Bayesian Statistics: Intro, and Parameter Estimation

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, the way that χ^2 analyses are often used is wrong, because they associate statistical uncertainty with the data (although in a true χ^2 test that's not the case).

But astronomers 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 and then do a simple example that compares Bayesian analysis using a Poisson log likelihood, with a chi squared analysis.

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:

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

Thus $P(B|A)$ (from our previous notation) is $\mathcal{L}$. Read as a probability distribution in d_i, $\mathcal{L}$ 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.

Now let's apply this approach to some data, which will allow us to compare this type of inference with how χ^2 is often (incorrectly) used. I promise we'll get to real astronomical data as soon as possible, but for these initial concepts it will often be useful for us to use synthetic examples.

Say that you flip a coin 10 times and you get 4 heads and 6 tails. Your model is that the probability of heads coming up in a given throw is a, and thus that the probability of tails coming up in a given throw is $1 - a$. Here we will fix the number of throws at 10 (i.e., the actual number!), which means that in our model we would expect $10a$ heads and $10(1 - a)$ tails; note that our two bins are the number of heads and the number of tails. The likelihood of the data given the model (with parameter a) is then

$$\mathcal{L}(a) = \left(\frac{(10a)^4}{4!} e^{-10a} \right) \times \left(\frac{(10(1-a))^6}{6!} e^{-10(1-a)} \right) . \tag{8}$$

We can rewrite this as

$$\mathcal{L}(a) = \frac{10^4}{4!} \frac{10^6}{6!} e^{-10} a^4 (1-a)^6 . \tag{9}$$

Because only the ratio of likelihoods matters in our estimation of a, we can cancel out all of the factors in front to leave

$$\mathcal{L}(a) \propto a^4 (1-a)^6 . \tag{10}$$

Note, though, that the likelihood is only one of the factors that we need to get the posterior probability density $P(A|B)$. We also have to multiply by the prior probability for a. Now, by its nature a has to be somewhere between 0 and 1. More on priors later, but for our current purposes let's say that a has an equal probability of being anywhere between 0 and 1. Then in that special case, the posterior probability density is simply proportional to the likelihood. We will label the posterior probability density $P(a) = \mathcal{L}(a)p(a)$ (where $p(a)$ is the prior probability density for a, which in this case we set to 1 across the whole range $a = 0$ to $a = 1$). Because P is a probability density, when it is properly normalized $\int_0^1 P(a)da = 1$ (in the same way that the prior probability density was normalized, $\int_0^1 p(a)da = 1$).

What can we do with that posterior probability density? As a first step, let's determine where we have our maximum probability (i.e., the mode). We get that by taking the derivative of $\mathcal{L}$ with respect to a and setting it to zero. This gives:

$$\begin{aligned} 4a^3(1-a)^6 - 6a^4(1-a)^5 &= 0 \\ 4(1-a) - 6a &= 0 \\ a &= 0.4 . \end{aligned} \tag{11}$$

That's intuitive; with no other information, our best guess is that the true probability exactly reflects the data.

But we almost always want more than just the best value; we also want to be able to say that, with some probability, a is in a particular range. In Bayesian parlance, we would like to know the "credible region" to some level of probability. To get an idea of what this means, we calculate and plot the normalized posterior probability density as a function of a in the figure. Note that the probability *density* can exceed 1; it is the integral of the probability density that must equal 1.

When we look at the figure we see that the probability density is not symmetric around the peak. For example, at $a = 0.2$ the probability density is about 0.95, whereas at $a = 0.6$ the probability density is about 1.2. This introduces an ambiguity in the definition of the credible region. Should we, for example, start at the peak and move symmetrically to smaller

