Statistics in Astronomy

Initial question: How do you maximize the information you get from your data?

Statistics is an art as well as a science, so it's important to use it as you would any other tool: you have to know what you want to get out of your statistics before you can choose the right way to apply them. Therefore, let's start with some philosophy.

First, statistics are not magic! If the data aren't good enough, no amount of statistical manipulation will bring out the results. If your data are exceptional, the main results are often evident even without lots of massaging. Nonetheless, there are plenty of cases in the middle (where only good statistics will work), and it's appropriate to be rigorous when possible anyway.

Second, in astronomy systematics are often a major issue. An important principle is that hypotheses are tested in bundles. When you make an observation, you aren't measuring the physical quantities of the system directly. You make implicit assumptions about your knowledge of the instrument, and often need to make explicit assumptions about a basic model, to get the quantities that really interest you. Often there are enormous uncertainties as a result. This means that the statistical significance you quote may not be representative. For example, suppose I claim an astronomical result at 3 sigma significance. If I know my instrument and ancillary assumptions perfectly, this will only happen by chance 3 times in 1000, so that sounds good. But long and painful experience shows that systematics can eat that up in a hurry, which is why people often adopt a much higher threshold for a significant result. Example: when the third BATSE catalog of gamma-ray burst positions was released, the team had revised its position-finding algorithm. They said that it corresponded well to the previous positions, because only 4% of the burst positions moved by more than 5σ. Now, 5σ is a 3×10^{-7} result, so obviously there were some significant systematics there!

Third, when quoting significance you need to be careful about the number of trials you've performed, and in particular you need to be careful about hidden trials. Suppose I'm looking for evidence of ESP by having subjects guess patterns on cards. At the end, I'm ecstatic because there are more hits than expected, at the 95% significance level. Even if my experiment is designed perfectly (and these usually aren't), I have to realize that there may well be 20 other labs doing the same thing, and only the positive results are quoted.

So what's a good approach? One thing you do *not* want to do is just use statistics blindly, any more than you would obliviously use the Saha equation for the center of our Sun (where it doesn't apply). An approach I find helpful is to try to think of how a certain statistical task would be done correctly (if only I had time and all the information I needed), then determine a reasonable approximation for my specific case. My upbringing as a postdoc, combined with a reasonable amount of experience, has made me comfortable with a generalized Bayesian

approach. Some of the characteristics of this approach are forward folding, the use of Poisson likelihood, and a clear distinction between parameter estimation and model comparison. Let's look at each of these in turn, then apply them.

The idea of forward folding is that we should try to evaluate models based on direct comparison with the data. You may ask: isn't this what we always do? Not necessarily. Say you want to estimate the temperature of a molecular cloud based on some observations. How do you do it? The way you might think of is, say, to get line ratios from your observations, then use an equilibrium model to get the temperature from the line ratios. This is backwards folding: take the data, then (figuratively) propagate it backwards to the molecular cloud. Forward folding would be having a model of the cloud including temperatures, densities, and so on, then a model of what emission this would produce, then a model of how that light would be received by your detector. You would then compare your "model data" with the real data. I must say that in practice the two methods are often nearly the same, mainly because the detector is well-understood. However, for a given case you should at least give some thought to whether some aspect of your detector is uncertain enough that backward folding could be a problem.

The Poisson likelihood can be used any time your data come in discrete intervals (which we'll call "counts"), and 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 likelihood of actually observing d_i counts in bin i of the data is, from the Poisson distribution,

$$\mathcal{L}_i = \frac{m_i^{d_i}}{d_i!} e^{-m_i} \ . \tag{1}$$

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

$$\mathcal{L} = \prod \frac{m_i^{d_i}}{d_i!} e^{-m_i} \ . \tag{2}$$

This becomes better and better approximated by a Gaussian as m_i increases.

