In response to Ellison (2003): an alternate explanation of frequentist versus Bayesian inference

Ellison (2003) argues that the fundamental difference between frequentist and Bayesian statistics is the view of probability as proportions (frequencies) versus viewing probability as subjective belief. For many non-Bayesians, I think this statement will be misleading or at least apt to be misunderstood. Bayesian statistics is widely used in the hard sciences, such as physics, signal analysis, and astronomy. It is not a "soft" subjective branch of statistics, rather it provides a way of making "reverse" inference: working backwards from a particular observation to put weights on the models (or parameters) that could have produced that observation. It is a method for solving the 'optimal decision making' problem, that is to figure out the odds that the optimal rational player would put on different model parameters if they had to place bets. It is solving not the problem "what is the frequency of my data under a particular set of parameters (the hypothesis)" but rather "what is the optimal odds function for the parameter given the data"?

The following are different statements of the basic difference from two physicists posting on the `sci.physics.research` listserv.

> "The basic difference between Bayesians and frequentists is this: Bayesians condition on the data actually observed, and consider the probability distribution on the hypotheses; they believe it reasonable to put probability distributions on hypotheses and they behave accordingly. Frequentists condition on a hypothesis of choice and consider the probability distribution on the data, whether observed or not; they do not think it reasonable to put probability distributions on hypotheses (in their opinion, one hypothesis is true, the rest are false, even if we do not know which is the case), and they behave accordingly." Bill Jeffreys, U Texas.
>
> "The frequentist fixes the true value of the parameter, and calculates from that a probability distribution for the data. The philosophy here is that the parameter must have a true value, even though we happen to be ignorant of it, but the data could be any number of things, even though we happen to have gotten one particular thing this time. This fits in with thinking of probabilities as relative frequencies of data in hypothetical ensembles.
>
> The Bayesian fixes the true value of the data, and calculates from that a probability distribution for the parameter. The philosophy here is that we know what the data is, even though we might just as easily have gotten something else, but we are in ignorance about what the parameter is, even if we agree that the parameter does have a precise value. This fits in with thinking of probabilities as degrees of belief concerning unknown facts." Toby Bartels, UC Riverside

These statements better reflect how I think about the differences between frequentist and Bayesian inference. Here I delve into these ideas more in depth and explore how I think

about Bayesian inference versus frequentist inference. I'm neither a Bayesian nor a frequentist, and this discussion is meant to be entirely pedagogical rather than persuasive.[1]

Inference about random variables[2]

Both frequentist and Bayesian inference are based on inductive inference, inference based on observations of random events. "The majority of species on earth are in the class Insecta. Given a randomly selected species, it's probably an insect." is an example of inductive inference. In its inherent uncertainty, it differs from deductive inference in which inferences must logically follow with certainty. "No insects are mammals. A beetle is an insect. Therefore a beetle is not a mammal." However, Bayesian and frequentist differ in the type of inductive reason they use. Before describing the difference, I need to define some terminology about random occurrences (whether frequentist or Bayesian):

> frequency: the fraction of times an event occurs in a really large set of identical replicates. Although it's easy to describe frequency with this physical analogy, it's really a conceptual idea that puts a quantitative 'weight' on the occurrence of different data. It's not really possible to replicate all the conditions under which the data were generated, yet we can still think about this conceptually. Some type of conditional is almost always stated (or implied): "the frequency of A given I do my data collection in a particular way" or "frequency of A given model B true".
>
> odds: Assume I observe some outcome *x*. Let the frequency with which model A produces *x* be *a*, and the frequency with which model B produces *x* be *b*, I'll refer to the ratio of *a* to *b* as the odds of A to B. The odds function is a relative measure (*a* to *b*) so can be normalized (divided by) a constant without loss of meaning. If *a*/*b* is big, it implies that A is more credible than B. It says nothing about the absolute frequency with which *x* appears under A or B.

