BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2017

1 Bayes's theorem and the logic of science

We start with a question. **What is the goal of doing (biological) experiments?** There are many answers you may have for this. Some examples:

- To further knowledge.
- To test a hypothesis.
- To explore and observe.
- To demonstrate. E.g., to demonstrate feasibility.

More obnoxious answers are

- To graduate.
- Because your PI said so.
- To get data.

This question might be better addressed if we zoom out a bit and think about the scientific process as a whole. In Fig. 1, we have a sketch of the scientific processes. This cycle repeats itself as we explore nature and learn more. In the boxes are milestones, and along the arrows in orange text are the tasks that get us to these milestones.

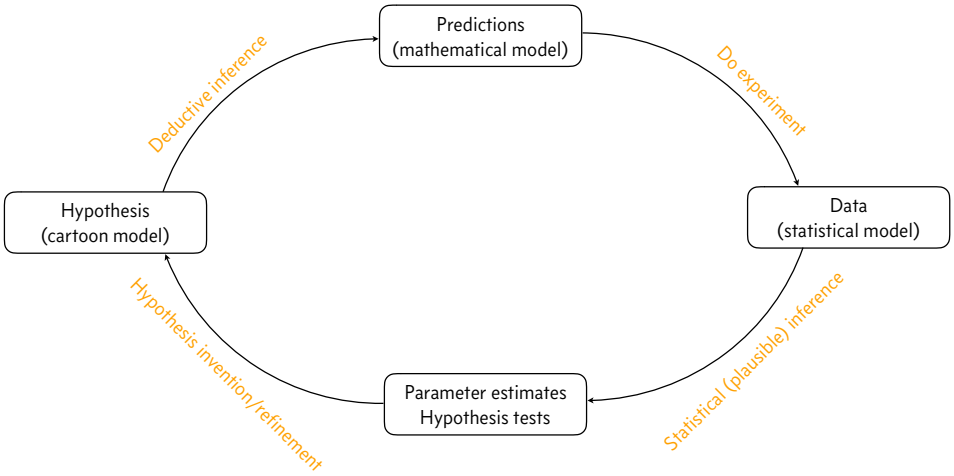

Figure 1: A sketch of the scientific process. Adapted from Fig. 1.1 of P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge, 2005.

Let's consider the tasks and their milestones. We start in the lower left.

- *Hypothesis invention/refinement.* In this stage of the scientific process, the researcher(s) think about nature, all that they have learned, including from their experiments, and formulate hypotheses or theories they can pursue with experiments. This step requires innovation, and sometimes genius (e.g., general relativity).
- *Deductive inference.* Given the hypothesis, the researchers deduce what must be true if the hypothesis is true. You have done a lot of this in your study to this point, e.g., *given X and Y, derive Z.* Logically, this requires a series of **strong syllogisms**:

  If *A* is true, then *B* is true.
  A is true.
  Therefore B is true.

  The result of deductive inference is a set of (preferably quantitative) predictions that can be tested experimentally.
- *Do experiment.* This requires *work*, and also its own kind of innovation. Specifically, you need to think carefully about how to construct your experiment to test the hypothesis. It also usually requires money. The result of doing experiments is data.
- *Statistical (plausible) inference.* This step is perhaps the least familiar to you, but *this is the step that this course is all about.* I will talk about what statistical inference is next; it's too involved for this bullet point. But the result of statistical inference is knowledge about how *plausible* a hypothesis and estimates of parameters under that hypothesis are.

1.1 What is statistical inference?

As we designed our experiment under our hypothesis, we used deductive logic to say, "If *A* is true, then *B* is true," where *A* is our hypothesis and *B* is an experimental observation. This was *deductive* inference.

Now, let's say we observe *B*. Does this make *A* true? Not necessarily. But it does make *A* more *plausible*. This is called a *weak syllogism*. As an example, consider the following hypothesis/observation pair.

> *A* = Wastewater injection after hydraulic fracturing, known as fracking, can lead to greater occurrence of earthquakes.
>
> *B* = The frequency of earthquakes in Oklahoma has increased 100 fold since 2010, when fracking became common practice there.
>
> Because *B* was observed, *A* is more plausible.

Statistical inference is the business of quantifying *how much more plausible* A is after obesrving B. In order to do statistical inference, we need a way to quantify plausibility. Probability serves this role.

So, **statistical inference requires a probability theory.** Thus, probability theory is a generalization of logic. Due to this logical connection and its crucial role in science, E. T. Jaynes says that probability is the "logic of science."

1.2 The problem of probability

We know what we need, a theory called probability to quantify plausibility. We will not formally define probability here, but use our common sense reasoning of it. Nonetheless, it is important to understand that there are two dominant *interpretations* of probability.

Frequentist probability. In the *frequentist* interpretation of probability, the probability $P(A)$ represents a long-run frequency over a large number of identical repetitions of an experiment. These repetitions can be, and often are, hypothetical. The event A is restricted to propositions about *random variables*, a quantity that can very meaningfully from experiment to experiment.[1]

Bayesian probability. Here, $P(A)$ is interpreted to directly represent the degree of belief, or plausibility, about A. So, A can be any logical proposition.

You may have heard about a split, or even a fight, between people who use Bayesian and frequentist interpretations of probability applied to statistical inference. There is no need for a fight. The two ways of approaching statistical inference differ in their interpretation of probability, the tool we use to quantify plausibility. Both are valid.

In my opinion, the Bayesian interpretation of probability is more intuitive to apply to scientific inference. It always starts with a simple probabilistic expression and proceeds to quantify plausibility. It is conceptually cleaner to me, since we can talk about plausibility of anything, including parameter values. In other words, Bayesian probability serves to quantify our own knowledge, or degree of certainty, about a hypothesis or parameter value. Conversely, in frequentist statistical inference, the parameter values are fixed, and we can only study how repeated experiments will convert the real parameter value to an observed real number.

We will use some frequentist approaches in class, especially when we study *nonparametric* methods, but we will generally focus on Bayesian analysis. For now, we will focus on some key properties of probability.

[1] More formally, a random variable transforms the possible outcomes of an experiment to real numbers.

1.3 Desiderata for Bayesian probability

In 1946, R. Cox laid out a pair of rules based on some desired properties of probability as a quantifier of plausibility. These ideas were expanded on by E. T. Jaynes in the 1970s. The *desiderata* are

I. Probability is represented by real numbers.

II. Probability must agree with rationality. As more information is supplied, probability must rise in a continuous, monotonic manner. The deductive limit must be obtained where appropriate.

III. Probability must be consistent.

   a) Structure consistency: If a result is reasoned in more than one way, we should get the same result.

   b) Propriety: All relevant information must be considered.

   c) Jaynes consistency: Equivalent states of knowledge must be represented by equivalent probability.

Based on these desiderata, we can work out important results that a probability function must satisfy. I pause to note that one can generally define probability without a specific *interpretation* in mind, and it is valid for both Bayesian and frequentist interpretations. See, for example, section 1.6 of Blitzstein and Hwang, *Introduction to Probability*, CRC Press, 2015.

Two results of these desiderata (worked out in chapter 2 of Gregory's book) are the *sum rule* and the *product rule*. They apply to both frequentist and Bayesian interpretations.

1.4 The sum rule, the product rule, and conditional probability

The *sum rule* says that the probability of all events must add to unity. Let $\bar{A}$ be all events *except* A. Then, the sum rule states that

$$P(A) + P(\bar{A}) = 1. \tag{1.1}$$

Now, let's say that we are interested in event A happening *given* that event B happened. So, A is *conditional* on B. We denote this conditional probability as $P(A \mid B)$. Given this notion of conditional probability, we can write the sum rule as

$$\textbf{(sum rule)} \qquad P(A \mid B) + P(\bar{A} \mid B) = 1, \tag{1.2}$$

for any B.