It's interesting to talk to some people and ask why they did certain things with their data. For example, maybe they've taken a spectrum and put it into coarser bins. When asked why, they may respond "I put it into coarser bins to make the bin-by-bin statistics better". That is, if originally there were few counts per bin, they rebin so there are lots of counts per bin and thus they can use Gaussian statistics. Here's a secret: when you make your bins coarser, you *lose* information, always! This has to be true, since by rebinning in this way you no longer know where in the bin each count came from. If one uses Poisson

likelihoods, small numbers are fine. In fact, if you can manage, the best way to represent your data is to have bins so tiny that you expect either 0 or 1 count per bin. Then, you're getting maximum information. There are times when your data don't come in this count-by-count way. For example, when observing at Keck you don't count individual photons. In that case, each bin of the data itself may have error bars associated with it. Then one can replace the Poisson likelihood with a standard Gaussian; this is because, effectively, there are many counts per bin.

What about parameter estimation and model comparison? In Bayesian statistics, there is a clear distinction drawn between these. For the first, you *assume* a particular model, which has some number of parameters. Given a data set, you want to know the best value of the parameters for that model, and a credible region around the best values (in Bayesian statistics they call it a credible region and not a confidence region). For model comparison, you need to precisely specify two models (there is no null hypothesis in Bayesian statistics), and see how each of these does against the data.

Let's be a little more concrete. Suppose we have a model with a single parameter a. We'd like to estimate the best value of a plus its uncertainties, from a data set. The key to the forward folding aspect of Bayesian statistics is that you compute the likelihood of the *data* given the *model*, and not the other way around. That is, for each value of a you can compute the likelihood that the data would come out as it does. So, for n data bins,

$$\mathcal{L} = \prod_{i=1}^{n} \frac{m_i(a)^{d_i}}{d_i!} e^{-m_i(a)} \ , \tag{3}$$

where $m_i(a)$ is the model number in bin i if the parameter value is a. Now, to find credible regions it turns out that the absolute value of $\mathcal{L}$ is unimportant; what matters is likelihood *ratios*. Looking at this equation, you can see that this allows simplifications. First, the product of $d_i!$ in the denominator is the same for all models, since it depends just on the data. This can be cancelled out. Second, the model is usually (but not always!) normalized so that the total number of model counts equals the total number of data counts. That means that $\sum m_i(a)$ is the same for all a. But the product of exponentials is just $\exp(\sum m_i(a))$, so this, too, can be cancelled out. One is left with

$$\mathcal{L} \propto \prod_{i=1}^{n} m_i(a)^{d_i} \ . \tag{4}$$

This product can be huge or tiny, depending on $m_i(a)$, so it is common to work with the log likelihood instead:

$$\ln \mathcal{L} = \sum_{i=1}^{n} d_i \ln m_i(a) + \text{const.} \tag{5}$$

Now it's the difference in log likelihoods that matters. Fortunately, if the total number of counts (*not* the counts per bin) is reasonably large, say $10 - 20$ or more, the distribution is

very similar to a chi squared distribution, and $2\Delta \ln \mathcal{L} \approx -\Delta\chi^2$, so you can use chi squared tables. The procedure is then that one finds the maximum in $\ln \mathcal{L}$, which gives your best model parameter value(s). One then computes $\ln \mathcal{L}$ for other parameter values to find the credible region, based on the number of parameters. Notice, by the way, that if $d_i = 0$ for some bin, that bin makes no direct contribution to the log likelihood. One can therefore confine attention to the bins where data are present; this results in a significant savings of time!

Proof of Wilks' theorem. 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

$$p_i = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-(d_i - m_i)^2/2\sigma_i^2} \tag{6}$$

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

$$\ln \mathcal{L} = -\sum_i \frac{(d_i - m_i)^2}{2\sigma_i^2} + \sum_i \ln\left(\frac{1}{\sigma_i\sqrt{2\pi}}\right) = -\chi^2/2 + \text{const} . \tag{7}$$

Therefore $2\Delta \ln \mathcal{L} = \Delta\chi^2$ in the Gaussian limit.

There's one aspect of this that we haven't addressed. In adopting the above procedure, we have implicitly assumed that, a priori, all values of the parameter a are equally likely. In practice this is not the case. For example, suppose you are estimating the radial pulsation speed of a star. You would not accept an answer that is three times the speed of light. Such restrictions on parameter values are called *priors*. Therefore, what you do in reality is (1) State your prior probability distribution $P(a)$, such that $P(a)$ integrates to unity over all possible values of a, (2) Compute the likelihood for each value of a given the data, (3) Multiply the likelihood by the prior probability distribution, then normalize, to get the final posterior probability distribution.

