Nested Sampling Journal Talk

Richard Udall

December 2021

1 What is a Bayesian Inference Problem?

1.1 Models and Uncertainty

- The laws of physics give us models of physical systems
- Experimentally, we want to verify those models, or use them to understand the underlying (unknown) parameters of the system
- If our experiment had 0 uncertainty, this would be easy, but all true experiments will have some level of noise

1.2 Frequentism

- Suppose we have a model $\mathcal{M}$ of our system, which has parameters θ and that there is some noise which we may also model as a random variable. Then our data collected will be

$$d = s(\theta, \mathcal{M}) + N \tag{1}$$

  where $s(\theta, \mathcal{M})$ is the signal and N is the noise
- Given this, we may ask the question, given this model and parameters, what the probability of the data we gather is. If we assume our model to be perfect[1], this is simply the probability of having a noise realization which shifts the predicted signal to the resulting data. We can write this as $p(d|\theta, \mathcal{M})$, that is the probability of getting the data given this set of parameters and our model. We will generally make simplifying assumptions about the noise (for example, in LIGO we assume all of our noise is stationary and Gaussian) that will make this quantity (relatively) easy to calculate.

[1] We can also account for an imperfect model by adding some parameterized systematic offset, say $C(\lambda)$ which we then model in conjunction with our physical model, though this risks allowing over-fit of the data

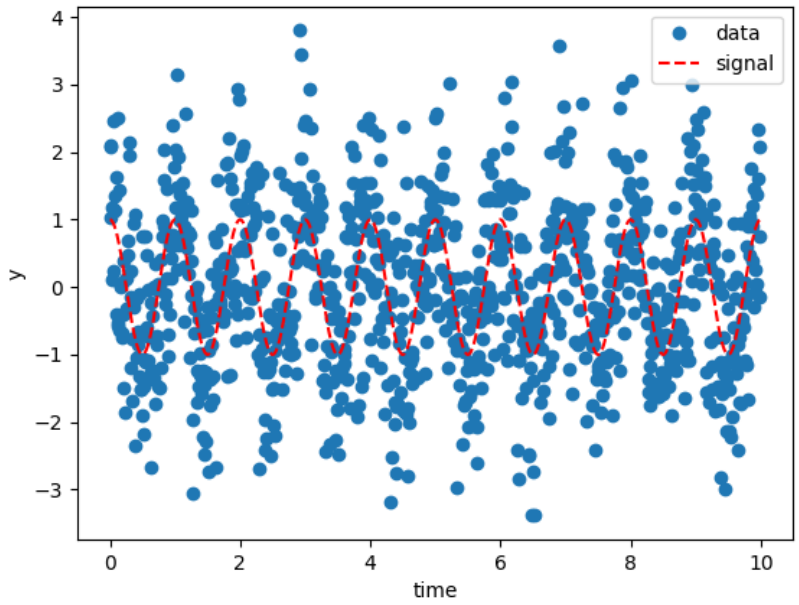

Figure 1: An Example of Noisy Data Given an Underlying Model

- From here, in frequentism we will usually ask what the probability of this is versus the probability of the null hypothesis, that is an assumption that the model is not correct (or not applicable). For example, if one is looking for a noisy sine wave, it is very improbable that the noise would *happen* to take on exactly the shape of a sine wave with your expected frequency and amplitude, and so you may compute just how improbable this is.
- Example: Dice - if you are rolling a six sided die, you may ask whether it is fair or not. The frequentist approach would be to roll it many times, and evaluate the probability of the resulting distribution given an assumption of a fair die. If this probability is sufficiently low, we would conclude that the die is not fair. But if it isn't fair, how do we determine what the actual probabilities would be? If we can roll forever, it's straightforward: given enough rolls, we will eventually converge to the exact result, especially since this is a discrete distribution. In nature, however, we can't roll forever, so we need something better

1.3 Bayesian Statistics

- The most important distinction in Bayesian statistics, which must be made at the beginning, is that Bayesian probabilities *are not the same*

as frequentist probabilities. They constitute statements of confidence in a conclusion, and cannot be directly mapped to the concept of repeated experiments. However, they often map more effectively to the way we treat probability in real life. The weather, for example, is essentially deterministic on short enough time scales: if there is an 80% chance of rain tomorrow, that is not a statement that the true atmospheric configuration at this moment produces rain in 24 hours 80% of the time, but rather that our uncertainty in knowing about it means we are only 80% confident in our conclusion. In physics, these are often the types of problems we will be more interested in, especially in astrophysical contexts.

