BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2017

2 Parameter estimation from repeated measurements

In the last lecture, 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 how to do a Bayesian estimate of a parameter. Before we do, a note on notation.

2.1 Notation of parts of Bayes's Theorem

In the last lecture, you probably noticed, and were perhaps frustrated by, the notational overloading of the letter P. Using P was useful in the last lecture to avoid confusion as we went from discussing the desiderata of a measure of plausibility and in discussing of probabilities of outcomes. To help aid in notation, we will use the following conventions going forward.

- Probability densities describing measured data are denoted with f.
- Probability densities describing parameter values, hypotheses, or other non-measured quantities, are denoted with g.
- A set of parameters for a given model are denoted θ.

So, if we were to write down Bayes's theorem for a parameter estimation problem, it would be

$$g(\theta \mid D, I) = \frac{f(D \mid \theta, I)\, g(\theta \mid I)}{f(D \mid I)}. \tag{2.1}$$

For, probabilities written with a g denote the prior or posterior, and those with an f denote the likelihood or evidence.

Furthermore, since the contents of I are always implicitly assumed to be part of any statistical model we will construct, we will henceforth not explicitly show it to reduce clutter. So, we write Bayes's theorem as

$$g(\theta \mid D) = \frac{f(D \mid \theta)\, g(\theta)}{f(D)}, \tag{2.2}$$

which is clearer notation, I think, for setting up our inference problems.

2.2 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 μ 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.

Our measurements of this parameter are $D = \{x_1, x_2, \ldots x_n\} \equiv \mathbf{x}$. Our "hypothesis" in this case, is the value of the parameter μ, so we have $\theta = \mu$. We wish to calculate $g(\mu \mid \mathbf{x})$, the posterior probability distribution for the parameter μ, given the data. Values of μ for which the posterior probability is high are more probable (that is, more plausible) than those for which is it low.

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

$$g(\mu \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \mu)\, g(\mu)}{f(\mathbf{x})}. \tag{2.3}$$

Since the evidence, $f(\mathbf{x})$ does not depend on the parameter of interest, μ, it is really just a normalization constant, so we do not need to consider it explicitly. We now have to specify the likelihood $f(\mathbf{x} \mid \mu)$ and the prior $g(\mu)$.

Specification of the likelihood/prior pair is what statistical modeling is all about. We will talk in most more depth about constructing these models in the next lecture. We need a little more background on probability distributions to do that, and we will get that in the tutorials for next week. For now, we will investigate an oft-used statistical model, that of a Gaussian likelihood with uninformative priors (with a precise definition of uninformative coming in the next lecture). The goal here is to show how you can compute and characterize the posterior distribution analytically.

2.3 The likelihood

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

$$g(\mathbf{x} \mid \mu) = \prod_{i=1}^{n} \delta(x_i - \mu), \tag{2.4}$$

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?

This question is made sharper if we think about the likelihood in terms of the *statistical model* we defined in the last lecture. It is the probability distribution that describes how the data relate to the parameter we are trying to measure. Indeed, specifying the likelihood is part of the modeling process. In Tutorial 3b, we will learn more about probability distributions, but for now we will introduce one useful distribution to use in our analyses.

2.4 The Gaussian distribution

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

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right]. \tag{2.5}$$

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 will make use of it when choosing likelihood distributions when we learn about building statistical models next week. 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 $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is

$$f(\mathbf{x} \mid \mu, \sigma) = (2\pi)^{-\frac{n}{2}} (\det \Sigma)^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\mu)^T \cdot \Sigma^{-1} \cdot (\mathbf{x}-\mu)\right], \tag{2.6}$$

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 x_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 x_i and x_j are correlated. In the case where all x_i are independent, all off-diagonal terms in the covariance matrix are zero, and the multidimensional Gaussian distribution reduces to

$$f(\mathbf{x} \mid \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right], \tag{2.7}$$

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

$$f(\mathbf{x} \mid \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i-\mu)^2\right]. \tag{2.8}$$