The *product rule* states that

$$P(A, B) = P(A \mid B)\, P(B), \tag{1.3}$$

where $P(A, B)$ is the probability of both A *and* B happening. The product rule is also referred to as the definition of conditional probability. It can similarly be expanded as we did with the sum rule.

$$\textbf{(product rule)} \qquad P(A, B \mid C) = P(A \mid B, C)\, P(B \mid C), \tag{1.4}$$

for any C.

1.5 Application to scientific measurement

This is all a bit abstract. Let's bring it into the realm of scientific experiment. We'll assign meanings to these things we have been calling A, B, and C.

$$A = \text{hypothesis (or parameter value)}, H_i, \tag{1.5}$$

$$B = \text{Measured data set}, D, \tag{1.6}$$

$$C = \text{All other information we know}, I. \tag{1.7}$$

Now, let's rewrite the product rule.

$$P(H_i, D \mid I) = P(H_i \mid D, I)\, P(D \mid I). \tag{1.8}$$

Ahoy! The quantity $P(H_i \mid D, I)$ is exactly what we want from our statistical inference. This is the probability that a hypothesis is true, or a probability density function (or probability mass function in the discrete case) for the values of a parameter, given measured data and everything we've learned. Now, how do we compute it?

1.6 Bayes's Theorem

Note that because "and" is commutative, $P(H_i, D \mid I) = P(D, H_i \mid I)$. So, we apply the product rule to both sides of the seemingly trivial equality.

$$P(H_i \mid D, I)\, P(D \mid I) = P(H_i, D \mid I) = P(D, H_i \mid I) = P(D \mid H_i, I)\, P(H_i \mid I). \tag{1.9}$$

If we take the terms at the beginning and end of this equality and rearrange, we get

$$\textbf{(Bayes's theorem)} \qquad P(H_i \mid D, I) = \frac{P(D \mid H_i, I)\, P(H_i \mid I)}{P(D \mid I)}. \tag{1.10}$$

This result is called **Bayes's theorem**. This is far more instructive in terms of how to compute our goal, which is the left hand side.[2] The quantities on the right hand side all have meaning. We will talk about the meaning of each term in turn, and this is easier to do using their names; each item in Bayes's theorem has a name.

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}. \tag{1.11}$$

The prior probability. First, consider the prior, $P(H_i \mid I)$. As probability is a measure of plausibility, or how believable a hypothesis is, we should be able to write this down based on I.[3] This represents the plausibility about hypothesis H_i given everything we know *before* we did the experiment to get the data.

The likelihood. The likelihood, $P(D \mid H_i, I)$, describes how likely it is to acquire the observed data, *given that the hypothesis H_i is true*. It also contains information about what we expect from the data, given our measurement method. Is there noise in the instruments we are using? How do we model that noise? These are contained in the likelihood.

The evidence. I will not talk much about this here, except to say that it can be computed from the likelihood and prior, and is also called the *marginal likelihood*, a name whose meaning will become clear in the next section.[4]

The posterior probability. This is what we are after. How plausible is the hypothesis, given that we have measured some new data? It is calculated directly from the likelihood and prior (since the evidence is also computed from them). Computing the posterior distribution constitutes the bulk of our inference tasks in this course.

[2]Do not be confused. Bayes's Theorem is a statement about probability and holds whether you interpret probability in a Bayesian or frequentist manner. The name "Bayesian" does not mean that it applies only to probability interpreted through the Bayesian lens.

[3]I say this flippantly. In fact, specifying prior probabilities is one of the most studied and most controversial aspects of Bayesian statistics.

[4]I have heard this referred to as the "fully marginalized likelihood" because of the cute correspondence of the acronym and how some people feel trying to get their head around the meaning of the quantity.

1.7 Marginalization

A moment ago, I mentioned that the evidence can be computed from the likelihood and the prior. To see this, we apply the sum rule to the posterior probability.

$$\begin{aligned} 1 &= P(H_j \mid D, I) + P(\bar{H}_j | D, I) \\ &= P(H_j \mid D, I) + \sum_{i \neq j} P(H_i \mid D, I) \\ &= \sum_i P(H_i \mid D, I), \end{aligned} \tag{1.12}$$

for some hypothesis H_j. Now, Bayes's theorem gives us an expression for $P(H_i \mid D, I)$, so we can compute the sum.

$$\begin{aligned} \sum_i P(H_i \mid D, I) &= \sum_i \frac{P(D \mid H_i, I)\, P(H_i \mid I)}{P(D \mid I)} \\ &= \frac{1}{P(D \mid I)} \sum_i P(D \mid H_i, I)\, P(H_i \mid I) \\ &= 1. \end{aligned} \tag{1.13}$$

Therefore, we can compute the evidence by summing over the priors and likelihoods of all possible hypotheses.

$$P(D \mid I) = \sum_i P(D \mid H_i, I)\, P(H_i \mid I). \tag{1.14}$$

This process of eliminating a variable (in this case the hypotheses) from a probability by summing is called *marginalization*.

Note that if the space of hypotheses is continuous, for example if the "hypothesis" is a parameter value which we'll call θ, we can replace the summation with an integral.[5]

$$P(D \mid I) = \int \mathrm{d}\theta\, P(D \mid \theta, I)\, P(\theta \mid I). \tag{1.15}$$

1.8 A note on the word "model"

You may have noticed the terms "cartoon model," "mathematical model," and "statistical model" in Fig. 1. Being biologists who are doing data analysis, the word

[5] There are some mathematical subtleties. These are discussed at length in Jaynes's book, *Probability Theory: the logic of science*.

"model" is used to mean three different things in our work. So, for the purposes of this course, we need to clearly define what we are talking about when we use the word "model."

Cartoon model. These models are the typical cartoons we see in text books or in discussion sections of biological papers. They are a sketch of what we think might be happening in a system of interest, but they do not provide quantifiable predictions.

Mathematical model. These models give quantifiable predictions that must be true if the hypothesis (which is sketched as a cartoon model) is true. In many cases, getting to predictions from a hypothesis is easy. For example, if I hypothesize that protein A binds protein B, a quantifiable prediction would be that they are colocalized when I image them. However, sometimes harder work and deeper thought is needed to generate quantitative predictions. This often requires "mathematizing" the cartoon. This is how a mathematical model is derived from a cartoon model. Oftentimes when biological physicists refer to a "model," they are talking about what we are calling a mathematical model.

Statistical model. Essentially, a statistical model specifies the likelihood and prior. Statisticians often use the word "model" in this context. As a simple example, consider the measurement of the length of a *C. elegans* eggs. A plausible statistical model would be that the egg lengths are Gaussian distributed (and therefore are described by a mean and a standard deviation). The statistical model can include any mathematization of cartoons we did to generate a mathematical model, and can also contain any information about any possible effects we might see in a measurement.

1.9 Bayes's theorem as a model for learning

We will close today's lecture with a discussion of Bayes's theorem as as model for learning. Let's say we did an experiment and got data set D_1 as an investigation of hypothesis H. Then, our posterior distribution is

$$P(H \mid D_1, I) = \frac{P(D_1 \mid H, I)\, P(H \mid I)}{P(D_1 \mid I)}. \tag{1.16}$$

Now, let's say we did another experiment and got data D_2. We already know D_1 ahead of this experiment, do our prior is $P(H \mid D_1, I)$, which is the posterior from the first experiment. So, we have

$$P(H \mid D_1, D_2, I) = \frac{P(D_2 \mid D_1, H, I)\, P(H \mid D_1, I)}{P(D_2 \mid D_1, I)}. \tag{1.17}$$

Now, we plug in Bayes's theorem applied to our first data set, equation (1.16), giving

$$P(H \mid D_1, D_2, I) = \frac{P(D_2 \mid D_1, H, I)\, P(D_1 \mid H, I)\, P(H \mid I)}{P(D_2 \mid D_1, I)\, P(D_1 \mid I)}. \tag{1.18}$$

By the product rule, the denominator is $P(D_1, D_2 \mid I)$. Also by the product rule,

$$P(D_2 \mid D_1, H, I)\, P(D_1 \mid H, I) = P(D_1, D_2 \mid H, I). \tag{1.19}$$

Inserting these expressions into equation (1.18) yields

$$P(H \mid D_1, D_2, I) = \frac{P(D_1, D_2 \mid H, I)\, P(H \mid I)}{P(D_1, D_2 \mid I)}. \tag{1.20}$$

So, acquiring more data gave us more information about our hypothesis in that same way as if we just combined D_1 and D_2 into a single data set. So, acquisition of more and more data serves to help us learn more and more about our hypothesis or parameter value.

