ASTR693B: Bayesian Statistics

Paul T. Baker

10 January 2018

Resources: I have found these three texts to be useful over the years. There are plenty of others out there.

- Allen B Downey, *Think Bayes: Bayesian Statistics Made Simple*, Green Tea Press (2012)
- Phil Gregory, *Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach*, Cambridge University Press (2010)
- William M. Bolstad, *Understanding Computational Bayesian Statistics*, Wiley (2009)

1 Introduction

Bayesian probability theory defines probability as a measure of the degree of belief in an outcome or proposition. This is in contrast to frequentist theory which defines probability as a ratio of instances of occurrence against the total number of trials. The frequentist definition restricts probability statements to propositions about random variables. The Bayesian definition is more robust, as it can apply to any logical proposition.

There are two fundamental tasks of science: model selection and parameter estimation (or measurement). For the first we wish to make statements about the probability that a hypothesis or model is correct. This is a natural statement in the Bayesian framework. In the frequentist framework this statement is much harder. Instead can must ask: what is the probability (*frequency*) we would observe this outcome, if our model were incorrect? If the resulting probability (sometimes called a 'p-value' in this context) is small, then the model is likely to be correct. There are other frequentist model selection techniques, but all require performing many identical observations to tease out the long term frequency of occurrences, which can be problematic. In Bayesian statistics we just directly calculate the probability that a model is correct.

For parameter estimation frequentist results are often presented in terms of confidence intervals. For example a star's mass may be estimated to be $(1.40\pm0.15)M_\odot$, where the uncertainty represents a 90% confidence interval (assuming a particular distribution of measurement uncertainty). This means that *if many new, independent measurements are made, 90% of new measurements would fall in this range.* This is not the same as *there is a 90% probability that the true mass is in this range.* In Bayesian statistics the we directly compute the probability distribution function for the parameter. There is no hidden assumption that the measurements follow a normal (or any other) distribution. And the Bayesian 90% **credible interval** really does mean there is a 90% probability that the true mass is in the given range.

2 Probability

The basic mathematical definitions of probability are as follows. A probability, p, must fall into the range $[0, 1]$, where 0 and 1 correspond to absolutely false and certainly true, respectively.

$$p \in [0, 1] \tag{1}$$

Probabilities describe exhaustive outcomes: if there are several, independent probable outcomes, the total probability of *any* outcome is 1. That is to say, one of the possible outcomes must occur. This can also be stated that sum of the probability that proposition A is true, $p(A)$, and the probability that A is false (or 'not A' is true) is unity.

$$\sum_i p_i = 1 \tag{2}$$
$$p(A) + p(\text{not}\, A) = 1$$

The probability of a proposition A can be conditional upon a second proposition B. The probability that A is true given B is true is $p(A \mid B)$. Because Bayesian probability applies to any logical proposition, we can apply the usual rules of deductive logic to them:

$$p(A\, \text{and}\, B) = p(A) \cdot p(B \mid A) \tag{3}$$
$$p(A\, \text{or}\, B) = p(A) + p(B) - p(A\, \text{and}\, B) \tag{4}$$

The logical *or* and *and* operators are commutative, associative, and distributive.

By using equation (3) and the commutativity of the logical *and* we can derive **Bayes' Theorem**:

$$p(A \mid B) = \frac{p(A) \cdot p(B \mid A)}{p(B)} \tag{5}$$

This doesn't seem like a terribly exciting result, but it is incredibly useful.

2.1 Cookie Problem

There are two indistinguishable cookie jars.

- Jar 1 has 20 chocolate cookies and 20 vanilla cookies.
- Jar 2 has 10 chocolate cookies and 30 vanilla cookies.

You randomly select a jar, then randomly take a cookie from that jar. The cookie is vanilla. What is the probability you are holding Jar 2?[1]

Solution: First, we write the question in terms of a conditional probability: the probability of holding Jar 2 given that we got a vanilla cookie. Now we use Bayes' Theorem:

$$p(J_2 \mid v) = \frac{p(J_2) \cdot p(v \mid J_2)}{p(v)}$$

All of the quantities on the right-hand-side are easily calculable. We have even odds of having selected Jar 2, so $p(J_2) = 1/2$. The probability of getting vanilla from Jar 2 is $p(v \mid J_2) = 3/4$. The probability of a getting vanilla under any Jar is $p(v) = 5/8$. Finally, $p(J_2 \mid v) = 3/5$.