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.

The preceding, however, is too abstract. Let's choose an easy example. Say you flip a coin 20 times and you get 8 heads and 12 tails. Your model is one of the probability of heads coming up, which we'll call a. The two bins are the number of heads and the number

of tails. The likelihood of the data given the model a is then

$$\mathcal{L} = \left(\frac{(20a)^8}{8!}e^{-20a}\right) \times \left(\frac{(20(1-a))^{12}}{12!}e^{-20(1-a)}\right) \ . \tag{8}$$

Dividing out common factors and taking the log,

$$\ln \mathcal{L}(a) = 8\ln(20a) + 12\ln[20(1-a)] \ . \tag{9}$$

We ignore the additive constant. Note, in fact, that we get $8\ln 20 + 12\ln 20 + 8\ln a + 12\ln(1-a)$, so we can ignore the $20\ln 20$ term as well. Maximizing, we find $a = 0.4$ gives the largest value. That's intuitive; with no other information, our best guess is that the true probability exactly reflects the data. For a single parameter, $\Delta\chi^2 = 1$ (or $\Delta \ln \mathcal{L} \approx 0.5$) gives our 1σ credible region of 0.3 to 0.51. Note, by the way, that if you are sure based on your knowledge of coins that a is really close to 0.5, this affects your prior and therefore the posterior probability distribution.

Let's try another, more complicated example. In the preceding we had a finite number of bins (two, in fact). In many circumstances we instead have to deal with a continuous variable, which therefore has a potentially unlimited number of bins. We can estimate parameters in the same way as before, but it's useful to give it a try because there are some apparent difficulties to overcome.

Suppose that we have measured radial velocities, and we fit a zero-centered Gaussian to the data. The single parameter of interest is the standard deviation. We have ten data points, which are velocities of 2.78944, -1.84623, 1.80029, 0.110899, -0.926109, -0.909967, 0.296846, 3.57825, -1.76493, and 2.55558 in units of km s^{-1}. How do we estimate the standard deviation? It might seem that you would *have* to bin the measurements, because otherwise it doesn't look like a distribution at all (if the bins are narrow, one only has zero or one points in a bin, so there are no peaks). This is, however, not the case, so let's see how it works.

First, we realize that our Gaussian has the form

$$N(v)dv = A\exp(-v^2/2\sigma^2)dv \ , \tag{10}$$

where σ is the standard deviation and A is a normalization factor. Therefore, it might appear that there are two parameters. However, it is usual to assume that the distribution is normalized so that the total number equals the number of data points, so in our case $\int_{-\infty}^{\infty} N(v)\,dv = 10$. We also realize that if v has units, our normalization constant must as well. This is unimportant for our purposes, so we'll adopt the convention that v is measured in units of 1 km s^{-1}. Then the normalization convention implies $A = 10/(\sigma\sqrt{2\pi})$.

Now we construct the log likelihood. As we showed earlier, the only important term is $\sum d_i \ln(m_i)$, where the sum is over all bins, d_i is the number of counts in bin i, and m_i is

the predicted number of counts in bin i from the model. If we imagine dividing the data space into an enormous number of narrow bins, though, we realize that the ones without counts don't contribute, because $d_i = 0$. Therefore, the sum really only needs to go over the bins that contain counts. Next, what is m_i? It is the expected number of counts in a bin. Suppose that a bin has width dv at velocity v. Then the expected number is $N(v)dv$. This appears to depend crucially on the bin width, but remember that we're just comparing *differences* of log likelihoods. Therefore, if we use the same bin widths for every value of σ (which we obviously will), the $\ln dv$ values will be in common between all models, and hence will cancel. If we make the further assumption that we've done the smart thing and chosen small enough bins that the ones with data all have $d_i = 1$, then we get finally

$$\ln \mathcal{L} = \sum_i \ln[N(v_i)] + \text{const} \ , \tag{11}$$

where the v_i are the measured velocities. The depends only on the values of the distribution function at the measured velocities.