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.

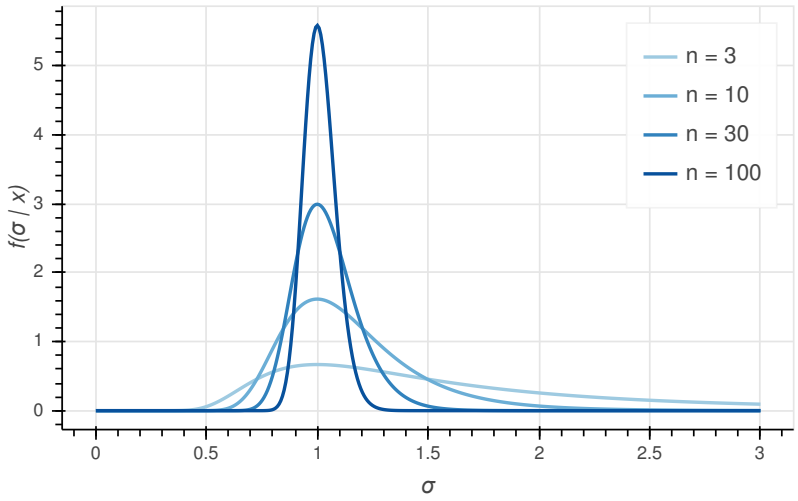

Figure 2: The posterior distribution of σ with $r = 1$ for various values of n. It becomes more symmetric as n grows.

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.

3 Constructing Bayesian models

In the last lecture, we saw how to perform parameter estimation for repeated measurements with a Gaussian likelihood and prior that goes like the inverse of the standard deviation of the Gaussian. Most of last lecture was then finding ways to summarize the posterior. We saw, and this is generally true, that we need only to specify the likelihood and prior to build the statistical model. In this lecture, we will discuss ways to build a statistical model. We will do this using two examples, learning general principles as we work through them.

3.1 Example 1: Mitotic spindle size

Matt Good and coworkers (Good, et al., *Science*, **342**, 856–860, 2013) developed a microfluidic device where they could create droplets of cytoplasm extracted from *Xenopus* eggs and embryos (see Fig. 3). A remarkable property about *Xenopus* extract is that mitotic spindles spontaneously form; the extracted cytoplasm has all the ingredients to form them. This makes it an excellent model system for studying spindles. With their device, Good and his colleagues were able to study how the size of the cell affects the dimensions of the mitotic spindle; a simple, yet beautiful, question. The experiment is conceptually simple; they made the droplets and then measured their dimensions and the dimensions of the spindles using microscope images.

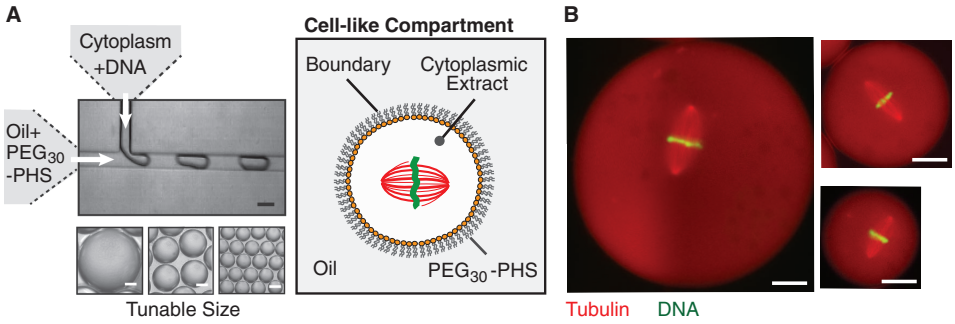

Figure 3: Schematic of spindle size experiment. Scale bars are 20 µm. Taken from Fig. 1 of Good, et al., *Science*, **342**, 856–860, 2013.

The question the authors were after was about how the spindle size scaled with the diameter of the droplet. The data they acquired are shown in Fig. 4.

3.1.1 The cartoon model

Recall in lecture 1 that we went through the process of developing a statistical model, starting with a cartoon model, mathematizing it, and then making a statistical model

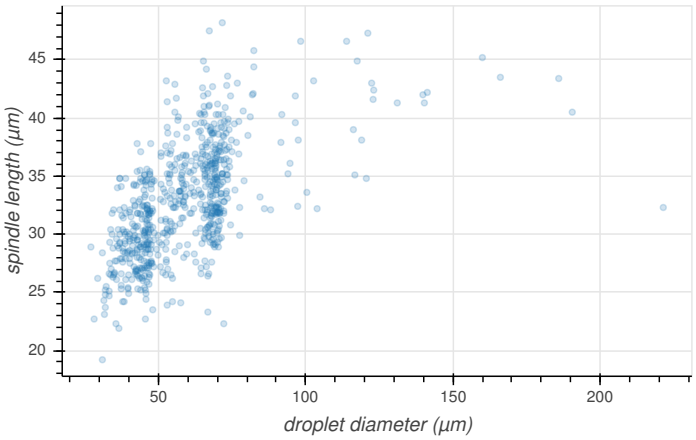

Figure 4: Spindle length versus droplet diameter.

describing how the data might vary from the mathematical model due to naturally occurring variability and that in the experiments. Good and coworkers hypothesized that the length of spindles is regulated by the total amount of tubulin available to make them. Specifically, the three key principles of their "cartoon" model are:

1. The total amount of tubulin in the droplet or cell is conserved.
2. The total length of polymerized microtubules is a function of the total tubulin concentration after assembly of the spindle. This results from the balances of microtubule polymerization rate with catastrophe frequencies.
3. The density of tubulin in the spindle is independent of droplet or cell volume.

3.1.2 The mathematical model

From these principles, we need to derive a mathematical model that will provide us with testable predictions. The derivation follows, and you may read it if you are interested. Since our main focus here is building a statistical model, you can skip ahead to equation (3.14), where we define a mathematical expression relating the spindle length, l to the droplet diameter, d, which depends on two parameters, γ and ϕ.

Principle 1 above (conservation of tubulin) implies

$$T_0 V_0 = T_1(V_0 - V_s) + T_s V_s, \tag{3.1}$$

where V_0 is the volume of the droplet or cell, V_s is the volume of the spindle, T_0 is the total tubulin concentration (polymerized or not), T_1 is the tubulin concentration in the cytoplasm after the the spindle has formed, and T_s is the concentration of tubulin in the spindle. If we assume the spindle does not take up much of the total volume

of the droplet or cell ($V_0 \gg V_s$, which is the case as we will see when we look at the data), we have

$$T_1 \approx T_0 - \frac{V_s}{V_0} T_s. \tag{3.2}$$

The amount of tubulin in the spindle can we written in terms of the total length of polymerized microtubules, L_{MT} as

$$T_s V_s = \alpha L_{MT}, \tag{3.3}$$

where α is the tubulin concentration per unit microtubule length. (We will see that it is unimportant, but from the known geometry of microtubules, $\alpha \approx 2.7$ nmol/μm.)