---

[1] This discussion is directed to those who are more familiar with the frequentist view of statistics. It's my attempt to help those folks get past "wrong, wrong, wrong" when they hear Bayesian statisticians describe things, and get to "oh, that's what's they're doing." This discussion will be largely non-mathematical. There are ample references for math elsewhere. For other perspectives written for ecologists see Hilborn & Mangel 1997, chap 1; Ellison 2003; Lewin-Koh et al. 2004; Goodman 2004. Note, it's not my intent to persuade anyone of what statistics to use since I have no idea what questions my reader thinks about and how my reader prefers to tackle questions. My impression from following engineering and physics listservs is that some stick with a Bayesian or frequentist approach uniformly, while others choose their approach depending on the objective at hand and which approach is more facile and widely accepted for said objective. I tend to take the latter approach, except that unlike many Bayesians, I take a very critical view of what constitutes a proper prior, in fact the priors I would acceptable are closer to what frequentists would accept – even if I use the prior in an entirely Bayesian way to solve P(param|data) and defining probability as the 'optimal rational player function'. As opposed to using Bayes theorem, to solve the frequentist question P(data|param) and defining probability as the 'frequency of N'. Isn't there just one probability? No, there are a multitude. A probability is a construct that obeys the axioms of probability theory. But I'm getting ahead of myself….

[2] My thinking on this has been influenced by Jayne's "Probability theory: the logic of science" chap 1 (http://bayes.wustl.edu/etj/prob/book.pdf) and Polya's books on plausible reasoning.

I will avoid use of the term 'probability' since frequentists and Bayesian statisticians use terminology differently, and this leads to endless confusion when trying to talk about both statistical approaches at the same time[3].

Frequentist statistics bases inference on frequency and is concerned with defining the hypotheses under which the observed data would occur with low frequency. Those hypotheses can then be rejected:

> A implies that *x* is unlikely (low frequency)
> *x* is true
> A not credible

Bayesian statistics uses inductive inference based on the odds, and it is concerned with the estimating the relative frequency of the data *x* under different hypotheses:

> A implies that *x* occurs with frequency *a*
> B implies that *x* occurs with frequency *b*
> *x* is true
> The odds of A being true to B being true are *a* to *b*
> If $a >> b$, it means A is more credible than B (all else being equal[4]).

Frequentist methods are focused on rejecting hypotheses that are outright unlikely to have produced the observed data. Bayesian methods are focused on the relative frequency with which different hypotheses produce the same observed data. This relative frequency, which in Bayesian statistics appears in the form of the ratio of likelihoods (or summations of likelihoods), is a measure of the data support[5]. Whether the observed data are infrequent under the hypotheses is irrelevant. If this last statement is puzzling, remember that in Bayesian inference, the support that data confer on hypotheses is based on measures of *relative* frequency of the data under those hypotheses.

Why does frequentist and Bayesian inference work?
Here's one simple, simple explanation that you'll find from philosophy of logic[6].

Suppose you have some set of empirically distinguishable hypotheses, one of which, h_t, is true (but we don't know which). Let f_i be the frequency of *x* if h_i is true. The weak law of large numbers [convergence in distribution] tells us that as the sample size of *x* goes to infinity, f_i for all h_i untrue goes to 0. Thus only the true hypothesis remains. This is why frequentist inference works. Collect enough data, eventually all untrue hypotheses are rejected and but the true hypothesis isn't.

---

[3] For a mind-twisting essay on what probability actually means, check out http://plato.stanford.edu/archives/sum2003/entries/probability-interpret/#3.5
[4] Yes, I've left out the prior. I know that.
[5] What about priors?(!) I'll get to that later.
[6] http://plato.stanford.edu/entries/logic-inductive/

Suppose again you have some set of empirically distinguishable hypotheses, one of which, h_t, is true (but we don't know which). Let f_i be the frequency of x given h_i true. Once again, the weak law of large numbers tell us that as the sample size of x goes to infinity, the odds on h_t relative to the untrue hypotheses go to infinity. This is why Bayesian inference works. Collect enough data, eventually only the true hypothesis has non-zero odds.