This is actually a general result for continuous distributions, in any number of dimensions. After you've normalized, the log likelihood is just the sum of the log of the distribution function at the measured locations.

For a general model, one would now calculate the log likelihood numerically for a set of parameter values, then maximize to get the best fit. In our particular case, we can do it analytically. Dropping the constant,

$$\ln \mathcal{L} = \sum_i \left[\ln(10/\sqrt{2\pi}) - \ln \sigma - v_i^2/2\sigma^2 \right] \ . \tag{12}$$

This sum is over the ten measured velocities. We note that the first term is in common between all models, so we drop it. We then have

$$\ln \mathcal{L} = -10 \ln \sigma - (1/2\sigma^2) \sum_i v_i^2 \ . \tag{13}$$

The sum of the squares of our velocities is 38.7. Taking the derivative with respect to σ and setting to zero (to maximize) gives

$$\begin{aligned} -10/\sigma_{\rm best} + 38.7/\sigma_{\rm best}^3 &= 0 \\ \sigma_{\rm best} = (3.87)^{1/2} &= 1.97 \ . \end{aligned} \tag{14}$$

The 68% confidence interval is computed using $\Delta \ln \mathcal{L} = 0.5$, and runs from $\sigma = 1.60$ to $\sigma = 2.50$. In fact, $\sigma = 2.5$ was used to generate the data.

If there are several parameters, one can do similar things but often one is only interested in the distribution of a subset of them (say, only one of them!). In that case, one *marginalizes* the posterior probability distribution. That is, suppose now we have lots of parameters

$a_1,\ a_2,\ \ldots, a_n$. Our posterior probability distribution is $P(a_1,\ a_2,\ \ldots, a_n)$, normalized so that $\int P(a_1,\ a_2,\ \ldots, a_n) da_1 da_2 \ldots da_n = 1$. If we only want to know the probability distribution for parameter a_1, independent of the values of the other parameters, we simply integrate over those other parameters:

$$p(a_1) = \int P(a_1,\ a_2,\ \ldots, a_n) da_2 \ldots da_n \ . \tag{15}$$

We then have $\int p(a_1) da_1 = 1$. Similarly, one could find the distribution for the two parameters a_1 and a_2 by integrating P over a_3 through a_n. The parameters you integrate over are called *nuisance* parameters.

One cautionary point: since the *value* of the likelihood never enters, one can happily calculate maximum likelihoods and credible regions for models that are awful! It's an automatic procedure. That's why Bayesians draw a distinction between parameter estimation and model comparison, which we will now treat.

Suppose we have a data set, and two models to compare. How do we determine which model is favored by the data? At first glance this may seem easy: just figure out which model matches the data better. But think about models with different numbers of parameters; intuitively, we should give the benefit of the doubt to the model with fewer parameters, based on Ockham's principle. In addition, one could imagine a situation in which the parameters of two models are qualitatively different. For example, some of the parameters could be continuous (e.g., temperature), and some could be discrete (e.g., the quantum spin of a particle). How are these to be taken into account?

This, in my opinion, is where Bayesian statistics shines. It provides a simple procedure that *automatically* takes into account different numbers of parameters in an intuitively satisfying way. As before we'll give the general principles, then try some examples.

Say we have two models, 1 and 2. Model 1 has parameters $a_1,\ a_2, \ldots,\ a_n$, and a prior probability distribution $P_1(a_1,\ a_2, \ldots,\ a_n)$. Model 2 has parameters $b_1,\ b_2, \ldots,\ b_m$ and a prior probability distribution $P_2(b_1,\ b_2, \ldots,\ b_m)$. For a given set of values $a_1,\ a_2, \ldots,\ a_n$, let the likelihood of the data given the model (defined above) for model 1 be $\mathcal{L}_1(a_1,\ a_2, \ldots,\ a_n)$, and similarly for model 2. Then the "Bayes factor" for model 1 in favor of model 2 is