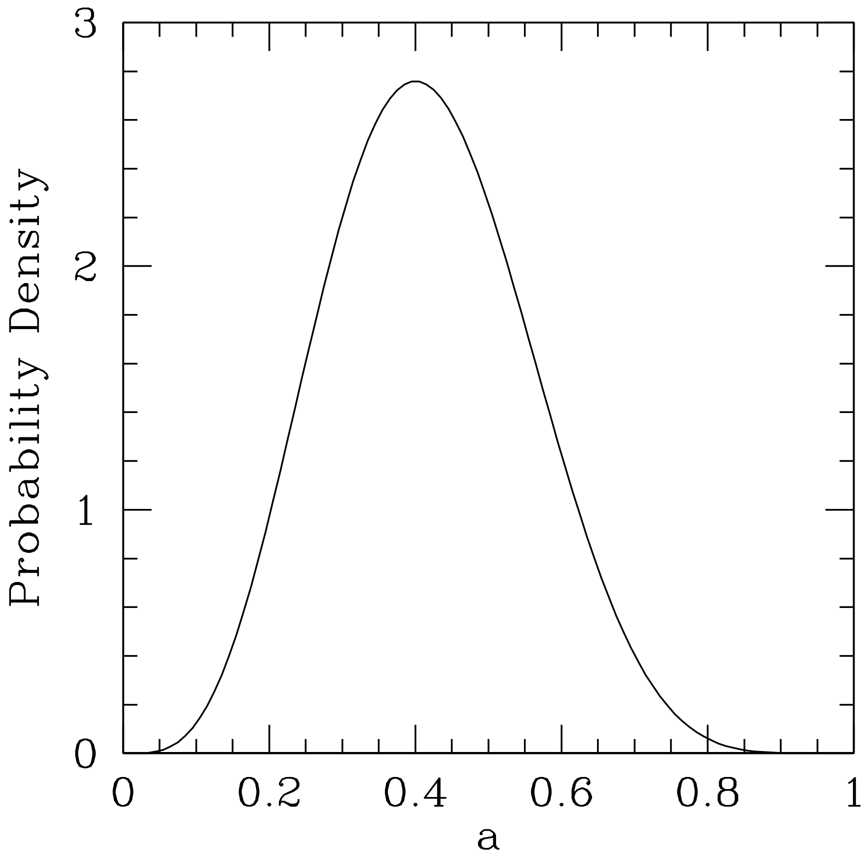

Fig. 1.— Posterior probability density for the probability a of heads after ten flips that produced four heads and six tails. Here our prior was that any value of a from 0 to 1 was equally likely. Note that, as a result, the posterior probability density peaks at $a = 0.4$, and that the probability density is asymmetric around that peak.

and larger values of a until we get to some total probability? Should we begin from $a = 0$ and find the value of a that gives us an integral equal to a specified probability? Should we find the smallest region that contains the specified probability? The smallest *contiguous* region that contains the specified probability?

We'll choose the last of these, for illustrative purposes. Suppose that we want a 68.3% credible region (which we choose because this corresponds to the probability between -1σ and $+1\sigma$ for a Gaussian distribution). Then the minimum-width contiguous range that includes this probability goes from $a = 0.264$ to $a = 0.547$, for a total width of $\Delta a = 0.283$.

What if we were to try to use χ^2 in the incorrect way it is often used, where uncertainty is associated with the data? Now, no one in their right mind would do this when there are only 4 counts in one bin and 6 in the other, but suppose that we blindly did it anyway. The way that most people in astronomy compute chi squared is to sum the ratio of the squared difference between the data and model at each data point, and divide by the variance that we associate with the data (the way it was introduced, you divide instead by the variance that you associate with the *model*, which is closer to the Bayesian approach although it's still very wrong if your model predicts a small number of counts in enough bins). If our data are simply counts, then in the Gaussian limit the variance in a given bin is equal to the number of counts in that bin of the data. Then for a heads fraction of a, the data-variance chi squared for our data is

$$\chi^2 = \sum_i \frac{(m_i - d_i)^2}{\sigma_i^2} = \sum_i \frac{(m_i - d_i)^2}{d_i} = \frac{(10a - 4)^2}{4} + \frac{(10(1-a) - 6)^2}{6} , \qquad (12)$$

which we can expand as $\chi^2 = \frac{5}{3}(5a - 2)^2$. The minimum χ^2, which in this particular case (but not in general) is $\chi^2 = 0$, is again $a = 0.4$. When we look up a chi squared table, we see that for one parameter (a in our case), the 1σ region is determined by looking for regions where the chi squared is 1 greater than the minimum: $\Delta\chi^2 = 1$. Performing this operation faithfully tells us that according to the χ^2 prescription, our 68.3% range should be from $a = 0.245$ to $a = 0.555$, for a total width of $\Delta a = 0.31$. What we see, therefore, is that our log likelihood procedure gets a somewhat tighter region than we get from a blind application of chi squared. The chi squared isn't *too* bad, even in this circumstance, but it doesn't get us the correct probability distribution.

