BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2015

2 Parameter estimation

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.

2.1 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 μ in multiple independent experiments. This could be beak depths of finches, fluorescence intensity on a cell, dissociation constant for two bound proteins, etc. The possibilities abound. You can have whatever your favorite measurement is in mind during this analysis.

Our measurements of this parameter are $D = \{x_1, x_2, \ldots x_n\}$. Our "hypothesis" in this case, is the value of the parameter μ. So, we wish to calculate $P(\mu \mid D, I)$, 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.

$$P(\mu \mid D, I) = \frac{P(D \mid \mu, I)\, P(\mu \mid I)}{P(D \mid I)}. \tag{2.1}$$

Since the evidence, $P(D \mid I)$ 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 $P(D \mid \mu, I)$ and the prior $P(\mu \mid I)$.

2.2 The likelihood

To specify the likelihood, we have to ask how 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

$$P(D \mid \mu, I) = \prod_{i \in D} \delta(x_i - \mu), \tag{2.2}$$

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 our "model 3" definition from last lecture. It is the probability distribution that describes how the data relate to the parameter we are trying to measure. In the next lecture, we will learn more about probability distributions, but for now, we will introduce one useful distribution to use in our analyses.

2.3 The Gaussian distribution

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

$$P(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{2.3}$$

The parameter μ is called the mean of the distribution and σ^2 is the variance, with σ being called the standard deviation.

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.

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.[4]

More generally, the multi-dimensional Gaussian distribution for $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ is

$$P(\mathbf{x} \mid \mu, \boldsymbol{\sigma}) = (2\pi)^{-\frac{n}{2}} \left(\det \boldsymbol{\sigma}^2\right)^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \cdot (\boldsymbol{\sigma}^2)^{-1} \cdot (\mathbf{x}-\mu)\right\}, \tag{2.4}$$

where $\boldsymbol{\sigma}^2$ is a symmetric positive definite matrix called the **covariance matrix**. If off-diagonal entry $(\boldsymbol{\sigma}^2)_{ij}$ 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

$$P(\mathbf{x} \mid \mu, \boldsymbol{\sigma}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma_i^2}\right\}, \tag{2.5}$$

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,

[4] It is also the **maximal entropy distribution** for a system with well-defined first and second moments, but we will not talk about entropy in this class.

the multi-dimensional Gaussian reduces to

$$P(\mathbf{x} \mid \mu, \sigma) = \left(\frac{1}{2\pi\sigma^2}\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2\right\}. \tag{2.6}$$

2.4 The likelihood revisited: and another parameter

The Gaussian distribution is a good choice for our likelihood for repeated measurements, and we will use it for the likelihood for our present problem of estimating a parameter form 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 independent and identically distributed (i.i.d.), so that only a single variance, σ^2, is specified. So, we choose our likelihood to be

$$P(D \mid \mu, \sigma, I) = \left(\frac{1}{2\pi\sigma^2}\right)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i\in D}(x_i - \mu)^2\right\}, \tag{2.7}$$

with $n = |D|$. 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 model (model 3) has *two* parameters, and Bayes's theorem now reads

$$P(\mu, \sigma \mid D, I) = \frac{P(D \mid \mu, \sigma, I)\, P(\mu, \sigma \mid I)}{P(D \mid I)}. \tag{2.8}$$

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

$$P(\mu \mid D, I) = \int_0^\infty \mathrm{d}\sigma\, P(\mu, \sigma \mid D, I). \tag{2.9}$$

2.5 Choice of prior

Because the evidence $P(D \mid I)$ 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 $P(\mu, \sigma \mid I)$. 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. But what if we known nothing?

One assumption we could make is that μ and σ are independent. In this case,

$$P(\mu, \sigma \mid I) = P(\mu \mid I)\, P(\sigma \mid I). \tag{2.10}$$

Now, we have to decide on prior probabilities for μ and σ. Our goal is to choose *uninformative priors*, i.e., we want to claim maximal ignorance in our choice of prior probability when we have to prior information.

For μ, we will choose a uniform prior, meaning that all values are equally likely.

$$P(\mu \mid I) = \begin{cases} (\mu_{\max} - \mu_{\min})^{-1} & \mu_{\min} < \mu < \mu_{\max}, \\ \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

We have put bounds on the allowed values of μ. You may argue that this is informative; we have chosen bounds. We intentionally choose the bounds to be so far away from the peak of the likelihood that the posterior is vanishingly small at the bounds. Thus, the parameters $\mu_{\min}$ and $\mu_{\max}$ are arbitrary and serve only to ensure the $P(\mu \mid I)$ is a proper probability distribution.[5]