The reverse inference problem: the importance of priors

In the reverse inference problem, you want to make inference about whether A or B is true given some observed data, x. What you know is that if A is true, x will be observed p_a% of the time, while if B is true, x will be observed p_b% of the time. What are the odds of A being true versus B being true? You know that one is true, you just don't know which one. So diagrammatically if we denote the frequency of x as f(x):

$$A \rightarrow f(x|A) = p_a \quad f(\text{not } x) = 1\text{-}p_a$$

$$B \rightarrow f(x|B) = p_b \quad f(\text{not } x) = 1\text{-}p_b$$

Let's say that x is the result of a coin toss from a coin your friend just pulled out of his pocket. x = heads. The question is whether the coin (which you didn't examine) is A, two-sided with heads on one side and tails on the other, or B, heads on both sides. Now $p_b/p_a = 2$. So are the correct odds, that the coin is two-headed versus normal 2-to-1? Of course not. I've never seen a two-headed coin in my life, and unless I have reason to suspect trickery on my friend's part, my prior assumption is that it's really unlikely the coin is two-headed. The odds have to include a prior assumption about the odds on A versus B:

$$[1] \qquad \text{(new odds on B versus A given } x) = \text{(prior odds on B vs A)} \times \frac{p_b}{p_a}$$

In this case, my prior odds on two-headed vs. two-sided are really small, and my new odds are still highly in favor of two-sidedness. Just observation of one head, isn't enough to overcome my prior assumption about the severe unlikelihood of a two-headed coin.

What if my friend proceeded to do 20 more flips in a row and each and every one came up heads? Now I start to get suspicious. I'm not sure that there isn't something fishy about the coin. The data are forcing me to update my assumptions about the coin. My friend then proceeds to do 500 flips and all are heads. At this point, I know there is something fishy about the coin. The odds I put on it being two-headed versus two-sided are overwhelming.

If you hadn't figured it out already, Eq. 1 is Bayes theorem (rearranged a bit) used in a Bayesian way[7]: the prior odds are the ratio of the so called prior probability of B, π(B),

[7] By that I mean, the type of coin was not a random variable. I could have said that my friend reached into a bag of 100 coins and pulled one out, flipped it, and it was heads. Could I reject the hypothesis that the

to the prior probability of A, $\pi(A)$. I used the term 'odds' so that the frequentist definition of 'probability' = 'frequency' would not automatically throw my frequentist reader off track. When I'm feeling particularly rebellious against Bayesian 'probability', I just think of $\pi(B)$ as the 'odds on B'/('odds on B' + 'odds on A'). Yet despite my strong comfort with the frequentist definition of probability, when I wrote this, my natural inclination was to write 'what is the *probability* that the coin is two-headed?' even though in this example the coin is either two-headed or not and its type is not a random variable in the frequentist sense.

Epilogue: I ask my friend to show me the coin, and it is a completely normal two-sided coin. Turns out my friend could do even or odd flips at will and simply looked at the top of the coin before flipping[8].

Translation of common Bayesian/frequentist arguments
There are few common examples that Bayesians use to convince frequentists the error of their ways. These examples invariably leave frequentists saying 'wrong', 'wrong', 'wrong' and the Bayesian amazed and astounded that the frequentist can hold such bizarre views. This is my top three list of Bayesian arguments that leave frequentists dumb-founded, along with my attempt at translation for both.

1. *The coin-flipping example*
A friend reports to me that he flipped a coin 12 times, obtaining 9 heads and 3 tails. He asks, is the coin fair or not? The frequentist calculate the tail area, i.e., the proportion of times one would get 9 or more heads in 12 flips (in an *infinite* series of identical 12-flip trials), given that the coin is fair. Pr(heads >= 9 in 12 flips) = 0.075, which is not rejected at the 5% level (2-tailed test). I report this to my friend.

But my friend tells me I have done this wrong. The reason is, that instead of deciding in advance to toss the coin 12 times and record the number of heads and tails, he decided in advance to toss the coin until he obtained 3 tails, and to record the number of heads and the total number of tosses. The probability is now different since we have a waiting time problem (negative binomial) Pr(# flips to get to 3 tails >= 12) = 0.0325, which is rejected at the 5% level.

To the Bayesian, this is sheer nonsense. The analysis changed depending on what was in your friends head even though the actual data are the same! With a Bayesian analysis, the posterior on p=1/2 is the same regardless of how your friend flipped.

---

bag has only two-headed coins at the α level? Now the coin type is a random variable, and knowing the number of coins in the bag, I can clearly talk about the probability, in a frequentist way, that the coin is two-headed. To solve this problem, I use Bayes theorem but as a frequentist and I'm asking about the frequency of coin types in the bag.