[1] shamelessly adapted from *Think Bayes*, A. Downey

Using Bayes' Theorem we rewrote the probability in terms of a **likelihood function** (which tells us the probability of observing our data, v, if our model, J_2 is correct) and the **prior probability** for J_2. The probability on the left-hand-side is called the **posterior probability**, the probability after the experiment is conducted.

2.2 A Medical Test's Accuracy

A rare condition, affects 1 in 100 000 people of your demographic group. There is a test for the condition that is 98% accurate. This means that 98% of test takers who have the condition test positive (the other 2% get false negative results), and 98% of those who do not have the condition test negative (the other 2% get false positive results). You decide to take the test and receive a positive result. What is the probability you have the condition?[2]

Solution: Again we use Bayes' Theorem:

$$p\,(c \mid +) = \frac{p\,(c) \cdot p\,(+ \mid c)}{p\,(+)}$$
$$p\,(+) = p\,(+ \mid c) \cdot p\,(c) + p\,(+ \mid h) \cdot p\,(h)$$

where c is the condition, h is healthy, and + is a positive test result. We determine the probability of getting a positive result by adding the conditional probabilities for both health cases. Finally, we see $p\,(c \mid +) \approx 0.005$.

3 Data Analysis

3.1 Bayes' theorem

The basic problem of data analysis is determining the degree of belief in a hypothesis $\mathcal{H}$ given some observed data d. Using Bayes' theorem this becomes:

$$p\,(\mathcal{H} \mid d) = \frac{p\,(\mathcal{H}) \cdot p\,(d \mid \mathcal{H})}{p\,(d)}. \tag{6}$$

We will consider each of the terms on the right-hand-side separately below.

3.1.1 Likelihood: $\mathcal{L} = p\,(d \mid \mathcal{H})$

The likelihood is the probability of observing your data, if your hypothesis were correct. This sounds backwards, but it is usually straightforward to determine (as in the cookie problem 2.1). A likelihood assumes that the data is drawn from a particular probability distribution function (PDF). If your data follows a known distribution like a powerlaw or a gamma distribution this is easy.

$$\mathcal{L}(x) = PDF(x)$$

When your data is described by a deterministic model (not a statistical one), things are a bit different. In these cases instrument noise will still be statistical, so the likelihood is based on the

[2] adapted from the excellent chapter on probability on *Statistical and Thermal Physics*, H. Gould and J. Tobochnik. Problems of the "wild false positive" genre appear in almost every book on Bayesian statistics too.

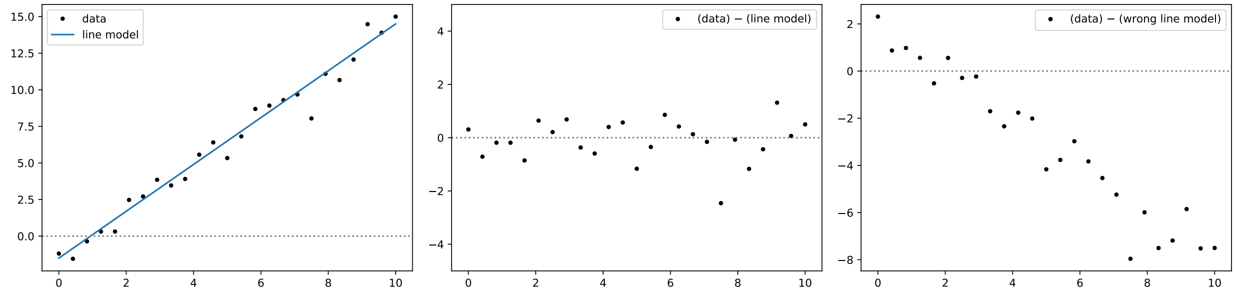

Figure 1: *left*: a deterministic line with Gaussian noise added. *middle*: subtracting the correct model from the data leaves Gaussian noise residuals. *right*: subtracting the wrong model leaves non-Gaussian residuals.

noise properties. Lets say our data, d, is built from a deterministic effect, $m^\star$, and instrument noise, n.

$$d = m^\star + n$$

We propose a model for the deterministic effect m. If our model is correct ($m = m^\star$), then the residual, r, is Gaussian noise.

$$r = d - m = n$$

And the likelihood that an individual datum is described by our deterministic model is the likelihood that the residual is described by the noise distribution!