For the parameter σ, the choice is not quite so simple. First, we know that $\sigma > 0$ by construction. We also note that we could have chosen to parametrize the Gaussian distribution with $\xi \equiv \sigma^{-1}$ instead of σ. We want there to be no bias as a result of this choice. In other words, we want

$$P(\sigma \mid I) = P(\xi \mid I) \left| \frac{\mathrm{d}\xi}{\mathrm{d}\sigma} \right|, \tag{2.12}$$

where we have employed the change of variables formula.[6] We will choose

$$P(\sigma \mid I) = \begin{cases} \ln(\sigma_{\max}/\sigma_{\min})\, \sigma^{-1} & \sigma_{\min} < \sigma < \sigma_{\max} \\ \\ 0 & \text{otherwise,} \end{cases} \tag{2.13}$$

and show that the condition given in equation (2.12) hold. For $\sigma_{\max}^{-1} < \xi < \sigma_{\min}^{-1}$, we have

$$P(\xi \mid I) \left| \frac{\mathrm{d}\xi}{\mathrm{d}\sigma} \right| = \ln(\sigma_{\max}/\sigma_{\min})\, \xi^{-1}\, \sigma^{-2} = \ln(\sigma_{\max}/\sigma_{\min})\, \sigma^{-1} = P(\sigma \mid I), \tag{2.14}$$

so the condition holds. It holds trivially outside of the bounds. This prior, $P(\sigma \mid I) \propto \sigma^{-1}$, is called a **Jeffreys prior**. It is an uninformative prior to use for **scale parameters** like σ, which can equivalently be represented by their reciprocals. Like the prior for μ, the bounds $\sigma_{\min}$ and $\sigma_{\max}$ can be chosen to be sufficiently large/small such that they are immaterial.

[5] There is really no problem with it not being proper. In fact, improper priors are commonly used.

[6] Remember your calculus: $f(x) = f(y)\, |\mathrm{d}y/\mathrm{d}x|$.

An aside about the Jeffreys prior. Again using the change of variables formula, we see that

$$P(\sigma \mid I) = P(\ln \sigma \mid I) \left| \frac{\mathrm{d} \ln \sigma}{\mathrm{d}\sigma} \right| = \frac{1}{\sigma} P(\ln \sigma \mid I). \tag{2.15}$$

Thus,

$$P(\ln \sigma \mid I) = \begin{cases} (\ln \sigma_{\max} - \ln \sigma_{\min})^{-1} & \ln \sigma_{\min} < \ln \sigma < \ln \sigma_{\max}, \\ \\ 0 & \text{otherwise.} \end{cases} \tag{2.16}$$

So, if a parameter has a Jeffreys prior, its logarithm has a uniform prior. This can be convenient when defining parameters and performing optimizations.

2.6 The posterior

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

$$P(\mu, \sigma \mid D, I) = \frac{c}{\sigma^{n+1}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i \in D} (x_i - \mu)^2 \right\}, \tag{2.17}$$

where we have absorbed all constants in to the normalization constant c. We could do this because we chose the bounds on the priors of μ and σ to be sufficiently far away from the peak of the likelihood that they contribute negligibly to the posterior. This also allows us to extend our bounds of integration to infinity when integrating.

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

$$\sum_{i \in D} (x_i - \mu)^2 = n(\bar{x} - \mu)^2 + nr^2, \tag{2.18}$$

where

$$r^2 = \frac{1}{n} \sum_{i \in D} (x_i - \bar{x})^2 \tag{2.19}$$

is the sample variance and

$$\bar{x} = \frac{1}{n} \sum_{i \in D} x_i. \tag{2.20}$$

Thus, we have

$$P(\mu, \sigma \mid D, I) = \frac{c\, \mathrm{e}^{-nr^2/2\sigma^2}}{\sigma^{n+1}} \exp\left\{ -\frac{n(\mu - \bar{x})^2}{2\sigma^2} \right\}. \tag{2.21}$$

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 most probable $\mu = \bar{x}$, instead of a plot.

2.6.1 The mean μ

We wanted to get $P(\mu \mid D, I)$ in the first place. As we said before, we can get that by marginalizing over σ.

$$\begin{aligned} P(\mu \mid D, I) &= \int_0^\infty \mathrm{d}\sigma\, P(\mu, \sigma \mid D, I) \\ &= c \int_0^\infty \frac{\mathrm{d}\sigma}{\sigma^{n+1}} \exp\left\{ -\frac{n(\mu - \bar{x})^2 + nr^2}{2\sigma^2} \right\}. \end{aligned} \tag{2.22}$$

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

$$P(\mu \mid D, I) \propto \left(1 + \frac{(\mu - \bar{x})^2}{r^2} \right)^{-\frac{n}{2}}. \tag{2.23}$$

We can also integrate this to get the normalization constant, giving

$$P(\mu \mid D, I) = \frac{\Gamma\left(\frac{n}{2}\right)}{\sqrt{\pi}\Gamma\left(\frac{n-1}{2}\right)} \frac{1}{r} \left(1 + \frac{(\mu - \bar{x})^2}{r^2} \right)^{-\frac{n}{2}}. \tag{2.24}$$

The normalization contains gamma functions. This distribution has a name. It is the **Student-t** distribution. As we now know, it describes the estimate of the value of samples drawn from a Gaussian distribution of unknown variance. 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 a confidence interval[7] for this parameter μ. Since we know its posterior, the confidence interval

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

is just some summary of the posterior distribution. We could report the confidence interval to contain the set of values of μ, centered on $\bar{x}$, that contain a given percentage of the probability. Indeed, I will strongly advocate for this style of credible region definitions (the Bayesian analog to confidence intervals).