- The key point of Bayesian statistics is Bayes' Theorem. In its simplest form, it is a statement about true joint probabilities. The probability of two events may be written

$$p(a, b) = p(a|b)p(b) = p(b|a)p(a) \tag{2}$$

That is, the probability of a and b happening together is the probability of a given b times the probability of b, and also the probability b given a times the probability of a. Then rearranging

$$p(b|a) = \frac{p(a|b)p(b)}{p(a)} \tag{3}$$

This is all well and good, but we can also write it for our probability of the data from before:

$$p(\theta|d, \mathcal{M}) = \frac{p(d|\theta, \mathcal{M})p(\theta|\mathcal{M})}{\pi(d|\mathcal{M})} \tag{4}$$

So, our posterior distribution for the parameters given the model and the data (that is, the confidence distribution after including the data) is the probability of the data given the parameters and the model, times the probability of the parameters given the model (the prior. denoted by π - we'll get to that), divided by the probability of producing the data given any configuration under the model. This probability is simply the integral over the parameter space, so if you perform this integral you will find that (since probability is normalized)

$$\mathcal{Z} = \int p(d|\theta, \mathcal{M})\pi(\theta|\mathcal{M})d\theta \tag{5}$$

- Priors are a sticky topic in Bayesian inference, because they necessarily require some uninformed decision about what is or is not reasonable or likely. Best practice is to usually choose something uninformative, such as "all configurations are equally likely" or "vectors on a sphere will have uniformly distributed magnitude and isotropic orientation." They may also be useful for other things though, such as testing underlying models.

