinitely often as sample size grew to infinity. Bayesians regard hypothesis testing at fixed significance levels independent of sample size as inherently unreasonable. A Bayesian who, say, began with prior beliefs that the true rate of defects in the factory was zero with probability $\alpha$ and non-zero, uniformly distributed between 0 and $g < 1$, with probability 1-$a$ (a so-called "spike-slab" prior) would actually find both samples strong evidence in favor of the factory having a zero defect rate under a fairly wide range of $a$ and $g$ choices. The Bayesian sees the large number of toasters that the spy had to test to achieve her stopping criterion as the same kind of evidence in favor of a low defect rate as the low blue-light rate of 1.2% (compared to the expected 1%) in the fixed-sample-size experiment.