The likelihood for the full data set is the product of the individual likelihoods

$$p\,(d \mid \mathcal{H}) = \prod_i p\,(d_i \mid \mathcal{H}) \tag{7}$$

For Gaussian noise with standard deviation, σ, the likelihood is:

$$p\,(n_i \mid \mathcal{H}) = \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{{n_i}^2}{2\sigma^2}\right), \tag{8}$$

so

$$p\,(d \mid \mathcal{H}) = \prod_i \frac{1}{\sqrt{2\pi}\,\sigma}\,\exp\left(-\frac{(d_i - m_i)^2}{2\sigma^2}\right) = \left(\frac{1}{\sqrt{2\pi}\,\sigma}\right)^N \exp\left(-\frac{\sum_i (d_i - m_i)^2}{2\sigma^2}\right) \tag{9}$$

More generally the noise may defined as a covariance matrix, $\mathbf{C}$. This is useful if the noise variance changes or is correlated between samples. In this case it is convenient to write the likelihood of equation (9) in the language of linear algebra. We define the residual as a vector $\vec{r} \rightarrow r_i = d_i - m_i$

$$p\,(d \mid \mathcal{H}) = \frac{1}{\sqrt{(2\pi)^N\,\det \mathbf{C}}}\,\exp\left(-\frac{1}{2}\,\vec{r}\cdot\mathbf{C}^{-1}\cdot\vec{r}\right) \tag{10}$$

3.1.2 Prior probability: $\mathcal{P} = p\,(\mathcal{H})$

The prior probability is probability your hypothesis is true, before you have conducted your experiment. Many people feel a bit uneasy about priors. The choice of prior does in fact affect the outcome, so how is this okay? First, frequentist methods make a myriad of assumptions about what is known about the data. At least the Bayesian framework lays these assumptions bare, for all to see and scrutinize. Second, given a large number of iterative experiments the posterior converges to the correct. This point illustrates another of Bayes' theorem's strengths: it is straightforward to iteratively apply Bayes' theorem to successive experiments.

Lets say we observe some data d_1 and compute the posterior for our hypothesis, $p\left(\mathcal{H} \mid d_1\right)$. Now we collect some new data d_2. What is the posterior probability for $\mathcal{H}$ after this? The outcome of the first experiment is now part of our prior knowledge for the second. We can see this by considering the posterior after observing $(d_1 \,\mathrm{and}\, d_2)$, which we leave as an exercise for the reader in problem 5.2.

When using past posteriors as priors one must be careful not to *double count* data. Say you collect 5 years of observations and compute a posterior, then you collect 2 new years of observations. If you want to use the 5 year posterior as a prior, the new analysis must use *only* the new 2 years of data. If you conduct a joint analysis of all 7 years, you must return to the original prior. The 7 year data set is not **independent** of the original 5 year data set.

There are a few common choices for priors, when starting from scratch. If you look up a parameter in the literature, and it's stated with uncertainty, that should be your prior on that parameter. For instance if $x = x_0 \pm \Delta x$ with uncertainty at the 1σ level, then

$$p\left(x\right) = \mathcal{N}(x_0, \Delta x^2).$$

An **ignorance prior** is usually a safe choice, where all outcomes are assumed to be equally likely. For a parameter this can be expressed as a uniform distribution. Uniform distributions are only properly defined for finite domains. This is perfect for bounded parameters like angles

$$p\left(\phi\right) = \mathcal{U}(0, 2\pi).$$

For unbounded parameters you can use a trick of the normal distribution

$$p\left(x\right) = \mathcal{U}(-\infty, \infty) = \lim_{\nu\to\infty} \mathcal{N}(0, \nu).$$

Practically, this is often achieved as

$$p\left(x\right) = \mathcal{N}(0, \mathtt{MAX}),$$

where `MAX` is the maximum `int` or `float` allowed by the computer.

In some cases a log-uniform prior is the preferred ignorance prior

$$p\left(\log_b x\right) = \mathcal{U}(e_0, e_1).$$

This is useful when a parameter could span a large range of decades (for $b = 10$, 10^{e_0} to 10^{e_1}) and you have no information about the scale of the effect. This prior is used for the gravitational wave background amplitude in some NANOGrav analyses.