For completeness, let's do this again by performing a chi squared test the way it should be performed: by having the denominator be the *model* variance. The χ^2 assumption is again that the variance is equal to the expected (not observed in this case!) value:

$$\chi^2 = \sum_i \frac{(m_i - d_i)^2}{\sigma_i^2} = \sum_i \frac{(m_i - d_i)^2}{m_i} = \frac{(10a - 4)^2}{10a} + \frac{(10(1-a) - 6)^2}{10(1-a)} . \qquad (13)$$

As you can see, compared with the data-variance version of the χ^2 test (which, again, is commonly used in astronomy!), extreme values of a are penalized much more ($a \to 0$ and $a \to 1$ both cause $\chi^2 \to \infty$). The minimum χ^2 is again 0 at $a = 0.4$. For the correct model-variance χ^2, $\Delta\chi^2 = 1$ gives us a range of $a = 0.2611$ to $a = 0.5571$.

Now it's your turn, using the data sets on the website. Based on the data sets, what are the posterior probability densities for p if we use the Poisson likelihood? What if we use Wilks' Theorem (see the Appendix) with the Poisson log likelihood? How about if we use χ^2? What is the 68.3% credible region using each method? Note that the χ^2 calculation can in this case be performed analytically, but I recommend that you save time and do it numerically. What conclusions do you draw?

In practice, likelihood analyses usually use the natural log of the likelihood rather than the likelihood itself. That's because products of exponentials and powers can often lead to values that are huge or tiny, which makes them difficult to use. Logs are better behaved. In that case, note that it is the *difference* between log likelihoods that we need to use, because that corresponds to the ratio between likelihoods.

Now let's return to the issue of priors. In our analysis above, we assumed that all allowed values of a (i.e., 0 to 1, since a is a probability) are equally likely. In practice this does not have to be the case. You've probably had plenty of experience in flipping coins, and you know that the probability of heads is pretty close to 0.5. Thus although in our example we concluded that $a = 0.4$ is the most probable value, you'd probably need a lot of convincing to conclude that a standard coin has a heads probability of $a = 0.4$. Would getting 40 out of 100 throws do it? Probably not. 400 out of 1000? Maybe you'd be suspicious at that point.

What this tells us is that we actually always have prior probabilities in mind. As a more astronomical example, if you estimate the radial pulsation speed of a star you would not accept an answer that is three times the speed of light. In Bayesian statistics, you have to specify your prior explicitly.

I must say that priors were a sticking point for me when I first encountered Bayesian statistics. The normal statistics you use give the impression that they'll just tell you the answer, with no subjective priors. This, however, is a bit misleading. There are plenty of things you assume as known (speed of light, Planck's constant, etc.) even without realizing it, and those are priors. One can say with some justification that if you try several reasonable priors and these give you wildly different answers, your data didn't contain enough information to judge between them, so you can't say much. It is appropriate to try to select priors that are as uninformative as possible so that the data speak for themselves.

For next time we will continue with parameter estimation, but this time in a more astronomically realistic setting where we have a potentially continuous distribution of data. This is where many astronomers get a nervous tic that compels them to bin data, but as we'll see that isn't necessary!

Appendix: Wilks' Theorem and its Proof

The likelihood is $\mathcal{L} = \prod_i p_i$, where p_i is the probability of the data given the model in bin i. If we are in the limit of Gaussian statistics, then

$$p_i = \frac{1}{\sigma_i\sqrt{2\pi}} e^{-(d_i-m_i)^2/2\sigma_i^2} \tag{14}$$

where $\sigma_i^2 \approx m_i$ for $m_i \gg 1$. Thus

$$\ln \mathcal{L} = -\sum_i \frac{(d_i - m_i)^2}{2\sigma_i^2} + \sum_i \ln\left(\frac{1}{\sigma_i\sqrt{2\pi}}\right) = -\chi^2/2 + \text{const} . \tag{15}$$

Note that for a decent fit, $d_i \approx m_i$, which means that you could switch d_i for m_i in the variance. Therefore $2\Delta \ln \mathcal{L} = -\Delta\chi^2$ in the Gaussian limit. For example, if you have one parameter and you are interested in the 1σ range, a look at a χ^2 table tells you that $\Delta\chi^2 = 1$ in that case. Thus to apply Wilks' Theorem to your log likelihood computation you find the maximum log likelihood and then determine the range of the parameter that is within $\Delta \ln \mathcal{L} = -0.5$ of the maximum.

One of the points of the computational exercise suggested for this class is for you to get an idea of how well this approximation does in specific cases. Enjoy!