$$\mathcal{B}_{12} = \frac{\int \mathcal{L}_1(a_1,\ a_2, \ldots,\ a_n) P_1(a_1,\ a_2, \ldots,\ a_n) da_1\, da_2 \ldots da_n}{\int \mathcal{L}_2(b_1,\ b_2, \ldots,\ b_m) P_2(b_1,\ b_2, \ldots,\ b_m) db_1\, db_2 \ldots db_m} \tag{16}$$

where the integration in each case is over the entire model parameter space. Therefore, it's just a ratio of the integrals of the likelihoods times the priors for each model. When you multiply the Bayes factor by the prior probability ratio you had for model 1 in favor of model 2 (which you could set to unity if you had no reason to prefer one model over another), you get the odds ratio $\mathcal{O}_{12}$ of model 1 in favor of model 2.

What does this mean? Don't tell a real Bayesian I explained it this way, but consider the following. Suppose you and a friend place a series of bets. In each bet, one has two possible models. You compute the odds ratio as above, and get $\mathcal{O}_{12}$ in each case. Ultimately, it will be determined (by future data, say) which of the two models is correct (we're assuming these are the only two possible models). If your friend puts down \$1 on model 2 in each case, how much money should you place on model 1 in each bet so that you expect to break even after many bets? You put down $\$\mathcal{O}_{12}$. That is, it really does act like an odds ratio. The reason a hard-core Bayesian might get agitated about this analogy is that Bayesian statistics emphasizes considering only the data you have before you, rather than imagining an infinite space of data (as happens in more familiar frequentist statistics). Still, I think this is a good description.

Why does this automatically take simplicity into account? Think of it like this. If your data are informative, then for a given set of data it is likely that only a small portion of the parameter space will give a reasonably large likelihood. For example, if you are modeling the interstellar medium in some region, you might have temperature and density as parameters; with good enough data, only temperatures and densities close to the right ones will give significant $\mathcal{L}$. Now, think about the priors. For a complicated model with many parameters, the probability density is "spread out" over the many dimensions of parameter space. Thus, the probability density is comparatively small in the region where the likelihood is significant. If instead you have few parameters, the prior probability density is less spread out, so it's larger where the likelihood is significant and therefore the integral is larger.

If the parameters have discrete instead of continuous values, you do a sum instead of an integral but otherwise it's the same. Note that we have to use the full Poisson likelihood here. When we did parameter estimation we could cancel out lots of things, but here we have an integral or sum of likelihoods so we can't do the cancellation as easily. The product $\prod(1/d_i!)$ will be the same for every likelihood, and if your model is normalized so that the total number of expected counts is set to the number of observed counts (which is common, but not always true) then $\prod \exp(-m_i)$ is the same for every likelihood. Thus those factors can be cancelled, but one still has a sum of likelihoods and so taking the log doesn't help.

Let's try an example. Consider a six-sided die. We want to know the probabilities of each of the six faces. Model 1 is that the probability is the same (1/6) for each face. Model 2 is that the probability is proportional to the number on the face. Normalized, this means a probability of 1/21 for 1; 2/21 for 2; and so on. We roll the die ten times and get 5, 2, 6, 2, 2, 3, 4, 3, 1, 4. What is the odds ratio for the two models?

We're starting with an easy one, in which there are no parameters, so we don't even have to do an integral, just a likelihood ratio. For model 1 the normalized model expectations per bin are $m_1 = 10/6$, $m_2 = 10/6$, and so on. For model 2 we have $n_1 = 10/21$, $n_2 = 20/21$,

$n_3 = 30/21$, and so on. Therefore,

$$\mathcal{L}_1 = \left(\frac{10}{6}\right)^1 \cdot \left(\frac{10}{6}\right)^3 \cdot \left(\frac{10}{6}\right)^2 \cdot \left(\frac{10}{6}\right)^2 \cdot \left(\frac{10}{6}\right)^1 \cdot \left(\frac{10}{6}\right)^1 = 165.4 \tag{17}$$

and

$$\mathcal{L}_2 = \left(\frac{10}{21}\right)^1 \cdot \left(\frac{20}{21}\right)^3 \cdot \left(\frac{30}{21}\right)^2 \cdot \left(\frac{40}{21}\right)^2 \cdot \left(\frac{50}{21}\right)^1 \cdot \left(\frac{60}{21}\right)^1 = 20.7 \; . \tag{18}$$

Thus, from this data,

$$\mathcal{O}_{12} = \mathcal{L}_1/\mathcal{L}_2 = 7.98 \; . \tag{19}$$

Model 1 is strongly favored.