2.5 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 i.i.d, so that only a single mean and variance are specified. So, we choose our likelihood to be

$$f(\mathbf{x} \mid \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}. \tag{2.9}$$

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 now have another parameter, σ, beyond the one we're trying to measure. So, our statistical model has *two* parameters, and Bayes's theorem now reads

$$g(\mu, \sigma \mid \mathbf{x}) = \frac{f(\mathbf{x} \mid \mu, \sigma)\, g(\mu, \sigma)}{f(\mathbf{x})}. \tag{2.10}$$

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

$$g(\mu \mid \mathbf{x}) = \int_0^\infty \mathrm{d}\sigma\, g(\mu, \sigma \mid \mathbf{x}). \tag{2.11}$$

2.6 Choice of prior

Because the evidence $f(\mathbf{x})$ is entirely determined by the likelihood, prior, and normalization condition of the posterior, we need only to specify the likelihood and prior to get the posterior. We have chosen a Gaussian distribution for our likelihood, so now we need to specify $g(\mu, \sigma)$. The prior encodes what we know about the parameters *before* the experiments. The prior may be informed by previous experiments, as we discussed in section 1.9. We will talk in depth in the next lecture about choices of priors. For the present, we will assume that μ and σ are independent such that

$$g(\mu, \sigma) = g(\mu)\, g(\sigma). \tag{2.12}$$

Further, we will assume a **Uniform prior** for μ and a **Jeffreys prior** for σ. Specifically,

$$g(\mu) = \begin{cases} (\mu_{\max} - \mu_{\min})^{-1} & \mu_{\min} < \mu < \mu_{\max}, \\ 0 & \text{otherwise}, \end{cases} \tag{2.13}$$

and

$$g(\sigma \mid I) = \begin{cases} (\ln(\sigma_{\max}/\sigma_{\min})\ \sigma)^{-1} & \sigma_{\min} < \sigma < \sigma_{\max} \\ 0 & \text{otherwise.} \end{cases} \tag{2.14}$$

For $g(\mu)$, all values between $\mu_{\min}$ and $\mu_{\max}$ are equally likely. We have put bounds on the values that μ can take, and we will work in the limit where these bounds are far from any peak in the likelihood in what follows. Similarly, for $g(\sigma)$, all values of the logarithm of σ are equally likely (as we will derive in the next lecture), and it, too, has bounds.

2.7 The posterior

Now that we have specified the likelihood and prior, we have the posterior.

$$g(\mu, \sigma \mid \mathbf{x}) = \frac{c}{\sigma^{n+1}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2\right], \tag{2.15}$$

where we have absorbed all constants in to the normalization constant c[6].

So, we are done! We have now updated our knowledge of μ and σ. We could just plot the posterior distribution. We could show it as a contour plot in the μ-σ plane, for instance.

But, it would be nice to get the posterior into a bit of a cleaner form. We can show, after some algebraic grunge, that

$$\sum_{i=1}^{n} (x_i - \mu)^2 = n(\bar{x} - \mu)^2 + nr^2, \tag{2.16}$$

where

$$r^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{2.17}$$

is the sample variance and

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.18}$$

[6] We do this here for convenience, but when we do model selection later on, we will have to compute the evidence, so we should be careful about the normalization constants of the priors throughout our calculations.

is the sample mean. Thus, we have

$$g(\mu\,,\,\sigma\mid\mathbf{x}) = \frac{c\,\mathrm{e}^{-nr^2/2\sigma^2}}{\sigma^{n+1}}\,\exp\left[-\frac{n(\mu-\bar{x})^2}{2\sigma^2}\right]. \tag{2.19}$$

In this form, we immediately see that, regardless the value of σ, the most probable value of μ is $\bar{x}$. This is perhaps not surprising that the most probable value of μ is the sample mean, but it is pleasing how nicely it falls out of the analysis.

Now, it would really like to get a summary of the posterior to be able to report some nice numbers, like the most probable value of $\mu\,,\,\bar{x}$, instead of a plot.

