BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2018

Caltech

Donna and Benjamin M.
Rosen Bioengineering Center

This document was prepared at Caltech with financial support from the Donna and Benjamin M. Rosen Bioengineering Center.

5 Introduction to Bayesian modeling

In the first lecture and in the homework problem on Bayes's theorem as a model for learning, we learned about Bayes's theorem as a way to update a hypothesis in light of new data. We use the word "hypothesis" very loosely here. Remember, in the Bayesian view, probability can describe the plausibility of any proposition. The value of a parameter is such a proposition. In this lecture, we will learn about the Bayesian approach to parameter estimation. This stands in contrast to what we have been doing with frequentist methods when we have been using the plug-in principle to get estimates about expectation values (or, more generally, statistical functionals) from a generative distribution by approximating it using an empirical distribution.

We will approach this problem by example, starting with a very simple problem, estimating the parameter in a one-parameter model.

5.1 Bayes's theorem as applied to simple parameter estimation

We will consider one of the simplest examples of parameter estimation. Let's say we measure a parameter θ in multiple independent experiments. This could be beak depths of finches, fluorescence intensity in a cell, a dissociation constant for two bound proteins, etc. The possibilities abound. To have a concrete example in mind for this example, let's assume we are measuring the length of *C. elegans* eggs.

Our measurements of this parameter are $y \equiv \{y_1, y_2, \ldots y_n\}$. Our "hypothesis" in this case, is the value of the parameter θ. We wish to calculate $g(\theta \mid y)$, the posterior probability distribution for the parameter θ, given the data. Values of θ for which the posterior probability is high are more probable (that is, more plausible) than those for which is it low. The posterior $g(\theta \mid y)$ codifies our knowledge about θ in light of our data y.

To compute the posterior probability, we use Bayes's theorem.

$$g(\theta \mid y) = \frac{f(y \mid \theta)\, g(\theta)}{f(y)}. \tag{5.1}$$

Since the evidence $f(y)$ does not depend on the parameter of interest, θ, it is really just a normalization constant, so we do not need to consider it explicitly. Specification of the likelihood and prior is sufficient for the posterior, since

$$f(y) = \int \mathrm{d}\theta\, f(y \mid \theta)\, g(\theta) \tag{5.2}$$

to ensure normalization of the posterior $g(\theta \mid y)$. So, we have just to specify the likelihood $f(y \mid \theta)$ and the prior $g(\theta)$.

Specification of the likelihood/prior pair is what statistical modeling is all about. We begin with the likelihood.

5.2 The likelihood

To specify the likelihood, we have to ask what we expect from the data, given a value of θ. If there are no errors or confounding factors at all in our measurements, we expect $y_i = \theta$ for all i. In this case

$$g(y \mid \theta) = \prod_{i=1}^{n} \delta(y_i - \theta), \tag{5.3}$$

the product of Dirac delta functions. Of course, this is really never the case. There will be some errors in measurement and/or the system has variables that confound the measurement. What, then should we choose for our likelihood?

That choice is of course dependent the story/theoretical modeling behind data generation. For our purposes here, we shall assume our data are generated from a Gaussian likelihood. Since this distribution gets heavy use, I will pause here to talk a bit more about it.

5.3 The Gaussian distribution

A univariate Gaussian, or Normal, probability distribution has a probability density function (PDF) of

$$f(y \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\,\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]. \tag{5.4}$$

The parameter μ is called the mean of the distribution and σ^2 is called the variance, with σ being called the standard deviation. Importantly, the mean and standard deviation in this context are *names of parameters* of the distribution; they are not what you compute directly from data.

The **central limit theorem** says that any quantity that emerges from a large number of subprocesses tends to be Gaussian distributed, provided none of the subprocesses is very broadly distributed. We will not prove this important theorem, but we make use of it when choosing likelihood distributions based on the stories behind the generative process. Indeed, in the simple case of estimating a single parameter where many processes may contribute to noise in the measurement, the Gaussian distribution is a good choice for a likelihood.