The common practice for getting the confidence interval 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.

$$\ln P(\mu \mid D, I) = \text{constant} - \frac{n}{2} \ln\left(1 + \frac{(\mu - \bar{x})^2}{r^2}\right) \tag{2.25}$$

$$\approx \text{constant} - \frac{n(\mu - \bar{x})^2}{r^2}. \tag{2.26}$$

Exponentiating and evaluating the normalization constant yields

$$P(\mu \mid D, I) \approx \frac{1}{\sqrt{2\pi r^2/n}} \exp\left\{-\frac{(\mu - \bar{x})^2}{2r^2/n}\right\}, \tag{2.27}$$

a Gaussian distribution with mean $\bar{x}$ and variance r^2/n. Recall that r^2 is the sample variance, do 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.

2.6.2 The variance σ^2

Often overlooked is an estimate for the variance. Remember, when we took measurements, we did not assume we know 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.

$$P(\sigma \mid D, I) = \int_{-\infty}^{\infty} \mathrm{d}\mu \, P(\mu, \sigma \mid D, I). \tag{2.28}$$

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

$$P(\sigma \mid D, I) = \frac{c}{\sigma^n} \exp\left\{-\frac{nr^2}{2\sigma^2}\right\}. \tag{2.29}$$

We can compute the normalization constant, but it is a messy expression including some incomplete gamma functions. We will not bother. Instead, we can find the most probable σ. This is found by finding the value of σ for which the derivative of the posterior is zero.

$$\frac{\mathrm{d}}{\mathrm{d}\sigma} P(\sigma \mid D, I) = c(nr^2\sigma^{-n-3} - n\sigma^{-n-1}) \exp\left\{-\frac{nr^2}{2\sigma^2}\right\} \tag{2.30}$$

$$= cn\sigma^{-n-1}\left(\frac{r^2}{\sigma^2} - 1\right) \exp\left\{-\frac{nr^2}{2\sigma^2}\right\}. \tag{2.31}$$

This is zero when $\sigma^2 = r^2$, or the variance is given by the sample variance.

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

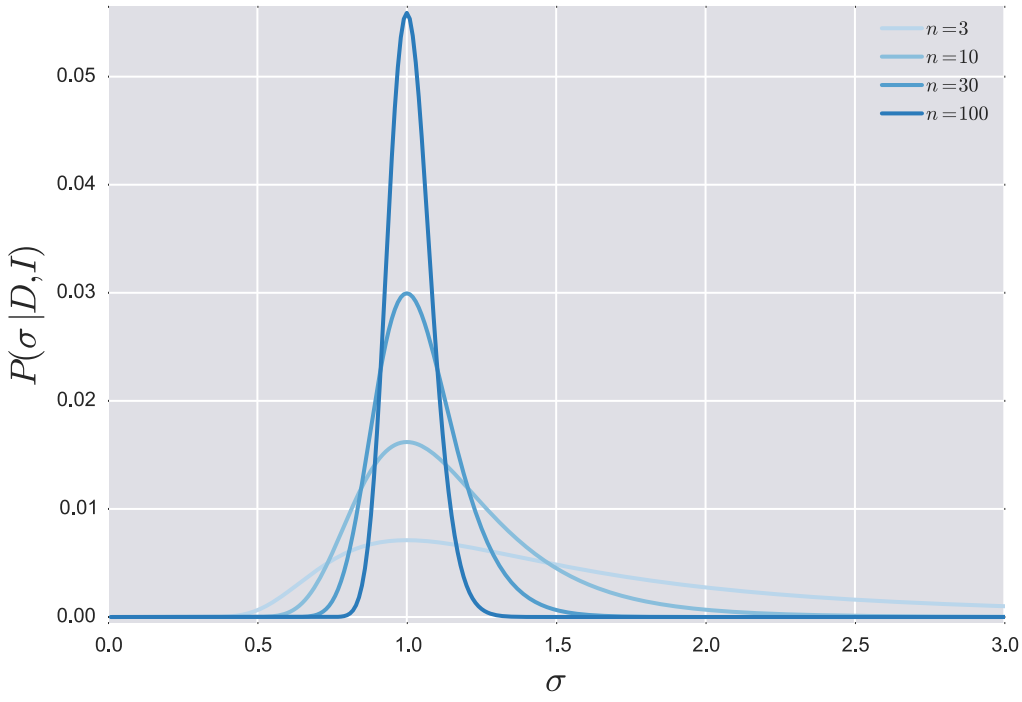

Figure 2: The posterior distribution of σ with $r = 1$ for various values of n. It becomes more symmetric as n grows.