To see why this helps, we can consider a parameter x that could be anywhere in the range 0.1 to 100. Assuming a uniform prior ($p(x) = \mathrm{const}$), we can compare the prior probability that x is in the smallest ($x \in [0.1, 1]$) and largest ($x \in [10, 100]$) decades.

$$\frac{\int_{10}^{100} p\left(x\right) \mathrm{d}x}{\int_{0.1}^{1} p\left(x\right) \mathrm{d}x} = \frac{90}{.9} = 100$$

The uniform prior heavily favors the largest decade. It states that the large values are favored over the small. In cases when we wish to detect a new phenomenon, we want the exact opposite. We expect a small value, or else we would have detected it already! It is these cases where the log-uniform prior is best.

3.1.3 Bayesian evidence: $\mathcal{E} = p(d)$

The Bayesian evidence is the probability of observing the data under any circumstances. It is best thought of as a normalization. It can be constructed by summing or integrating over the space of hypotheses. In most cases this is easier said than done. Explicitly this can look like:

$$p(d) = \sum_i p(d \mid \mathcal{H}_i) \cdot p(\mathcal{H}_i), \tag{11}$$

which is exactly what one must do to normalize the posterior. The Bayesian evidence is sometimes called the **marginal likelihood** as the normalization processes is equivalent to marginalizing the likelihood over **all** parameters.

The evidence is often challenging to compute because the tails of the distribution can be very important. Common numerical integration methods, like trapezoid and Simpson's sums, can become intractable for as few as three free parameters. Many computational methods, such as Markov chain Monte Carlo, avoid this problem by working with ratios of the posterior, so the evidence cancels out.

The evidence is important in problems of model selection, where multiple parameterized models must be compared. This occurs, for example, when trying to decide whether data are best fit by a straight line or a quadratic.

3.2 Marginalization

Sometimes we want to calculate a quantity like $p(d \mid \mathcal{H})$, but we only know how to calculate $p(d \mid \mathcal{H}, x)$. This can happen if our hypothesis is parameterized. For example, if we are trying to fit a line to data the slope of the line will appear in the likelihood. We can calculate the likelihood for any given x, but which x do we choose? In these cases x is said to be a **nuisance parameter**. To compute $p(d \mid \mathcal{H})$, we must first **marginalize** over x. That is, we average over all possible x making the likelihood a statement about any x instead of a single choice.

$$p(d \mid \mathcal{H}) = \int \mathrm{d}x \, p(x \mid \mathcal{H}) \, p(d \mid \mathcal{H}, x) \tag{12}$$

Note that the average is weighted by the prior on x.

Sometimes it is possible to compute these integrals analytically, but more often than not one must numerically integrate. As the number of nuisance parameters grows, the methods for computing these integrals must become more and more sophisticated.

For parameterized hypotheses notation like $\mathcal{H}(x, y)$ is common. In the case of two parameters we can marginalize over one only, if we choose:

$$p(d \mid \mathcal{H}(y)) = \int \mathrm{d}x \, p(x \mid \mathcal{H}) \, p(d \mid \mathcal{H}(x, y)) \,. \tag{13}$$

Another common notation replaces the hypothesis with the parameters entirely. Using a parameter vector, we can write the marginal likelihood (or evidence) as

$$p(d) = \int p(\vec{x}) \, p(d \mid \vec{x}) \; \mathrm{d}^N x = \mathcal{E}, \tag{14}$$

where $\vec{x}$ is the vector of N parameters that define the model.

3.3 Parameter Estimation

In cases of parameter estimation, we want to determine the measured or best fit parameters from observed data. This may be the result of fitting a curve to data, where the fit parameters correspond to physical observables. If we calculate the posterior probability distribution (either analytically or numerically) for the parameters, then the inference problem is a matter of interpreting this result.

The **best fit** parameters are not rigorously defined. The true output of the analysis *is* the posterior. Any best fit or uncertainty is a summary of the full result. Many prefer the **mean** to report as the best fit parameter, calculated

$$\langle x \rangle = \int x \, p\left(x \mid d\right) \, \mathrm{d}x. \tag{15}$$

where $p\left(x \mid d\right)$ is the posterior probability of parameter, x, given the data, d.