More generally, the multi-dimensional Gaussian distribution for $y = (y_1, y_2, \cdots, y_n)$ is

$$f(y \mid \mu, \sigma) = (2\pi)^{-\frac{n}{2}} \left(\det \Sigma\right)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(y-\mu)^T \cdot \Sigma^{-1} \cdot (y-\mu)\right], \tag{5.5}$$

where $\mu = \{\mu_1, \mu_2, \ldots, \mu_n\}$ is an array of means (again, here "mean" is the name of the *parameter* of the Gaussian, not of the mean of a measurement, which does not even make sense here, since y_i is a single measurement). The parameter Σ is a symmetric positive definite matrix called the **covariance matrix**. If off-diagonal entry Σ_{ij} is nonzero, then y_i and y_j are correlated. In the case where all y_i are independent, all off-diagonal terms in the covariance matrix are zero, and the multidimensional Gaussian distribution reduces to

$$f(y \mid \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma_i^2}\right], \tag{5.6}$$

where σ_i^2 is the ith entry along the diagonal of the covariance matrix. This is the variance associated with measurement i. So, if all independent measurements have the same variance and mean, which is to say that the measurements are **independent and identically distributed** (i.i.d.), the multi-dimensional Gaussian reduces to

$$f(y \mid \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]. \tag{5.7}$$

5.4 The likelihood revisited: and another parameter

For the purposes of this demonstration of parameter estimation, we assume the Gaussian distribution is a good choice for our likelihood for repeated measurements. We have to decide how the measurements are related to specify how many entries in the covariance matrix we need to specify as parameters. It is often the case that the measurements are i.i.d, so that only a single mean and variance are specified. So, we choose our likelihood to be

$$f(y \mid \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right]. \tag{5.8}$$

By choosing this as our likelihood, we are saying that we expect our measurements to have a well-defined mean μ with a spread described by the variance, σ^2.

But wait a minute; we had a single parameter, θ, that we sought to estimate, and now we now have another parameter, σ, beyond the one we're trying to measure (which we are now calling μ). So, our statistical model has *two* parameters, and Bayes's theorem now reads

$$g(\mu, \sigma \mid y) = \frac{f(y \mid \mu, \sigma)\, g(\mu, \sigma)}{f(y)}. \tag{5.9}$$

After we compute the posterior, we can still find the posterior probability distribution we are after by marginalizing.

$$g(\mu \mid y) = \int \mathrm{d}\sigma\, g(\mu, \sigma \mid y). \tag{5.10}$$

5.5 Choice of prior

Now that we have defined a likelihood, we know what the parameters are and we can define a prior, $g(\mu, \sigma)$. As is often the case, we assume μ and σ are independent of each other, so that

$$g(\mu, \sigma) = g(\mu)\, g(\sigma). \tag{5.11}$$

How might we choose prior distributions for μ and σ? Remember, the prior probability distribution captures what we know about the parameter before we measure data. I often like to sketch how I think the probability density function of a parameter will look and then find a named distribution that looks like that. Generally, I think it is wise to choose a **weakly informative prior**. I think the idea is well-described in the Stan wiki on priors, which say, "the [weakly informative] prior rules out unreasonable parameter values but is not so strong as to rule out values that might make sense." In other words, you want to draw your prior distribution broad enough such that it covers all parameter values that are even somewhat reasonable, but rules out absurd parameter values.

For the current contrived example of *C. elegans* eggs, we can guess that the egg length should be about 50 µm, but we are not to sure about this. So, we take $g(\mu)$ to be Gaussian with a mean of 50 µm, but a variance of 20 µm. That is,

$$g(\mu) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} \exp\left[-\frac{(\mu - \mu_\mu)^2}{2\sigma_\mu^2}\right], \tag{5.12}$$

with $\mu_\mu = 50$ µm and $\sigma_\mu = 20$ µm. This means that getting a very tiny egg length of, say, 10 µm is unlikely, as is a very large egg of 90 µm.