We now formalize assumption 2 into a mathematical expression. Microtubule length should grow with increasing T_1. There should also be a minimal threshold T_{min} where polymerization stops. We therefore approximate the total microtubule length as a linear function,

$$L_{MT} \approx \begin{cases} 0 & T_1 \leq T_{min} \\ \beta(T_1 - T_{min}) & T_1 > T_{min}. \end{cases} \tag{3.4}$$

Because spindles form in *Xenopus* extract, $T_0 > T_{min}$, so there exists a T_1 with $T_{min} < T_1 < T_0$. Thus, going forward, we are assured that $T_1 > T_{min}$. So, we have

$$V_s \approx \alpha \beta \frac{T_1 - T_{min}}{T_s}. \tag{3.5}$$

With insertion of our expression for T_1, this becomes

$$V_s \approx \alpha \beta \left(\frac{T_0 - T_{min}}{T_s} - \frac{V_s}{V_0} \right). \tag{3.6}$$

Solving for V_s, we have

$$V_s \approx \frac{\alpha \beta}{1 + \alpha \beta / V_0} \frac{T_0 - T_{min}}{T_s} = \frac{V_0}{1 + V_0 / \alpha \beta} \frac{T_0 - T_{min}}{T_s}. \tag{3.7}$$

We approximate the shape of the spindle as a prolate spheroid with major axis length l and minor axis length w, giving

$$V_s = \frac{\pi}{6} l w^2 = \frac{\pi}{6} k^2 l^3, \tag{3.8}$$

where $k \equiv w/l$ is the aspect ratio of the spindle. We can now write an expression for the spindle length as

$$l \approx \left(\frac{6}{\pi k^2} \frac{T_0 - T_{min}}{T_s} \frac{V_0}{1 + V_0 / \alpha \beta} \right)^{\frac{1}{3}}. \tag{3.9}$$

For small droplets, with $V_0 \ll \alpha\beta$, this becomes

$$l \approx \left(\frac{6}{\pi k^2} \frac{T_0 - T_{\min}}{T_s} V_0 \right)^{\frac{1}{3}} = \left(\frac{T_0 - T_{\min}}{k^2 T_s} \right)^{\frac{1}{3}} d, \tag{3.10}$$

where d is the diameter of the spherical droplet or cell. So, we expect the spindle size to increase linearly with the droplet diameter for small droplets.

For large V_0, the spindle size becomes independent of droplet size;

$$l \approx \left(\frac{6\alpha\beta}{\pi k^2} \frac{T_0 - T_{\min}}{T_s} \right)^{\frac{1}{3}}. \tag{3.11}$$

We can define two parameters to describe the data,

$$\gamma = \left(\frac{T_0 - T_{\min}}{k^2 T_s} \right)^{\frac{1}{3}} \tag{3.12}$$

$$\phi = \left(\frac{6\alpha\beta}{\pi} \right)^{\frac{1}{3}}. \tag{3.13}$$

We assume that γ and ϕ are the same for all data. We can rewrite the general model expression in terms of these parameters as

$$l(d; \gamma, \phi) \approx \frac{\gamma d}{(1 + (d/\phi)^3)^{\frac{1}{3}}}. \tag{3.14}$$

For small and large droplets, respectively, we have

$$l \approx \gamma d \quad \text{for } d/\phi \ll 1, \tag{3.15}$$

$$l \approx \gamma \phi \quad \text{for } d/\phi \gg 1. \tag{3.16}$$

Note that the expression for the linear regime gives bounds for γ. Obviously, $\gamma > 0$. Because $l \leq d$, lest the spindle not fit in the droplet, we also have $\gamma \leq 1$. The parameter ϕ is independent of the system geometry, so it only has the physical lower bound of $\phi > 0$.

3.1.3 A comment on the model parameters

We went through some algebraic manipulations to get our mathematical model in a form with two parameters. We want to try to identify *independent* parameters in your mathematical before doing regression analysis. In a trivial example, imagine someone proposed the following model to use in a regression on (x, y) data:

$$y = ax + bx + c. \tag{3.17}$$

Obviously, it would be silly to have both a and b as regression parameters, and we should instead define a new parameter $d = a + b$ and use that as a regression parameter. In the case of spindle length, we had parameters T_0, $T_{\min}$, T_s, k, α, and β, but, as we saw, we can only resolve two parameters, γ and ϕ. Furthermore, if we happen to be in the linear regime, ϕ does not enter the expressions, so we obviously cannot resolve it. Similarly, we can only determine ϕ if we are in the plateau regime.

3.1.4 The statistical model: The likelihood

We have a mathematical model, so now we are left to specify the likelihood and prior. We will start with the likelihood. The data are pairs of droplet diameters and spindle lengths. We denote one such pair as (d_i, l_i), and the whole data set as $D = \mathbf{d}, \mathbf{l}$. The parameters are $\theta = \gamma, \phi$. So, the likelihood is $f(D \mid \theta) = f(\mathbf{d}, \mathbf{l} \mid \gamma, \phi, \theta_s)$, where θ_s are the parameters associated with the statistical model (as opposed to γ and ϕ, which are associated with the mathematical model).

We need a probabilistic model about how the observe data might vary stochastically about the mathematical model. We can write

$$l_i = l(d_i; \gamma, \phi) + e_i, \tag{3.18}$$

where e_i is how much the measured spindle length differs from the predicted length for the measured drop diameter. So, we are left to choose how e_i is distributed.

Because many processes come together to make a spindle, and then to measure its length, it is reasonable to assume that e_i is Gaussian distributed. The mean of this Gaussian should be zero, since on average, the model should fit the data. One way to write this is

$$e_i \sim \text{Norm}(0, \sigma_i). \tag{3.19}$$

This reads as, "The error e_i is Normally distributed with mean zero and standard deviation σ_i." This notation is commonly used to make a sentence like the one I just quoted more concise. We could also write out the full PDF.

$$f(e_i \mid \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \, \mathrm{e}^{-e_i^2/2\sigma_i^2}. \tag{3.20}$$

Thus, for a single data point, we have

$$f(d_i, l_i \mid \gamma, \phi, \sigma_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \, \exp\left[-\frac{(l_i - l(d_i; \gamma, \phi))^2}{2\sigma_i^2}\right]. \tag{3.21}$$

This could be equivalently written as

$$l_i \sim \text{Norm}(l(d_i; \gamma, \phi), \sigma_i). \tag{3.22}$$

Now, if each measurement is independent, the likelihoods of each data point multiply, giving

$$f(\mathbf{d}, \mathbf{l} \mid \gamma, \phi, \{\sigma\}) = \frac{1}{(2\pi)^{n/2} \prod_i \sigma_i} \exp\left[-\sum_i \frac{(l_i - l(d_i; \gamma, \phi))^2}{2\sigma_i^2} \right], \tag{3.23}$$

where n is the number of observations of d_i, l_i pairs we have and $\{\sigma\}$ represents the σ_i values. If these are all equal, we have a single σ, which gives a likelihood of

$$f(\mathbf{d}, \mathbf{l} \mid \gamma, \phi, \sigma) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_i (l_i - l(d_i; \gamma, \phi))^2 \right]. \tag{3.24}$$

This can equivalently be written as

$$l_i \sim \text{Norm}(l(d_i; \gamma, \phi), \sigma) \quad \forall i. \tag{3.25}$$

We thus have our likelihood. We have assumed that each measurement is independent of the others and that the variation from the model is **homoscedastic**, which means that the magnitude of the error of measured data from the model is the same for all data points (as opposed to **heteroscedastic**).

3.1.5 Choice of prior

We are now left to the choice of the prior. Before we embark on the journey of choosing the prior, I quote Efron and Hastie from their book, *Computer Age Statistical Inference*.