Now try another example, with the same data. Model 1 is the same as before, but now model 2 has a parameter. In model 2, the probability of a 1 is $1 - p$, and the probability of a 2, 3, 4, 5, or 6 is $p/5$. Therefore, model 2 encompasses model 1, so by maximum likelihood alone it will do better. But will it do enough better to be favored? Let's assume as a prior that p is equally probable from 0 through 1. The numerator is the same as before, but for the denominator we need to do an integral. For probability p and our given data, the Poisson likelihood of the data given the model is

$$\mathcal{L}_2(p) = [10(1-p)] \cdot (2p)^3 \cdot (2p)^2 \ldots = 10(1-p)(2p)^9 \; . \tag{20}$$

Therefore the denominator is

$$\int_0^1 5120(1-p)p^9 \, dp = 46.5 \tag{21}$$

and the odds ratio is

$$\mathcal{O}_{12} = 165.4/46.5 = 3.55 \; , \tag{22}$$

so the first model is still preferred. Note that the maximum likelihood for model 2 occurs for $p = 0.9$ and gives 198.4, so as expected the more complicated model has a higher *maximum* likelihood; it's just not enough to make up for the extra complication.

Model comparison in Bayesian statistics is always between two precisely defined models. There is no analogue to the idea of a null hypothesis. Hard-core Bayesians consider this to be a strength of the approach. For example, suppose that you try to define a null hypothesis and do a standard frequentist analysis, finding that the null hypothesis can be rejected at the 99% confidence level. Should you, in fact, reject the null hypothesis? Not necessarily, according to Bayesians. Unless you know the full space of possible hypotheses, it could be that there are 10,000 competing hypotheses and of those your null hypothesis did the best. For example, suppose I think that gamma-ray bursts should come from isotropically distributed positions in the sky; that's my null hypothesis. A hundred positions are measured, and they are all found to cluster within 1° of each other. Surely I can reject my null hypothesis? Well, if I

compare it with another hypothesis that says that all bursts should come from within 1" of each other, my null hypothesis does much better!

I'm not happy with this line of argument. To me, the testing of a null hypothesis as it's done in frequentist statistics is important because it gives you a way to tell if your model is reasonably close or not. That is, a standard chi squared per degree of freedom can give you an idea of whether you need to work a lot harder to get a good model, or if you're nearly there. In my opinion, it's important to have that kind of information, but there is reasoned disagreement on this issue.

Where does all of this leave us? In this class I've emphasized over and over that when evaluating a model or derivation or whatever you should use quick, easy methods first (e.g., order of magnitude estimation) before settling in for more detailed treatments. The same goes for statistics. I recommend doing a quick analysis (chi squared, Kolmogorov-Smirnov test, or whatever) first, to see if your data are informative. If they are, then you may be justified in spending time with a more rigorous method to get the most out of your data. In all cases, however, you have to know the limitations of your method! If your model is terrible, getting detailed confidence or credible regions around your best fit isn't meaningful. If you have 40 degrees of freedom (defined as the number of data points minus the number of parameters in your model) and your total chi squared is 4000, you can't say much. On the other hand, if you get a reduced chi squared much *less* than one, doing delta chi squareds is also not really meaningful; actually, what it means is that you overestimated your error bars. So, the lesson as always is that you need to understand your method. That's why I've found it helpful to think in the Bayesian way. In many circumstances it means I can figure out what *should* be done, then I have a better sense of how good an approximation a simpler method is.

If you want a much more thorough discussion of Bayesian methods, I recommend highly Tom Loredo's Bayesian page:

http://astro.cornell.edu/staff/loredo/bayes/