Introduction to Bayesian Statistics

I'll begin with a categorical statement: observed data have neither uncertainties nor errors.

If you had never thought about these issues before, that might strike you as reasonable. The data are the data; you measure what you measure, so why should there be uncertainties or errors?

But you almost certainly *have* thought about these issues before, and given your previous thought I suspect that you are now bubbling over with objections. Of *course* there are uncertainties in the data! Your measuring instruments aren't perfect, and in any case we know about fluctuations in data, so the data must at least have uncertainties. In addition, it is hard to imagine that any real measuring instrument wouldn't be at least somewhat biased. For example, the calibration of an instrument isn't perfectly understood, so if we measure (say) the flux from a source, the real flux will be something different. So surely data have errors as well as uncertainties?

In fact, no. Confusion can easily arise because many of the things that we think of as *measurements* are actually *inferences* from the true, raw data. Take measurement of flux as an example. We never measure flux directly, although it may seem that we do. In reality, we measure the direct response of the detector (voltages through a CCD, or counts in a detector, or something like that, are closer to the direct measurement, although it's still not there) and then use a model of the detector to make our best estimate of the flux. In terms of biases, the fault isn't in the data, it's in the model you have of your detector. The direct measurement gives specific values, with no uncertainties or errors. Put another way, your detector responds somehow to the photons (or whatever) that interact with it. There is systematic error in the way that you model your detector, but that's not the same thing as saying that there are errors in the *data*.

One consequence of this is that most plots you've ever seen in astronomy that have error bars are misleading in a fundamental way because they suggest that the data have errors. In addition, most χ^2 analyses in astronomy are wrong, because they associate statistical uncertainty with the data.

But people who use χ^2 aren't idiots, so what gives? The answer is that there are some circumstances in which the correct analysis can be reasonably well-approximated by associating a "standard error" with the data (shudder; better to call this "standard uncertainty"), and indeed by using Gaussian statistics. But the path we're encouraging in this course, and for any analysis you do, is to make any such choices with eyes open, so that you know what corners you are actually cutting and how that might affect the accuracy and precision of your analysis.

So what's a good way to perform statistical analyses? My experience has made me comfortable with a generalized Bayesian approach. Indeed, more and more areas of astronomical analyses are using Bayesian statistics. Many people, reasonably enough, like the philosophical clarity of Bayesian statistics in contrast to the statistics you might normally encounter (which is called "frequentist" statistics). However, from my standpoint the philosophy isn't nearly as important as the practical results. What has convinced me over the years that Bayesian analysis is a good way to go is that it provides clear, correct, and precise inference in a wide variety of tasks. Thus in this class we'll go over some of the basics of Bayesian statistics.

Bayes' Theorem

Suppose that we have two events, A and B, each with some probability. We consider the probability that both happen: $P(A \text{ and } B)$. This is equal to the probability of B by itself, times the probability that A happens given that B happened:

$$P(A \text{ and } B) = P(A|B)P(B) \; , \tag{1}$$

where $P(A|B)$ is the *conditional* probability that A happens if B happens. Note that A isn't required to depend on B; for example, if A and B are completely independent, then $P(A|B) = P(A)$, but that isn't the general form. We can also write this the other way around:

$$P(A \text{ and } B) = P(B|A)P(A) \; . \tag{2}$$

Therefore $P(A|B)P(B) = P(B|A)P(A)$, which also means that we can write

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \; , \tag{3}$$

as long as $P(B) \neq 0$. This is Bayes' Theorem.

The power of this for statistical analysis comes from replacing A with a particular hypothesis (e.g., that the temperature of a blackbody is 7,312 K) and B with the data you have in hand. Then the factors in this equation may be interpreted as follows:

- $P(A|B)$ is the probability of the hypothesis given the data *and prior information*.
- $P(B|A)$ is the probability that the data would be observed if the hypothesis were true.
- $P(A)$ is the prior probability of the hypothesis being true (in other words, the probability you assigned to the hypothesis before you took the data).
- $P(B)$ can be considered as a normalizing constant, given that probabilities must integrate to unity.

$P(B|A)$ is sometimes called "the likelihood of the data given the model".

Let's now be a little more specific, and then let's go into a particular example. Say that your data come in discrete intervals (which we'll call "counts"), and that the counts are independent of each other. Schematically, we imagine dividing data space up into "bins", which could be bins in energy channel of our detector, location on the sky, time of arrival, or any of a number of other things. Suppose that in a particular model m, you expect there to be m_i counts in bin i. Then if the model is correct the probability of actually observing d_i counts in bin i of the data is, from the Poisson distribution,

$$\mathcal{L}_i = \frac{m_i^{d_i}}{d_i!} e^{-m_i} \ . \tag{4}$$

Note that m_i can be any positive real number, whereas d_i must be a nonnegative integer. Note also that the sum of $\mathcal{L}_i$ from $d_i = 0$ to ∞ is 1 for any m_i, and that the integral of $\mathcal{L}_i$ from $m_i = 0$ to $m_i = \infty$ is 1 for any d_i. The likelihood for the whole data set is the product of the likelihoods for each bin:

$$\mathcal{L} = \prod_i \frac{m_i^{d_i}}{d_i!} e^{-m_i} \ . \tag{5}$$

Thus $P(B|A)$ (from our previous notation) is $\mathcal{L}$. Read as a probability distribution in d_i, as we saw earlier $\mathcal{L}$ becomes better and better approximated by a Gaussian as m_i increases.

If we want to estimate the parameters of a model, only the *ratios* between the likelihoods matter. Because the factor $\prod \frac{1}{d_i!}$ is independent of the model (it depends only on the data), we can factor that out to write

$$\mathcal{L} \propto \prod m_i^{d_i} e^{-m_i} \ . \tag{6}$$

In some circumstances (but not all! Be careful...) we normalize the model so that it has the same total number of counts as the data. If we do that, then because

$$\prod e^{-m_i} = e^{-\sum m_i} \tag{7}$$

this is also a common factor that we can divide out.