> For 200 years, however, two impediments stood between Bayesian theory's philosophical attraction and its practical application.
>
> 1 In the absence of relevant past experience, the choice of a prior distribution introduces an unwanted subjective element into scientific inference.
> 2 Bayes' rule looks simple enough, but carrying out the numerical calculation of a posterior distribution often involves intricate higher-dimensional integrals.

We will deal with the second impediment in coming weeks when we use **Markov chain Monte Carlo** to handle the intricate integrals. Our goal now is to come up with a prior distribution that avoids subjectivity. As Efron and Hastie called this process an impediment, we proceed with trepidation.

The prior encodes our knowledge about the parameters of the statistical model. In this case, we have three parameters, γ and ϕ, which entered through the physical model, and σ which entered through our modeling of the variability inherent in the system and in measurement. So, we need to specify $g(\gamma, \phi, \sigma)$.

Independence of priors. Our first step on the journey to specifying $g(\gamma, \phi, \sigma)$ is to note that these parameters should be independent of each other. The parameter γ depends only on the aspect ratio of spindles, and the total concentration of tublin in the cell, the concentration of tubulin in the cytoplasm, and the critical concentration of tubulin where microtubule growth arrests. The parameter ϕ depends on the concentration of tubulin in a single microtubule (known from the geometry of microtubules) and a constant of proportionality between microtubule length and cytoplasmic tubulin concentration. Because they depend on distinct, independent physical quantities, the parameters γ and ϕ are independent of each other. Similarly, the parameter σ describes how much the spindle length differs from the prediction. It is a bit harder to state that this is independent of γ and ϕ. However, doing so is a less egregious approximation, perhaps, than assuming homoscedasticity in the first place. So, we will proceed assuming all three parameters are independent, so

$$g(\gamma, \phi, \sigma) = g(\gamma)\, g(\phi)\, g(\sigma). \tag{3.26}$$

Uninformative priors. If we want to reduce subjectivity in our prior, we want to remain as ignorant as possible about the parameters before we see the data. However, we are not *completely* ignorant. For example, we know for sure that $0 \leq \gamma \leq 1$ based on physical arguments stated at the end of section 3.1.2. This should also be encoded in our prior, such that $g(\gamma) = 0$ for all negative γ and for all $\gamma > 1$.

If we want to avoid subjectivity, we might say, then, that any value of γ on the interval from zero to one is equally likely as any other. In this case, we have

$$g(\gamma) = \begin{cases} 1 & 0 \leq \gamma \leq 1, \\ 0 & \text{otherwise}, \end{cases} \tag{3.27}$$

or

$$\gamma \sim \text{Uniform}(0, 1), \tag{3.28}$$

which says that the prior distribution of γ is Uniform on the interval $[0, 1]$[8]. This notion of assigning equal probability to all possibilities is often referred to as **Laplace's**

[8]Strictly speaking, we should have that $\gamma > 0$, not $\gamma \geq 0$, since a zero value for γ under this model would mean that spindles always have zero length.

Principle of Insufficient Reason.

Unlike γ, ϕ has no upper bound. Yes, it must be positive, but the upper bound is not apparent. Remember, though, that the prior contains all information we know before the experiment. For example, we know that the mitotic spindle in a one-cell *Xenopus* embryo do not span the entire cell, and that the cell is about 2 mm across (huge!). So, the absolute maximum we could expect for $\gamma\,\phi$ is 2 mm. So there is a reasonable maximum we could choose.

So, we could choose a Uniform prior for ϕ on the interval of, say zero to ten mm. But is this really uninformative? John Venn and Ronald Fisher, famous attackers of a Bayesian approach, would say no. They could argue that we could equally well have chosen to to parametrize the model in terms of $\xi = \phi^{-3}$ instead so that the theoretical expression for spindle length is

$$l(d;\gamma,\phi) = \frac{\gamma d}{(1+\xi d^3)^{\frac{1}{3}}}. \tag{3.29}$$

If we chose a Uniform prior on ϕ, then the prior on ξ is no longer Uniform. Recall the **change of variables formula** from multivariate calculus.

$$g(\xi) = \left|\frac{\mathrm{d}\phi}{\mathrm{d}\xi}\right| g(\phi). \tag{3.30}$$

Taking $g(\phi)$ to be a constant (which it is for a Uniform distribution), we perform the change of variables, to get

$$g(\xi) = \left|-\frac{\xi^{-\frac{4}{3}}}{3}\right| g(\phi) = \text{constant}\cdot\,\xi^{-\frac{4}{3}}, \tag{3.31}$$

which is no longer flat. So perhaps in cases like this, a Uniform prior is not actually uninformative; we are biasing toward a certain parametrization. We desire **transformation invariance**, meaning that the prior should be the same functional dependence on ϕ if we transform the parameter in a certain way.

Generically, this means that if we have a set of parameters θ that are transformed into a new set of parameters ζ, we should choose $g(\theta)$ such that

$$\left|\frac{\partial(\zeta_1,\zeta_2,\ldots)}{\partial(\theta_1,\theta_2,\ldots)}\right| g(\zeta(\theta)) \tag{3.32}$$

has the same functional form as $g(\theta)$, up to a multiplicative constant. The first factor in this equation denotes the Jacobian, which is the absolute value of the determinant of the Jacobi matrix.

So, in the present example let's say we want our prior to be invariant if we transform ϕ to a new variable ξ such that $\xi = \phi^a$. That is, we want

$$g(\xi(\phi)) = \left|\frac{\mathrm{d}\phi}{\mathrm{d}\xi}\right| g(\phi) = a\phi^{a-1} g(\phi) \tag{3.33}$$

to have the same ϕ dependence as $g(\phi)$. If we pick $g(\phi) = c/\phi$, where c is a constant, we see that this is indeed the case.

$$g(\xi(\phi)) = \frac{ac}{\phi}, \tag{3.34}$$

which has the same ϕ-dependence.

This kind of prior, which is uninformative maintaining transformational invariance like we have just described, is a case of a **Jeffreys prior** (discussed very briefly at the end of this lecture). In fact, in portions of the literature, including in Sivia's book, such a prior, $g(\theta) \propto 1/\theta$, is just called "a Jeffreys prior." *For the purposes of this course, this is what we mean when we refer to a Jeffreys prior.*

It makes sense, then, to also parametrize σ with a Jeffreys prior, since we could also have chosen to parametrize the likelihood with $\tau = \sigma^{-1}$.

Proper and improper priors. Our prior for γ, being Uniform on the interval from zero to one, is **proper**, in the sense that it is properly normalized. If we did not have bounds on it, we would call it an **improper prior**, since it cannot be normalized. The same is true for the Jeffreys prior. If we do not define bounds for a prior of the form $g(\theta) \propto 1/\theta$, it cannot be normalized, since

$$\int_b^\infty \frac{\mathrm{d}\theta}{\theta} \tag{3.35}$$

diverges for any positive b, as does

$$\int_0^b \frac{\mathrm{d}\theta}{\theta}. \tag{3.36}$$

Usually, this is not a problem for the problem of parameter estimation, that is computing $g(\theta \mid D)$. This is because for extreme values of the parameters θ, the likelihood typically is vanishingly small. Recall, the posterior is

$$g(\theta \mid D) = \frac{f(D \mid \theta)\, g(\theta)}{\int \mathrm{d}\theta\, f(D \mid \theta)\, g(\theta)}. \tag{3.37}$$

Since $f(D \mid \theta)$ typically is tiny for extreme parameter values, it overwhelms the finite $g(\theta)$ in the numerator, and in the integral in the denominator. Furthemore, any normalization constants for $g(\theta)$ cancel out with those appearing in the denominator while computing the posterior.