2.7.1 The mean μ

We wanted to get $g(\mu\mid\mathbf{x})$ in the first place. As we said before, we can get that by marginalizing over σ.

$$\begin{aligned} g(\mu\mid\mathbf{x}) &= \int_0^\infty \mathrm{d}\sigma\, g(\mu\,,\,\sigma\mid\mathbf{x}) \\ &= c\int_0^\infty \frac{\mathrm{d}\sigma}{\sigma^{n+1}}\,\exp\left[-\frac{n(\mu-\bar{x})^2+nr^2}{2\sigma^2}\right]. \end{aligned} \tag{2.20}$$

This integral is a little gnarly, but we can evaluate it. We end up getting

$$g(\mu\mid\mathbf{x}) \propto \left(1+\frac{(\mu-\bar{x})^2}{r^2}\right)^{-\frac{n}{2}} \propto \left(\sum_{i=1}^{n}(x_i-\mu)^2\right)^{-\frac{n}{2}}. \tag{2.21}$$

I have written the expression in two equivalent forms because it is sometimes more convenient to use one or the other. They are proportional, which you can verify for yourself. For now, we'll use the first expression, since it is convenient for computing the marginalized posteriors. We can integrate this to get the normalization constant, giving

$$g(\mu\mid\mathbf{x}) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{n-1}{2}\right)}\,\frac{1}{r}\left(1+\frac{(\mu-\bar{x})^2}{r^2}\right)^{-\frac{n}{2}}. \tag{2.22}$$

The normalization contains gamma functions. This distribution has a name. It is the **Student-t** distribution, albeit with a nonstandard parametrization. As we now know, it describes the mean of a Gaussian distribution with unknown variance from which the data were drawn. As written, the Student-t distribution above is said to have $n-1$ degrees of freedom.

As we have already determined, the most probable value of μ is $\bar{x}$. We would like to describe an error bar[7] for this parameter μ. Since we know its posterior, the error

[7] I'm using the term "error bar" loosely here. We will sharpen this definition later in the course.

bar is just some summary of the posterior distribution. We could report the error bar to contain the set of values of μ, centered on $\bar{x}$, that contain a given percentage of the probability.

The common practice for getting the error bar is to approximate the posterior distribution as Gaussian and report intervals based on the standard deviation of the Gaussian approximation. To get a Gaussian approximation, we expand the logarithm of posterior probability distribution function in a Taylor series about its maximum.

$$\ln g(\mu \mid \mathbf{x}) = \text{constant} - \frac{n}{2} \ln \left(1 + \frac{(\mu - \bar{x})^2}{r^2} \right) \tag{2.23}$$

$$\approx \text{constant} - \frac{n(\mu - \bar{x})^2}{2r^2}. \tag{2.24}$$

Exponentiating and evaluating the normalization constant yields

$$g(\mu \mid \mathbf{x}) \approx \frac{1}{\sqrt{2\pi r^2/n}} \exp \left[-\frac{(\mu - \bar{x})^2}{2r^2/n} \right], \tag{2.25}$$

a Gaussian distribution with mean $\bar{x}$ and variance r^2/n. Recall that r^2 is the sample variance, so the variance of the Gaussian approximation of the posterior distribution is the sample variance divided by n. The quantity $r/\sqrt{n}$ is referred to as the **standard error of the mean**, which is often how error bars are reported. We now know that it describes the width of the (Gaussian approximation of the) posterior distribution describing the parameter value we sought to measure.

2.7.2 The variance σ^2

Often overlooked is an estimate for the variance. Remember, when we took measurements, we did not assume we knew the variance of the measurements. We would also like an estimate of it.