When the posterior is computed numerically it can be challenging to compute the mean, especially if there are many fit parameters. In these cases it is more convenient to report the **median** with the added benefit that the cumulative sums used to compute the median will produce **credible intervals** too. The median, $x_{50\%}$, is found by inverting

$$0.50 = \int_{-\infty}^{x_{50\%}} p\left(x \mid d\right) \, \mathrm{d}x. \tag{16}$$

One can perform a similar inversion to find any percentile of the posterior.

To state the uncertainty on the parameters credible intervals are used. For a 90% credible interval we compute the 5% and 95% percentiles of the parameter. It is becoming more common to quote a median with pseudo-1σ uncertainty by computing the 18%, 50%, and 84% percentiles. This gives a 66% credible interval which is approximately the size of a 1σ frequentist confidence interval. One can compute upper or lower limits by similarly integrating the posterior.

Another commonly quoted best fit is the maximum probability parameters. These are sometimes abbreviated as MAP for *maxima a posteriori*. The MAP is easy to calculate, and it ensures the best fit parameters come from the same point in parameter space. Individually computing the median or mean for each of several parameters independently can lead to a worse combined result. This is rare and usually only happens when there are nontrivial correlations in the parameters. In any case, the MAP simultaneously maximizes all parameters, and can never have this problem.

3.4 Model Selection

In many cases there is more than one hypothesis or model that fits the data to some extent. How do we decide which is preferred? A commonly used method is the maximum likelihood ratio. In this method one computes the likelihood of the best fit parameters for each model and compares. This method has an important shortcoming: it ignores the complexity of the models. A model with more free parameters should always achieve a better fit. So why not fit 10 data points with a 10th order polynomial? It fits perfectly every time. This complexity consideration is sometimes referred to as the **Occam factor**.

Bayesian analysis naturally takes care of the Occam factor by using the **marginal likelihood ratio** or evidence ratio (same thing). The marginalization integral takes into account the probability across the whole prior volume, penalizing models with many or widely ranging parameters.

The probability for a model $\mathcal{H}_i$ is calculated from Bayes theorem

$$p\left(\mathcal{H}_i \mid d\right) = \frac{p\left(\mathcal{H}_i\right) p\left(d \mid \mathcal{H}_i\right)}{p\left(d\right)} \tag{17}$$

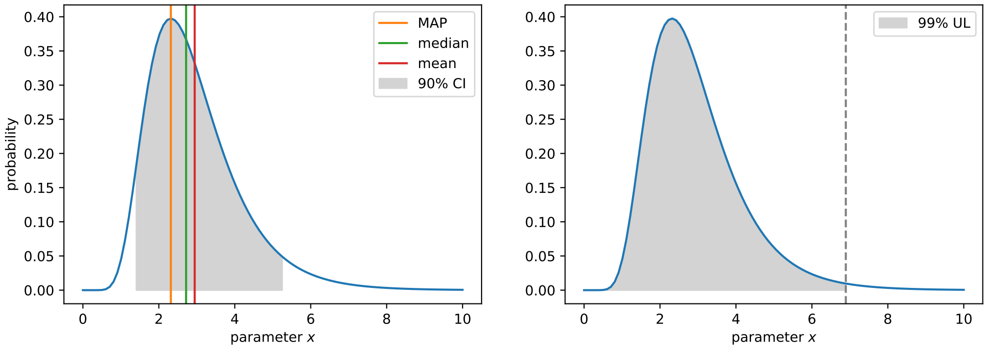

Figure 2: Example parameter estimation for a log-normal distribution. The distribution is skewed so the mean, median, and MAP do not agree. By integrating the probability distribution we can determine credible intervals (*left*) or an upper limit (*right*).

If we want to compare two models, 0 and 1, we can compute the **odds ratio**

$$\mathcal{O}_{1,0} = \frac{p\left(\mathcal{H}_1 \mid d\right)}{p\left(\mathcal{H}_0 \mid d\right)} = \frac{p\left(\mathcal{H}_1\right)}{p\left(\mathcal{H}_0\right)} \frac{p\left(d \mid \mathcal{H}_1\right)}{p\left(d \mid \mathcal{H}_0\right)} = \frac{p\left(\mathcal{H}_1\right)}{p\left(\mathcal{H}_0\right)} \mathcal{B}_{1,0} \tag{18}$$

It is common to break the odds ratio into the prior odds ratio and the marginal likelihood ratio or **Bayes factor**, $\mathcal{B}_{1,0}$.

$$\mathcal{B}_{1,0} = \frac{p\left(d \mid \mathcal{H}_1\right)}{p\left(d \mid \mathcal{H}_0\right)} \tag{19}$$

It is typical to report Bayes factors rather than full odds ratios, as the priors on various models are usually seen as a personal choice. In the case that there is no *a priori* reason to favor either model, there prior odds ratio is 1 and the odds ratio is the Bayes factor.