While this is convenient for the parameter estimation problem, as we will see in later lectures, we do need to *exactly* compute the evidence,

$$\int \mathrm{d}\theta\, f(D \mid \theta)\, g(\theta), \tag{3.38}$$

when doing model selection. So, we should bound and normalize the priors with reasonable bounds. We can write our prior for our example problem with spindle lengths as

$$g(\gamma, \phi, \sigma) = g(\gamma)\, g(\phi)\, g(\sigma), \tag{3.39}$$

with

$$g(\gamma) = \begin{cases} 1 & 0 \leq \gamma \leq 1, \\ 0 & \text{otherwise}, \end{cases} \tag{3.40}$$

$$g(\phi) = \begin{cases} \frac{1}{\phi\, \ln(\phi_{\max}/\phi_{\min})} & \phi_{\min} \leq \phi \leq \phi_{\max}, \\ 0 & \text{otherwise}, \end{cases} \tag{3.41}$$

$$g(\sigma) = \begin{cases} \frac{1}{\sigma\, \ln(\sigma_{\max}/\sigma_{\min})} & \sigma_{\min} \leq \sigma \leq \sigma_{\max}, \\ 0 & \text{otherwise}. \end{cases} \tag{3.42}$$

Alternatively, we could write this as

$$\gamma \sim \text{Uniform}(0, 1), \tag{3.43}$$

$$\phi \sim \text{Jeffreys}(\phi_{\min}, \phi_{\max}), \tag{3.44}$$

$$\sigma \sim \text{Jeffreys}(\sigma_{\min}, \sigma_{\max}). \tag{3.45}$$

3.1.6 Choosing bounds

We saw that we could choose the bounds on γ with physical arguments. We would like to make similar choices for bounds for ϕ and σ. We already made a physical argument based on the size of *Xenopus* embryos that the maximal ϕ cannot be more than a few millimeters. Its lower bound cannot be zero because this would mean that the spindle length would always be zero. We might instead choose a lower bound to be something like 10 nanometers, about the size of a microtubule nucleus.

Choosing bounds on σ can be a bit more challenging, because it is describing variability in the experiment. We might choose an upper bound close to the maximal size of a spindle, since we would not get variation bigger than that. So, one millimeter is plenty big for an upper bound. For the lower bound, we might again choose 10 nanometers, as this is about the size of four or five tubulin diameters, which should be the smallest fluctuation we could imagine seeing.

3.1.7 Computing the posterior

Our specification of the posterior is now complete. We have specified the likelihood and prior. The evidence can be calculated by integrating the product of the likelihood and prior over all parameter values. Actually computing, plotting, and summarizing the posterior is a separate challenge. Specifically, it is impediment number 2 laid out by Efron and Hastie. This is the subject of the next few weeks of the course.

3.2 Example 2: Worm reversals

In Homework 3.3, we consider reversals upon exposure to blue light of *C. elegans* that have a Channelrhodopsin in a specific neuron. There is some probability p of reversal. Say we do n trials and observe r reversals. The likelihood is Binomially distributed according to the story of the Binomial distribution. So, Bayes theorem reads

$$g(p \mid n, r) = \frac{f(r \mid p, n)\, g(p)}{f(r \mid n)}, \tag{3.46}$$

where

$$f(r \mid p, n) = \frac{n!}{(n-r)!r!}\, p^r (1-p)^{n-r}, \tag{3.47}$$

which we could alternatively write as

$$r \mid p, n \sim \text{Binom}(n, p), \tag{3.48}$$

(Note that I wrote $g(p)$ instead of $g(p \mid n)$ because they are equal; n has no bearing on p.)

As we consider our choice of prior, $g(p)$, Think back to the first lecture when we talked about Bayes's theorem as a model for learning. The idea there was that we know something before (*a priori*) acquiring data, and then we update our knowledge after (*a posteriori*). So, we come in with the prior and out with the posterior after acquiring data. It might make sense, then, that the prior and the posterior distributions are the same. That is to say they are the same distribution, but with different parameters. The parameters get updated going from the prior to the posterior. When this is the case, the prior is said to be **conjugate** to the likelihood. This makes sense: the likelihood determines the relationship between the prior and the posterior, so it should determine the functional form of the prior/posterior such that they are the same.

3.2.1 Conjugate priors

What functional form can we choose for the prior $g(p)$ such that the posterior $g(p \mid n, r, I)$ has the same functional form? This requires some serious mathematical work,

but the answer is the Beta distribution. The Beta distribution is parametrized by two positive parameters, a and b,

$$g(p \mid a, b) = \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)}, \tag{3.49}$$

where

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{3.50}$$

is the Beta function. The distribution is defined on the interval $0 \leq p \leq 1$. Importantly, or $a = b = 1$, we get a Uniform distribution. The Uniform distribution on the interval from zero to one is therefore a special case of the Beta distribution.

Now, if we insert a Beta distribution for the posterior and prior, we have

$$g(p \mid n, r, a, b) = \frac{f(r \mid p, n)\, g(p \mid a, b)}{f(r \mid n)} \tag{3.51}$$

$$= \frac{1}{f(r \mid n)} \frac{n!}{(n-r)!r!} p^r (1-p)^{n-r} \frac{p^{a-1}(1-p)^{b-1}}{B(a,b)} \tag{3.52}$$

$$= \frac{1}{f(r \mid n)\, B(a,b)} \frac{n!}{(n-r)!r!} p^{r+a-1}(1-p)^{n-r+b-1}. \tag{3.53}$$

In looking at this expression, the only bit that depends on p is $p^{r+a-1}(1-p)^{n-r+b-1}$, which is exactly the p-dependence of a Beta distribution with parameters $r+a$ and $n-r+b$. Because the posterior must be normalized, the posterior must be a Beta distribution and

$$\frac{1}{f(r \mid n)\, B(a,b)} \frac{n!}{(n-r)!r!} = \frac{1}{B(r+a, n-r+b)}. \tag{3.54}$$

We have just normalized the posterior without doing any nasty integrals! So, the posterior is

$$g(p \mid n, r, a, b) = \frac{p^{r+a-1}(1-p)^{n-r+b-1}}{B(r+a, n-r+b)}, \tag{3.55}$$

or,

$$p \mid n, r, a, b \sim \text{Beta}(r+a, n-r+b). \tag{3.56}$$

So, we can see that conjugacy is useful. For a given likelihood, if we know its conjugate prior, we can just immediately write down the posterior in a clear form. The Wikipedia page on conjugate priors has a useful table of likelihood-conjugate pairs.

Note though that a closed form conjugate does not always exist for a given likelihood, especially for complicated models, and when they do exist, they may be very difficult to find. This does limit their utility. Further, there is no reason why a posterior and prior should have the same functional form; all analysis is completely valid without conjugacy. Sivia has stinging words about using conjugate priors: "While we might expect our initial understanding of the object of interest to have a bearing on the experiment we conduct, it seems strange that the choice of the prior pdf should have to wait for, and depend in detail upon, the likelihood function."

3.3 The impediment is not resolved

We tried to be as objective as possible in choosing our priors. We intentionally tried to be uninformative, and took into account transformation invariance. This has flaws, since it is mathematically impossible to come up with a prior that can be invariant to *all* transformations. There are other strategies for choosing uninformative priors. Among them are

- Using what is generically called a Jeffreys prior by computing the Fisher information from the likelihood.
- Using the principle of maximum entropy. Entropy can be thought of as a formal metric of ignorance, which we wish to maximized when being objective. Sivia talks about this in Chapter 5.

Both of these methods are outside the scope of this course, but they are important to consider when choosing priors.

Now consider Sivia's comment I just quoted. And now consider the title of a recent paper by Gelman, Simpson, and Betancourt, "The prior can generally only be understood in the context of the likelihood." Some of the section headings in that paper are also gold, like "Uniform priors are not a panacea and can do unbounded damage."

So, obviously there is disagreement about constructing priors. One the one hand, we want to be as objective as possible in predicting priors. That said, we almost always do know *something* about parameter values *a priori*. We really *should* encode that in the prior. Furthemore, a probability density function is just a pdf. It only becomes a prior when it is connected to a likelihood. So we may need to this with some degree of pragmatism.