We take a similar approach. We marginalize the full posterior over μ.

$$g(\sigma \mid \mathbf{x}) = \int_{-\infty}^{\infty} \mathrm{d}\mu \; g(\mu, \sigma \mid \mathbf{x}). \tag{2.26}$$

The integral is again doable, but also again a bit gnarly. The result is

$$g(\sigma \mid \mathbf{x}) = \frac{c}{\sigma^n} \exp \left[-\frac{nr^2}{2\sigma^2} \right]. \tag{2.27}$$

We can compute the normalization constant, which involves a little messy integration, giving

$$g(\sigma \mid \mathbf{x}) = \frac{\left(nr^2\right)^{(n-1)/2}}{2^{(n-3)/2}\Gamma\left(\frac{n-1}{2}\right)\sigma^n} \exp \left[-\frac{nr^2}{2\sigma^2} \right]. \tag{2.28}$$

We can find the most probable σ (note that the normalization constant is not necessary for this calculation). This is found by finding the value of σ for which the derivative of the log posterior is zero.

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} \ln g(\sigma \mid \mathbf{x}) = \frac{\mathrm{d}}{\mathrm{d}\sigma}\left(-n \ln \sigma - \frac{nr^2}{2\sigma^2}\right) = -\frac{n}{\sigma} + \frac{nr^2}{\sigma^3}. \tag{2.29}$$

This is zero when $\sigma^2 = r^2$, or

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{2.30}$$

We can also compute a confidence interval on the parameter σ. Note, though, that its distribution, $g(\sigma \mid \mathbf{x})$, is not symmetric, as seen in Fig. 2.

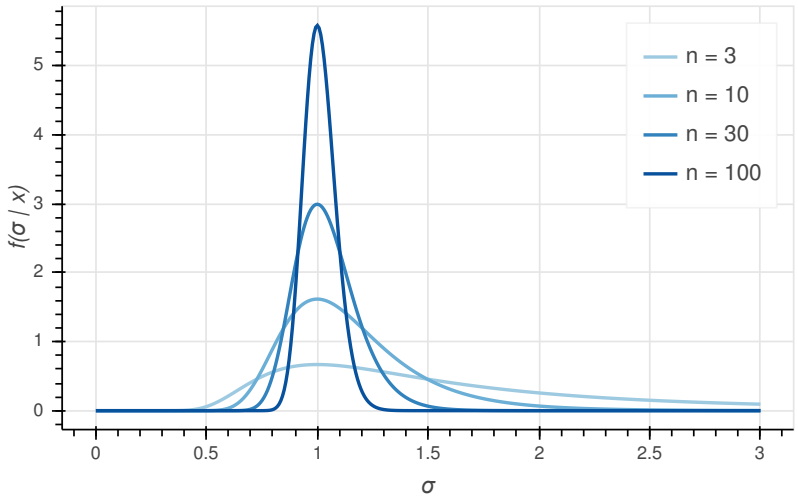

Figure 2: The posterior distribution of σ with $r = 1$ for various values of n. It becomes more symmetric as n grows.

Given that the distribution is not symmetric, we might want to provide a point estimate for σ using expectation values, instead of finding the most probable value. The integrals are nasty, but can be evaluated.

$$\langle \sigma \rangle = \int_0^\infty \mathrm{d}\sigma\ \sigma\, g(\sigma \mid \mathbf{x}) = \frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sqrt{\frac{n}{2}}\, r. \tag{2.31}$$

Alternatively, we could compute the expectation value for σ^2,

$$\langle \sigma^2 \rangle = \int_0^\infty \mathrm{d}\sigma\ \sigma^2\, g(\sigma \mid \mathbf{x}) = \frac{n}{n-1}\, r^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \tag{2.32}$$

which may be familiar to you as the so-called sample variance, or the unbiased estimate of the variance. Really, by choosing to report the most probable value of σ, the $\langle \sigma \rangle$, or $\sqrt{\langle \sigma^2 \rangle}$, we are just choosing one property of $g(\sigma \mid \mathbf{x})$ to report. We actually know the whole distribution, though, so whatever we choose is just a summary of it. These summaries are nevertheless useful, since they can concisely describe the posterior. For a Gaussian example like this, everything is nicely behaved. As we will later see, computing summary statistics without investigating the whole posterior can be a risky enterprise, and not advised.