[8] If as a frequentist, you're inclined to say 'ha, this shows the problem with Bayesian statistics', I would like to point out that the frequentist would have rejected the hypothesis that the coin is two-sided. The problem is not the method of statistical inference, but rather the formulation of the possible hypotheses in the first place which should have been more specific: 'two-sided and random-flipping'. Careful and proper formulation of hypotheses/models is important for statistical analysis, period.

Translation:
The goal of the frequentist's experiment was to test the null hypothesis. The absolute frequently with which the data occur under the experiment is critical to the frequentist. If the observed data occur easily, then the null cannot be rejected. In the case of the coin flipping, the absolute frequency of 3 tails out of 12 is higher if you flip until you get to 3 tails versus if you just flip to 12. That's why the frequentist rejects under one type of coin flipping and not the other. Testing null hypotheses may seem weird to some Bayesians, but there is a long tradition within empirical fields for doing science this way and this really is what some scientists intend with a particular experimental test.

Although, Bayesians often use this example to argue that Bayesian statistics conditions only on the data, this is not true. The likelihood function depends not only on the data but also how the data were collected. The likelihood for p=1/2 is different in the two experiments. It's just that in this particular case the likelihood differs by a constant. This constant falls out when we take the likelihood ratio, and the Bayesian is concerned with the ratios of the frequency of data (likelihood ratio) not absolute frequency. If my friend instead said, actually I can determine whether the coin lands head or tail and I decided ahead of time to flip 12 times with 3 tails, well, clearly the Bayesian analysis will must be changed – based on what was in the friends head. In this case, the likelihood is independent of p and the data provide no information on the value of p.

2. People want to know the probability that the parameter is in a particular interval, thus Bayesian inference is more natural and thus superior.

The frequentist is likely to grumble that people *want* significant results but does not mean we should adjust our statistical framework to provide this.

Translation:
When Bayesians use this argument they are usually referring to the frequent mis-interpretation of frequentist confidence intervals as a statement about where the parameter is likely to be. Frequentist CIs are supposed to be interpreted relative to the null hypothesis test statistic only[9]. If they do not cover the null, reject the null. One example, the Bayesians like to bring up is that the CIs for the mass of the neutrino include negative values. However, to the frequentist the CI is not a statement about the true mass of the neutrino, it's statement about whether we can reject the hypothesis that the neutrino's mass is 0 (or whatever the null is). It's a statement about the power of the experiment to reject the null.

Bayesians are right that scientists often need to estimate parameters of models. The Bayesian or likelihood (if you're allergic to priors) framework is a natural way to do this and to show parameter estimation uncertainty in a uniform framework. Frequentist statistics isn't designed for that. Just like Bayesian statistics isn't designed for rejecting

[9] There is a rather subtle, and I would say, fascinating difference between correct versus accurate frequentist CIs. Correct CIs show proper coverage: CI (a random variable) covers true parameter $(1-\alpha)\%$ of the time. Accurate CIs do that plus they properly show the observed data's power to reject the null hypothesis. See appendix so some references and discussion about this.

hypotheses. It's true that when using Bayesian approaches you have to be careful to make sure that the candidate model(s) can actually produce your data; otherwise, you get the best parameter estimates for a bad model and your predictions will be bad. But good model selection is an integral part of proper practice in Bayesian statistics, and all texts discuss this.

While Bayesians are right that their framework is more, often much more, facile for specifying parameter estimation uncertainty, it often seems that Bayesian proponents are unaware that many scientists do experimental science where they have no underlying model. I was at talk by Mary Power the other day. It was on some experimental tests of stream ecosystem functioning and the idea was to test the response of various ecosystem components to addition or removal of particular elements. This was an experiment to test whether there was an effect. How do you write this down as a parameter estimation problem and why would one want to? Frequentist statistics allows one to analyze this experiment without specifying the underlying model. We can ask 'Is the mean significantly different?' – without having to specify any likelihood functions. I should note that I have discussed this with Bayesians, and they always insist that classical tests (such as some type of ANOVA) do assume a model for the data and thus a likelihood function. I have never been successful at arguing that classical frequentist approaches are 'model-free' and that maximum-likelihood estimation is philosophically controversial for similar reasons as Bayesian statistics – that being that a model is being assumed when the underlying model is unknown.