It is very hard to be truly informative in more complicated models, such as the very powerful hierarchical models we will work with later in the class. Furthermore, when using flat priors, the other impediment comes back in. Flat priors can really wreak havoc on Markov chain Monte Carlo (MCMC) samplers in hierarchical mod-

els, thereby exacerbating the difficulty in computing the posterior. (If you cannot compute it, what good is it?)

It's probably no coincidence that Gelman and coworkers are lead developers on one of the major MCMC packages, called Stan. In fact, the Stan wiki has some guidelines about choice of priors, which run quite contradictory to what we have just discussed here. Specifically, as of October 11, 2017, there is this: "Some principles we don't like: invariance, Jeffreys, entropy." The Stan developers are obviously going to be more pragmatic in their views, since they are in the business of actually computing posteriors. They tend to favor **weakly informative priors**; things like broad Gaussians. Their reasons, again quoting the Wiki,

> - Weakly informative prior should contain enough information to regularize: the idea is that the prior rules out unreasonable parameter values but is not so strong as to rule out values that might make sense
> - Weakly informative rather than fully informative: the idea is that the loss in precision by making the prior a bit too weak (compared to the true population distribution of parameters or the current expert state of knowledge) is less serious than the gain in robustness by including parts of parameter space that might be relevant.

In the end, my view is that you want to encode all of the information you confidently have about parameters, and not more, into the prior. For example, *before seeing the data*, if you think that the variability in measured spindle length should be about a 10 microns, you could choose a weakly informative prior, like a Gaussian with mean of one micron and standard deviation of 3 microns, and you would probably be fine. As you can see in homework 3.4, the choice of prior often has very little effect on the end result of your inference.

4 The theory of Markov chain Monte Carlo

4.1 Why MCMC?

When doing Bayesian analysis, our goal is very often to compute a posterior distribution, $g(\theta \mid D)$, where $\theta = \{\theta^{(1)}, \theta^{(2)}, \ldots\}$ is a set of possibly many parameters. However, just having an analytical expression for the posterior is of little use if we cannot understand any properties about it. Importantly, we often want to marginalize the posterior; that is, we want to integrate over parameters we are not interested in and get simpler distributions for those we are. This is often necessary to understand all but the simplest models. Doing these marginalizations requires what David MacKay calls "macho integration," which is often impossible to do analytically.

Furthermore, we may also want to compute **expectations** out of the posterior. For example, we might want the mean, or expectation value, of parameter $\theta^{(1)}$. If we know the posterior, this is

$$E[\theta^{(1)}] = \int \mathrm{d}\theta \; \theta^{(1)} \, g(\theta \mid D). \tag{4.1}$$

Generally, we can compute the expectation of any function of the parameters, $h(\theta)$, and we often want to. This is

$$E[h(\theta)] = \int \mathrm{d}\theta \, h(\theta) \, g(\theta \mid D). \tag{4.2}$$

So, pretty much anything we want to know about the posterior requires computation of an integral.

MCMC allows us to **sample** out of an arbitrary probability distribution, which includes pretty much any posterior we could write down.[9] By sampling, we mean that we can choose values of the parameters, θ, where the probability that we choose a given value is proportional to the posterior probability. Note that each sample consists of a complete set of parameters θ; that is a sample contains a value for $\theta^{(1)}$, a value for $\theta^{(2)}$, ... We are more likely to choose samples of high probability than of low. Using MCMC, we collect a large number of these samples.

From these samples, we can trivially perform marginalizations. Say we are interested in the marginalized distribution

$$g(\theta^{(1)} \mid D) = \left(\int \mathrm{d}\theta^{(2)} \int \mathrm{d}\theta^{(3)} \cdots \right) g(\theta \mid D). \tag{4.3}$$

[9] Well, not *any*. For some cases, we may not be able to make a transition kernel that satisfies the necessary properties, which I describe in the following pages.

Given a set of MCMC samples out of $g(\theta \mid D)$, to get a set of samples out of $g(\theta^{(1)} \mid D)$, we simply ignore the values of $\theta^{(2)}$, $\theta^{(3)}$, ...! Then, given the samples of the marginalized posterior, we can plot the CDF of the marginalized posterior as an ECDF of the samples, and the PDF of the marginalized posterior as a histogram of the samples.

To compute expectations, the MCMC samples are again very convenient. Now, we just approximate the integral with an average over samples.

$$E(h(\theta)) = \int \mathrm{d}\theta \, h(\theta) \, g(\theta \mid D) \approx \frac{1}{N} \sum_{i=1}^{N} h(\theta_i), \tag{4.4}$$

where θ_i is the ith of N MCMC samples taken from the posterior.

It is now abundantly clear why the ability to generate samples from the posterior is so powerful. But generating samples that actually come from the probability distribution of interest is not a trivial matter. We will discuss how this is accomplished through MCMC.

4.2 The basic idea behind MCMC

We often draw *independent* samples from a **target distribution**. For example, we could use `np.random.uniform(0, 1, 100)` to draw 100 independent samples from a uniform distribution on the domain [0, 1]. Generating independent samples for complicated target distributions is difficult.

But the samples need not be independent! Instead, we only need that the samples be generated from a process that generates samples from the target distribution in the correct proportions. In the case of the parameter estimation problem, this distribution is the posterior distribution parametrized by θ, $g(\theta \mid D)$. For notational simplicity in what follows, since we know we are always talking about a posterior distribution, we will use $P(\theta)$ for shorthand notation for an arbitrary distribution of theta.

The approach of MCMC is to take random walks in parameter space such that the probability that a walker arrives at point θ is proportional to $P(\theta)$. This is the main concept and is important enough to repeat.

> The approach of MCMC is to take random walks in parameter space such that the probability that a walker arrives at point θ is proportional to $P(\theta)$.

If we can achieve such a walk, we can just take the walker positions as samples from the distributions. To implement this random walk, we define a **transition kernel**, $T(\theta_{i+1} \mid \theta_i)$, the probability of a walker stepping from position θ_i in parameter space to position θ_{i+1}. The transition kernel defines a **Markov chain**, which you

can think of as a random walker whose next step depends only on where the walker is right now; i.e., it has no memory.

The condition that the probability of arrival at point θ_{i+1} is proportional to $P(\theta_{i+1})$ may be stated as

$$P(\theta_{i+1}) = \int \mathrm{d}\theta_i \, T(\theta_{i+1} \mid \theta_i) \, P(\theta_i). \tag{4.5}$$

Here, we have taken θ to be continuous. Were it discete, we just replace the integral with a sum. When this relation holds, it is said that the target distribution is an **invariant distribution** or **stationary distribution** of the transition kernel. When this invariant distribution is unique, it is called a **limiting distribution**. We want to choose our transition kernel $T(\theta_{i+1} \mid \theta_i)$ such that $P(\theta)$ is limiting. This is the case if equation (4.5) holds and the chain is **ergodic**. An ergodic Markov chain has the following properties:

1. It is **aperiodic**. A periodic Markov chain can only return to a given point in parameter space after k, $2k$, $3k, \ldots$ steps, where k is the period. An aperiodic chain is not periodic.
2. It is **irreducible**, which means that any point in parameter space is accessible to the walker from any other.
3. It is **positive recurrent**, which means that the walker will surely come revisit any point in parameter space in a finite number of steps.

So, if our transition kernel satisfies this checklist and equation (4.5), it will eventually sample the posterior distribution. We will discuss how to come up with such a transition kernel in a moment; for now we focus on the important concept of "eventually" in the preceding sentence.