For $g(\sigma)$, we might think that the egg length may vary about five or ten microns, but not much more than that. We could again choose a Gaussian prior, with

$$g(\sigma) = \frac{1}{\sqrt{2\pi\sigma_\sigma^2}} \exp\left[-\frac{(\mu - \mu_\sigma)^2}{2\sigma_\sigma^2}\right], \tag{5.13}$$

with $\mu_\sigma = 5$ µm and $\sigma_\sigma = 2$ µm.

The exact functional form of the prior is not so important. In this case, we have the obvious issue that there is nonzero probability that μ or σ could be negative, which we know is unphysical. We could refine our prior distribution to make sure this does not happen. With any approach we choose, the prior should roughly match what we would sketch on a piece of paper and cover any reasonable parameter values and exclude any that are unreasonable (or unphysical).

5.6 Succinctly stating the model

Our model is complete, which means that we have now completely specified the posterior. We can write it out.

$$
\begin{aligned}
g(\mu, \sigma \mid y) = \frac{1}{f(y)} \Bigg\{ \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} &\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right] \\
&\times \frac{1}{\sqrt{2\pi 20^2}} \exp\left[-\frac{(\mu - 50)^2}{2\cdot 20^2}\right] \\
&\times \frac{1}{\sqrt{2\pi 2^2}} \exp\left[-\frac{(\mu - 5)^2}{2\cdot 2^2}\right] \quad \Bigg\},
\end{aligned} \tag{5.14}
$$

with

$$
f(y) = \int \mathrm{d}\mu \int \mathrm{d}\sigma \; \{\text{term in braces in the above equation}\}. \tag{5.15}
$$

Oh my, this is a mess, even for this simple model! Even though we have the posterior, it is very hard to make sense of it. Essentially the rest of the course involved making sense of the posterior, which is the challenge. It turns out that writing it down was relatively easy!

One of the first things we can do to make sense of our model, and also to specify it, is to use a shorthand for model specification. First of all, we do not need to specify the evidence, since it is always given by integrating the likelihood and prior; that is by fully marginalizing the likelihood. So, we will always omit its specification. Now, we would like to have a notation for stating the likelihood and prior. English works well.

> The parameter μ is Gaussian distributed with mean 50 µm and standard deviation 20 µm.
>
> The parameter σ is Gaussian distributed with mean 5 µm and standard deviation 2 µm.

> The egg lengths are i.i.d. and are Gaussian distributed with mean μ and standard deviation σ.

This is much easier to understand. We can write this with a convenient, and self evident, shorthand.[7]

$$\mu \sim \text{Norm}(50, 20), \tag{5.16}$$

$$\sigma \sim \text{Norm}(5, 2), \tag{5.17}$$

$$y_i \sim \text{Norm}(\mu, \sigma)\ \forall i. \tag{5.18}$$

Here, the symbol $\sim$ may be read as "is distributed as." The above three lines are completely sufficient to specify our model. Because we will be using a probabilistic programming language in practice, we will almost never need to code up any nasty mathematical expressions in our modeling.

5.7 A Bayesian workflow

In coming tutorials, we will learn numerical techniques for gaining useful information, usually expectations and marginalized distributions (which are a kind of expectation) from the posterior. The most powerful technique is **Markov chain Monte Carlo** (MCMC), which we will do extensively using Stan. But before you start computing expectations of the posterior you should make sure that your generative model can produce data sets that make sense and that it accurately captures your prior belief about the generative process. This procedure is known as **prior predictive checking**. Once you are pleased with that, you should check to make sure that the numerical techniques (usually MCMC) you use to compute the expectations can provide reliable results. You should also check that the addition of your data changes the posterior beyond what was already known from the prior. When all of those checks are in place, you can proceed to compute expectations from the posterior, which is the process we ofter refer to as parameter estimation.

We will learn how to do prior predictive checks in the tutorials next week, and we will jump into summarizing the posterior right after that. Throughout the next few weeks, we will discuss techniques to ensure that your inferences are sound. For a more detailed discussion on a principled Bayesian workflow, I encourage you to read Michael Betancourt's excellent blog post on the topic.

[7] I understand that I should be providing units on all parameters that I am specifying with numbers. I am not doing this here, nor throughout the course, to avoid notational clutter and to maintain focus on the modeling.