After computing the odds, we can reverse engineer the probability for a particular model. Starting from the definition of probability (equation 2), and using the definition of $\mathcal{O}_{i,0} = p\left(\mathcal{H}_i \mid d\right) / p\left(\mathcal{H}_0 \mid d\right)$, we can show that

$$p\left(\mathcal{H}_i \mid d\right) = \frac{\mathcal{O}_{i,0}}{\sum_i \mathcal{O}_{i,0}}. \tag{20}$$

For two models $\mathcal{O}_{0,0} = 1$ so

$$p\left(\mathcal{H}_1 \mid d\right) = \frac{\mathcal{O}_{1,0}}{1 + \mathcal{O}_{1,0}} = \frac{1}{1 + \frac{1}{\mathcal{O}_{1,0}}}$$

Table 1 compares Bayes factors, their probability for two model comparison, and the frequentist σ used to summarize p-value style model selection. A Bayes factor less than 10 is pretty weak evidence.

Table 1: Comparing Bayes factors to frequentist 'σ'.

Bayes factor, $\mathcal{B}$	probability, p	$n\,\sigma$
3	0.75	$1.15\,\sigma$
10	0.91	$1.69\,\sigma$
100	0.99	$2.58\,\sigma$
370	0.997	$3\,\sigma$
1 744 277	0.999 999 4	$5\,\sigma$

4 Example

4.1 Model Selection: Is there a spectral line?

We have some noisy data containing a possible novel spectral line feature. The spectrum is measured as a temperature anomaly, dT, from the mean observed temperature in each frequency channel. Our instrument has 64 frequency channels and is known to have white, Gaussian radiometer noise at a level of $\sigma_N = 1$ mK.

Some theory predicts the line will have a gaussian shape:

$$dT = T_0 \exp\left(\frac{-(\nu - \nu_0)^2}{2{\sigma_L}^2}\right),$$

where T_0 is the amplitude of the feature that according to the theory can range from 0.01-100 mK. The frequency ν is measured in channel number for our instrument, and the theory says line should appear in channel $\nu_0 = 37$ with width $\sigma_L = 2$.

Of course, this theory could be wrong. In which case we should expect to see only the radiometer noise in our data. If we collect the data shown in Fig 3 with our instrument, what does it say about the theory?

We begin by stating our hypotheses. Let $\mathcal{H}_1$ be *there is a spectral line as described by the theory*, and let $\mathcal{H}_0$ be *there is no spectral line*. We wish to calculate the odds ratio between these two hypotheses. Since we have no *a priori* reason to favor one or the other we set the prior odds between the hypotheses to 1, and the odds ratio reduces to the Bayes factor

$$\mathcal{O}_{1,0} = \mathcal{B}_{1,0} = \frac{p\,(d \mid \mathcal{H}_1)}{p\,(d \mid \mathcal{H}_0)}. \tag{21}$$

First, we calculate the likelihood for the noise only model, $\mathcal{H}_0$. We know the properties of the radiometer noise (white, Gaussian), so

$$p\,(d \mid \mathcal{H}_0) = \frac{1}{\sqrt{2\pi}\,\sigma_N} \exp\left(-\frac{\sum {d_i}^2}{2{\sigma_N}^2}\right). \tag{22}$$