3. The bad statistician argument

Both Bayesians and frequentists like to bring up examples that show how each side is fatally flawed. The examples typically involve what would be considered poor statistical practice within each respective field.

In summary, frequentist or Bayesian?

Surprising as this may be to my Bayesian friends, I often feel the need to ask 'What is the frequency of my data under a null hypothesis?' Frequentist confidence intervals is the natural way (for me) to do this – even if there is a Bayesian equivalent. I have clear understanding of what the frequentist CI means and that's what I want to answer this question. At the same time, I have to do projections, i.e. to estimate where the population will be in the future. For this question, estimation, the Bayesian approach feels right. I want a measure of data support for different parameters values and I want to weight or average[10] those values in an appropriate way and weighting by the posterior or by the likelihood ratio is well supported. The question I am still trying to understand is whether to use likelihood ratios alone or to use them in a Bayesian construct with a prior. Currently it's difficult for me to conceive of the meaning of the likelihood ratio in an absolute way without reference to a prior.

But both frequentist and Bayesian statistics seems to me to involve a 'hand-waving' step: for frequentists this is the 'consistent as *n* goes to infinity' step since *n* is never infinite.

[10] I can imagine weighting versus averaging depending on the exact needs of the analysis.

For Bayesians, the hand-waving is in prior specification - typically. Usually people specify vague priors and say that that's ok as long as the prior doesn't influence the posterior. Each statistical camp seems to view the other's hand-waving step as fatal.

If Bayesian, subjective versus objective?

At first glance this may seem like a no-brainer. Obviously, we want to be objective and "let only the data speak!" Before taking such a rigid position, let me describe an example of how subjective priors have been used to help interpret experiments to measure the mass of the neutrino (from Efron's 2005 lecture, "Bayesians, Frequentists, and Physicists"). The mass of a neutrino is really small, close to 0, and when physicists measure it the estimates come up negative about half the time due to measurement error. If the estimate is sufficiently negative, then the classically calculated confidence intervals are entirely negative; the upper bound on the mass of the neutrino is negative. There is nothing incorrect about this since the definition of a α-CI is "if the CI is constructed with such a procedure, (1-α)% of the constructed CIs will contain the true value." It should be pretty clear that the entirely negative CIs are one of those 5% of cases where the CI doesn't include the true value. Correctness notwithstanding, physicists found CIs that only include patently impossible values for the mass of the neutrino to be unsatisfying. In a Bayesian analysis, one can easily incorporate restriction on the mass of the neutrino by specifying a prior with 0 weight in negative masses. What you get from the Bayesian analysis is a "credible interval" in which (1-α)% of the area of the likelihood function is found. It never goes negative. You've gained a consistent measure of data support for the neutrino mass given the experimentally measured value. Note though you've given up the notion of an interval in which the true parameter appears with fixed frequency[11]

Ok, so that's a little example of an informative prior based on physical constraints. One can easily come up with lots of examples like that. However, physicists also argue for the importance of subjective priors in the process of scientific discovery. The idea is that a scientist (of any statistical bent) does an experiment (or otherwise acquires data), and then analyzes that data and decides on the next step. That decision on the next step is based on the scientist's prior beliefs. Think of the situation of a group of physicists talking about the results of their neutrino experiment and discussing how to design the next experiment. If you imagine that discussion, it's got involve how the results of this experiment updated whatever prior beliefs (based on theory, logical arguments, prior experiments) that they have about the mass of the neutrino. But just because you use Bayesian statistics with subjective priors to help you in the process of scientific discovery doesn't require that you publish results with CIs calculated using priors – you might still argue that it's proper to do just likelihood profiles (say) of some sort. Some arguments[12] for a mixed approach are Feldman & Cousins (1998) and Efron (2005) [I'm positive there are lots more pertinent ones, Berger?].

---

[11] It's important to remember that Bayesian (1–α) credible intervals are not the same as frequentist confidence intervals. A (1-α) credible interval will NOT cover, in general, the true value (1-α)% of the time (cf Feldman & Cousins 1998 article on a unified approach to frequentist CI construction for example). Rather it is a statement about the RELATIVE probability of the data if the parameter is in the credible interval versus not in the interval (modified appropriately by the prior weight that you put on this ratio).

[12] There are many out there and these are probably not the best, nonetheless.