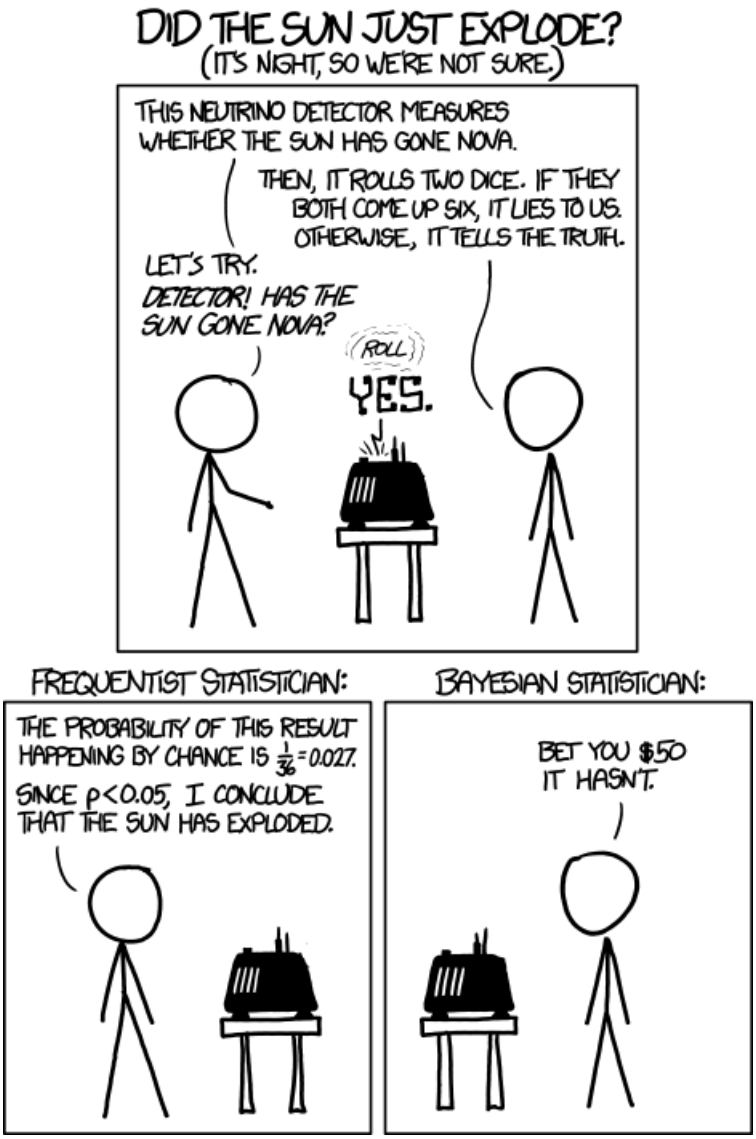

Figure 2: Priors are Good, Actually

Importantly, though, priors are again something which correspond much better to the way we regularly deal with probability - we already have assumptions about what and is not reasonable. It is important not to rely on our priors to strongly - confirmation bias is the end of that road - but they also help guide us where we should be looking.

1.4 Likelihood

- An important preliminary is to note that in (4) above the probablities are, by definition, normalized. That's not really something we can enforce a priori on our analysis methods, so instead we'll introduce the likelihood $\mathcal{L}(\theta, \mathcal{M})$ which is proportional to $p(d|\theta, \mathcal{M})$. Similarly, we will have the evidence $\mathcal{Z}(\mathcal{M})$, which is proportional to $p(d|\mathcal{M})$ by the same factor. We'll also frequently talk about $\ln\mathcal{L}$ and $\ln\mathcal{Z}$, the natural logarithms of these quantities, since these are much more tractable in computational implementations. These get put together into the formula we'll be using:

$$p(\theta|d, \mathcal{M}) = \frac{\mathcal{L}(\theta|\mathcal{M})\pi(\theta|\mathcal{M})}{\mathcal{Z}(\mathcal{M})} \tag{6}$$

2 Nested Sampling

2.1 Sampling

- The task now set is to evaluate (6) over the entirety of parameter space (that is, all possible values of θ. In most cases, there will be no clean analytic solution, and so instead we will perform Bayesian sampling. Essentially, we will seek to find some approximate of the left side of the equation, and simultaneously will be performing the integral over the numerator to determine the evidence. There are many methods for doing this, but as the title indicates I'll be focusing on Nested Sampling

2.2 Basics

- Nested sampling begins with the following reformulation of (5):

$$\mathcal{Z} = \int \mathcal{L} dX \tag{7}$$

X here is the prior volume - $dX = \pi(\theta)d\theta$. Since the prior is a normalized distribution, we have $\int dX = 1$. Now, we can arrange the parts of the prior volume however we want for this integral, so lets arrange them by increasing order:

$$X(\lambda) = \int_{\mathcal{L}(\theta)>\lambda} \pi(\theta)d\theta \tag{8}$$

This formulation has the interesting property that it is monotonically decreasing: if $\lambda = -\infty$, then it is equal to 1, and if $\lambda > \max \mathcal{L}$ then it is equal to 0. This allows us to put meaningful integration bounds on (7)

$$\mathcal{Z} = \int_0^1 \mathcal{L}(X) dX \tag{9}$$

where I've suppressed the chain rule on lambda.

- An example (from Skilling) is the following:

$$L = \begin{array}{|r|r|r|r|} \hline 0 & 8 & 15 & 3 \\ \hline 11 & 24 & 22 & 10 \\ \hline 19 & 30 & 26 & 16 \\ \hline 9 & 23 & 18 & 6 \\ \hline \end{array}$$

Figure 3: A Multidimensional Likelihood Distribution

$$Z = \frac{30}{16} + \frac{26}{16} + \frac{24}{16} + \frac{23}{16} + \frac{22}{16} + \frac{19}{16} + \frac{18}{16} + \frac{16}{16} + \frac{15}{16} + \frac{11}{16} + \frac{10}{16} + \frac{9}{16} + \frac{8}{16} + \frac{6}{16} + \frac{3}{16} + \frac{0}{16} = 15$$

Figure 4: The Corresponding Evidence Computation

- To compute a real integral, we can imagine this ordering:

$$0 < X_m < X_{m-1} < ... < X_2 < X_1 < 1 \tag{10}$$

Where increasing m corresponds to increasing likelihood per (8). Then we can estimate:

$$\mathcal{Z} \approx \sum_i^m w_i \mathcal{L}_i \tag{11}$$

Where w_i are some associated prior weights $X_i - X_{i+1}$. Since X is monotonically decreasing, left and right Riemann sums will bound the true value from above and below respectively, and one may of course also apply higher order e.g. trapezoidal sums.