Next, we calculate the likelihood for the line model, $\mathcal{H}_1$. Under $\mathcal{H}_1$ the data in the i-th frequency channel is given by $d_i = dT_i + n_i$, where dT_i is the modeled temperature. After subtracting the modeled line the residuals are noise, so the likelihood is

$$p\,(d \mid \mathcal{H}_1, T) = \left(\frac{1}{\sqrt{2\pi}\,\sigma_N}\right)^2 \exp\left(-\frac{\sum (d_i - m_i)^2}{2{\sigma_N}^2}\right) \tag{23}$$

$$m_i = dT_i = T\,\exp\left(-\frac{\nu_i - {\nu_0}^2}{2{\sigma_L}^2}\right).$$

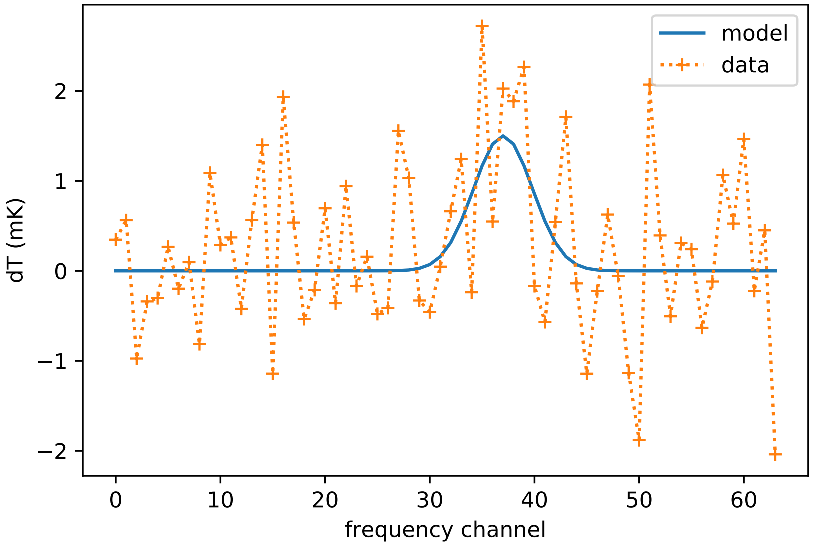

Figure 3: Measured and modeled spectrum. The amplitude of the modeled spectrum matches that of the simulation that was used to generate the data.

To compute $p\,(d \mid \mathcal{H}_1)$ we must first marginalize over the nuisance parameter, T.

$$p\,(d \mid \mathcal{H}_1) = \int \mathrm{d}T\, p\,(T \mid \mathcal{H}_1)\, p\,(d \mid \mathcal{H}_1, T) \tag{24}$$

Because T can fall into a huge range, $[10^{-2}, 10^{2}]$, we adopt a log-uniform prior

$$p\,(\log T \mid \mathcal{H}_1) = \frac{1}{\log T_{\mathrm{max}} - \log T_{\mathrm{min}}} \tag{25}$$

The companion Jupyter notebook[3] walks through the numerical calculation of these expressions for the data shown in Fig 3. The final result is an odds ratio of $\mathcal{O}_{1,0} \approx 124$. This means the line model is favored over the noise only model at 124 : 1 or about 99.2%. This is good, but not definitive evidence.

4.2 Parameter Estimation: What is the best fit value of T?

After determining the odds for the spectral line being present, we can now calculate the best fit amplitude for the feature. We effectively calculated the posterior for T (assuming the spectral line model) during the model selection process. By Bayes' theorem it is proportional to the integrand of equation (24).

$$p\,(T \mid \mathcal{H}_1, d) \propto p\,(T \mid \mathcal{H}_1)\, p\,(d \mid \mathcal{H}_1, T)$$

This function is determined numerically for the data shown in Fig 3 in the companion Jupyter notebook. We integrate this function numerically to determine the median and 90% credible interval: $T = 1.7 \pm 0.7$.