- Another important preliminary is the concept of information. Information of the posterior against the prior is defined by

$$H = \int \log\left(\frac{dP}{dX}\right) dX \tag{12}$$

Roughly, the volume of the true posterior will be a factor e^{-H} less than that of the prior. For high information (and it's usually high), this means the posterior occupies only a vanishingly small part of prior space. Thus, it's important that we sample in log prior space, rather than prior space itself. This may be achieved by attempting to decrease the prior space by some roughly constant amount each iteration, i.e. $X_m = t_m X_{m-1}$ where $t_m < 1$

2.3 The Algorithm

- We want to get this ordered set of points, with the prior volume always decreasing. Because of how we ordered our X's, we may do this by selecting some point of higher likelihood, $\mathcal{L}_{m+1} > \mathcal{L}_m$, and indeed doing so allows us to skip the process of sorting entirely. Making an iid sample from the prior space which satisfies this constraint will be equivalent to drawing a new X with $X_i = t_i X_{i-1}$, so we will be sampling in log X as desired.

- To make this work, we will have N "live" points, which we will order by likelihood. Then they will satisfy the recurrence relation

$$X_0 = 1,\ X_i = t_i X_{i-1}, \Pr(t_i) = N t_i^{N-1} \tag{13}$$

with t_i being the largest of N draws from Uniform$(0, 1)$ a bit of calculus gives

$$E(\ln(t_i)) = -\frac{1}{N} \tag{14}$$

That is, $\ln(X_{i+1}) - \ln(X_i) \approx -\frac{1}{N}$, giving us a way to approximate the weights required for our evidence sum.

- The algorithm itself:

  1. Draw N live points
  2. Order by likelihood
  3. Initialize $\mathcal{Z} = 0$, $X_0 = 1$
  4. While Converging $(i = 1, i{+}{+})$:
     (a) Choose worst of the live points
     (b) assign it weight:

$$\ln(w_i) = \ln(X_i) + \ln\mathcal{L}_i \tag{15}$$

where

$$\ln(X_i) \approx -\frac{i}{N} \tag{16}$$

and put it into the collection of "dead" points
     (c) increment the evidence by this weight

(d) Draw a new iid sample from the prior, under the requirement $\mathcal{L} > \mathcal{L}_i$,

5. Once convergence ends, assign weights using prior mass $-i_{max}/N$ to remaining live points, and add these to the evidence / dead points.

Remarkably, this algorithm has not only provided us a very good estimate of the evidence, the dead points also form a posterior, since the weights w_i are exactly posterior weights from (6).

- A few steps of the above may stand out. The most important is, how do we defined convergence? Generally, we want to be sure we have sampled most of the evidence. We can approximate an upperbound on how much evidence is left to accrue with:

$$\Delta\mathcal{Z}_i \approx \mathcal{L}_{max} X_i \tag{17}$$

Then we can define a quantity dlogz, such that

$$\Delta\ln\mathcal{Z}_i = \ln(\mathcal{Z}_i + \Delta\mathcal{Z}_i) - \ln\mathcal{Z}_i \tag{18}$$

Generally it's best to think of a fractional stopping criteria:

$$f \leq \frac{\Delta\ln\mathcal{Z}_i}{\ln(\mathcal{Z}_i + \Delta\mathcal{Z}_i)} = 1 - \frac{\ln\mathcal{Z}_i}{\ln(\mathcal{Z}_i + \Delta\mathcal{Z}_i)} \tag{19}$$

that is, when the remaining evidence to accrue constitutes less than f of the total, stop (where $f \in (0, 1)$, usually a number like 0.02)