[3] https://www.github.com/paulthebaker/bayes_talk

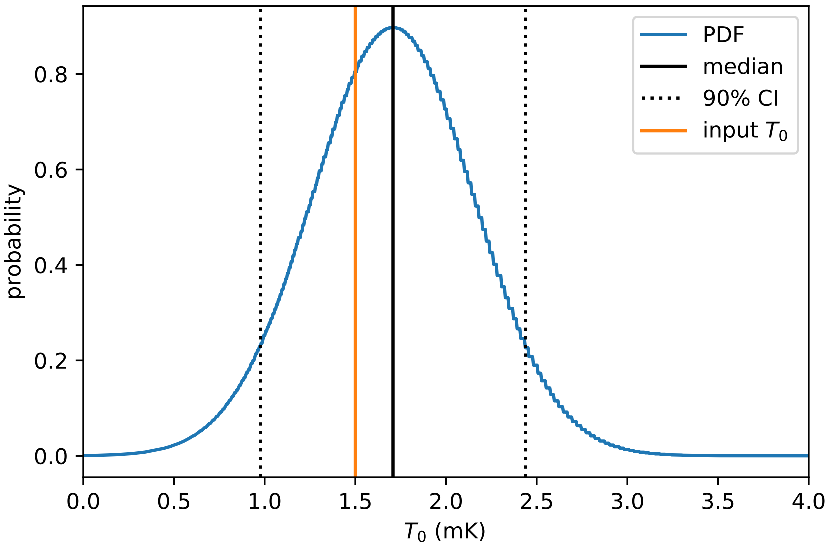

Figure 4: Posterior distribution function (PDF) for spectral line amplitude, T. The median and 90% credible interval are shown along with the input T_0 used to simulate the data.

5 Problems

5.1 Monty Hall

A favorite game from a classic a game show presented the contestant with three doors. Behind one was a fabulous prize, while the others held only junk. First, the contestant chooses a door. Next, the host opens one of the unchosen doors to reveal junk[4]. Finally, the contestant is given the opportunity to switch doors, taking home whatever prize is behind the chosen door.

1. What is the initial probability that the prize is in the door selected by the contestant?
2. After the host opens a door, what is the probability that the prize is in the remaining door?
3. What if there were 5 doors (still one prize), and the host opened three to reveal junk before the contestant's opportunity to switch?

5.2 Multiple observations

Show that

$$p\left(\mathcal{H} \mid d_1, d_2\right) = \frac{p\left(\mathcal{H} \mid d_1\right) \cdot p\left(d_2 \mid \mathcal{H}\right)}{p\left(d_2\right)},$$

where d_1 and d_2 are data from two **independent** observations and a comma is used to denote the logical *and.*

[4] The host knows which door the prize is in and always reveals junk. If the contestant selected the prize door, he opens one of the others at random.

5.3 Cookie Jar Redux

Starting from the Cookie Problem of example 2.1, you draw a second cookie from the same jar, and it's vanilla too. What is the probability you are holding Jar 2 now?

5.4 Using Hubble's law to find distance

Hubble's law relates the distance to a galaxy, x, to it's apparent recessional velocity, v, due to the expansion of the universe.

$$v = H_0\, x,$$

where H_0 is the Hubble constant.

You observe a galaxy with a recessional velocity of $v = 100 \pm 5$ km/s, where the uncertainty represents 1σ Gaussian noise. Determine the posterior probability for the distance to this galaxy given the velocity observation. Write down the equation for $p\,(x \mid v)$. Set up any integrals, but don't evaluate them. In each case your answer should be a function of v only.

1. What is $p\,(x \mid v)$, assuming a point estimate of $H_0 = 70$ km/s/Mpc? Remember to account for the uncertainty in v.

2. What is $p\,(x \mid v)$, accounting for 1σ uncertainty in $H_0 = 72 \pm 3$ km/s/Mpc[5] by marginalizing? Start by determining $p\,(x \mid v, H_0)$.

3. What is $p\,(x \mid v)$, accounting for uniform uncertainty in $H_0 = 70 \pm 10$ km/s/Mpc?

5.5 Using Hubble's law to find distance, calculated

Warning: this problem requires a computer.

1. Evaluate and plot the posteriors from parts 1, 2, and 3 of the previous problem (5.4).

2. Determine the median and a 90% CI on the distance, x, for each case.

3. What if you use the Hubble constant from the Planck mission[6], $H_0 = 67.7 \pm 0.5$ km/s/Mpc?

5.6 Spectral line redux

Warning: this problem requires a computer and is **very** challenging.

1. Redo the calculation of the odds for the spectral line problem from section 4. This time let there be uncertainty in the predicted channel for the line $\nu_0 = 37 \pm 5$, where the uncertainty is given at the 1σ level.

2. Use the joint posterior on T and ν_0 to determine the best fit amplitude and channel for the line.

[5] V. Bonvin, *et al.* "H0LiCOW – V. New COSMOGRAIL time delays of HE 0435–1223: H_0 to 3.8% precision from strong lensing in a flat ΛCDM model". *MNRAS*, **465** (4): 4914–4930 (2016).

[6] Planck Collaboration, P.A.R. Ade, *et al.* "Planck 2015 results – XIII. Cosmological parameters". *A&A*, **594** A13 (2016).