4.3 Tuning

Imagine for a moment that we devised a transition kernel that satisfies the above properties. Say we start a walker at position θ_0 in parameter space and it starts walking according to the transition kernel. Most likely, for those first few steps, the walker is traversing a part of parameter space that has incredibly low probability. Once it got to regions of high probability, the walker would almost never return to the region of parameter space in which it began. So, unless we sample for an incredibly long time, those first few samples visited are over-weighted. So, we need to let the walker walk for a while without keeping track of the samples so that it can arrive at the limiting distribution. This is called **tuning**, otherwise known as **burn-in** or **warm up**[10].

[10] When using NUTS with PyMC3, the tuning is a bit more than just burn-in, where we simply neglect samples. The algorithm is actively choosing stepping strategies during the tuning phase.

There is no general way to tell if a walker has reached the limiting distribution, so we do not know how many burn-in steps are necessary. There are several heuristics. For example, Gelman and coauthors proposed generating several tuning chains and computing the **Gelman-Rubin** $\hat{R}$ statistic,

$$\hat{R} = \frac{\text{variance between the chains}}{\text{mean variance within the chains}}. \tag{4.6}$$

Limiting chains have $\hat{R} \approx 1$, so you can use this as a metric for having achieved stationarity. Gelman and his coauthors in their famous book *Bayesian Data Analysis* suggest that $|1 - \hat{R}| < 0.1$ as a good rule of thumb for stationary chains.

4.4 Generating a transition kernel: The Metropolis-Hastings algorithm

The **Metropolis-Hastings algorithm** covers a widely used class of algorithms for MCMC sampling. I will first state the algorithm here, and then we will show that it satisfies the necessary conditions for the walkers to be sampling out of the target posterior distribution.

4.4.1 The algorithm/kernel

Say our walker is at position θ_i in parameter space.

1. We randomly choose a candidate position θ' to step to next from an arbitrary **proposal distribution** $K(\theta' \mid \theta_i)$.
2. We compute the **Metropolis ratio**,

$$r = \frac{P(\theta')\,K(\theta_i \mid \theta')}{P(\theta_i)\,K(\theta' \mid \theta_i)}. \tag{4.7}$$

3. If $r \geq 1$, accept the step and set $\theta_{i+1} = \theta'$. Otherwise, accept the step with probability r. If we do reject the step, set $\theta_{i+1} = \theta_i$.

The last two steps are used to define the transition kernel $T(\theta_{i+1} \mid \theta_i)$. We can define the acceptance probability of the proposal step as

$$\alpha(\theta_{i+1} \mid \theta_i) = \min(1, r) = \min\left(1, \frac{P(\theta_{i+1})\,K(\theta_i \mid \theta_{i+1})}{P(\theta_i)\,K(\theta_{i+1} \mid \theta_i)}\right). \tag{4.8}$$

Then, the transition kernel is

$$T(\theta_{i+1} \mid \theta_i) = \alpha(\theta_{i+1} \mid \theta_i)\,K(\theta_{i+1} \mid \theta_i). \tag{4.9}$$

4.4.2 Detailed balance

This algorithm seems kind of nuts! How on earth does this work? To investigate this, we consider the joint probability, $P(\theta_{i+1}, \theta_i)$, that the walker is at θ_i and θ_{i+1} at sequential steps. We can write this in terms of the transition kernel,

$$\begin{aligned}
P(\theta_{i+1}, \theta_i) &= P(\theta_i)\, T(\theta_{i+1} \mid \theta_i) \\
&= P(\theta_i)\, \alpha(\theta_{i+1} \mid \theta)\, K(\theta_{i+1} \mid \theta_i) \\
&= P(\theta_i)\, K(\theta_{i+1} \mid \theta)\, \min\left(1, \frac{P(\theta_{i+1})\, K(\theta_i \mid \theta_{i+1})}{P(\theta_i)\, K(\theta_{i+1} \mid \theta_i)}\right) \\
&= \min\left[P(\theta_i)\, K(\theta_{i+1} \mid \theta_i), P(\theta_{i+1})\, K(\theta_i \mid \theta_{i+1})\right] \\
&= P(\theta_{i+1})\, K(\theta_i \mid \theta_{i+1})\, \min\left(1, \frac{P(\theta_i)\, K(\theta_{i+1} \mid \theta_i)}{P(\theta_{i+1})\, K(\theta_i \mid \theta_{i+1})}\right) \\
&= P(\theta_{i+1})\, \alpha(\theta_i \mid \theta_{i+1})\, K(\theta_i \mid \theta_{i+1}) \\
&= P(\theta_{i+1})\, T(\theta_i \mid \theta_{i+1}). \qquad (4.10)
\end{aligned}$$

Thus, we have

$$P(\theta_i)\, T(\theta_{i+1} \mid \theta_i) = P(\theta_{i+1})\, T(\theta_i \mid \theta_{i+1}). \qquad (4.11)$$

This says that the rate of transition from θ_i to θ_{i+1} is equal to the rate of transition from θ_{i+1} to θ_i. In this case, the transition kernel is said to satisfy **detailed balance**.

Any transition kernel that satisfies detailed balance has $P(\theta)$ as an invariant distribution. This is easily shown.

$$\begin{aligned}
\int \mathrm{d}\theta_i\, P(\theta_i)\, T(\theta_{i+1} \mid \theta_i) &= \int \mathrm{d}\theta_i\, P(\theta_{i+1})\, T(\theta_i \mid \theta_{i+1}) \\
&= P(\theta_{i+1}) \left[\int \mathrm{d}\theta_i\, T(\theta_i \mid \theta_{i+1})\right] \\
&= P(\theta_{i+1}), \qquad (4.12)
\end{aligned}$$

since the bracketed term is unity because the transition kernel is a probability.

Note that all transition kernels that satisfy detailed balance have an invariant distribution. (If the chain is ergodic, this is a limiting distribution.) But not all kernels that have an invariant distribution satisfy detailed balance. So, detailed balance is a sufficient condition for a transition kernel having an invariant distribution.

4.4.3 Choosing the transition kernel

There is an art to choosing the transition kernel. The original Metropolis algorithm (1953), took $K(\theta_{i+1} \mid \theta_i) = 1$. As a rule of thumb, you want to choose a proposal distribution such that you get an acceptance rate of about 0.4. If you accept every step, the walker just wanders around and it takes a while to get to the limiting distribution. If you reject too many steps, the walkers never move, and it again takes a long time to get to the limiting distribution. There are tricks to "tune" the walkers to achieve the target acceptance rate.

Gibbs sampling, which is popular, though we will not go into the details, is a special case of a Metropolis-Hastings sampler, as is the No U-turn sampler (NUTS), which is an example of a **Hamiltonian Monte Carlo** sampler. These both result in significant performance improvements for important subclasses of problems. The sampler employed by `emcee`, the affine invariant ensemble sampler (Goodman and Weare, *J. Comp. Sci.*, **5**, 65–80, 2000), utilizes many walkers walking at the same time, sharing information between them. It is technically not a Metropolis-Hastings sampler, but many of the ideas presented in this lecture there apply for ensuring that the sampler is indeed sampling the appropriate posterior distribution.

Finally, importantly, the No U-Turn sampler and the affine invariant sample can only handle continuous variables; they cannot sample discrete variables. Depending on your problem, this could be a serious limitation.

In this class, we will use PyMC3, which uses NUTS. We will not delve into the algorithmic details, but it helps to have a feel for how the algorithm works. To educate yourself more‘ recommend Michael Betencourt's conceptual introduction to Hamiltonian Monte Carlo and this lecture by him on that topic.

5 Model comparison

We have spent a lot of time in the past couple of weeks looking at the problem of parameter estimation. Really, we have been stepping through the process of bringing our thinking about a biological system into a concrete statistical model that defines a likelihood for the data and the parametrization thereof. Writing down Bayes's theorem then gives the posterior,

$$g(\theta \mid D) = \frac{f(D \mid \theta)\, g(\theta)}{P(D)}, \tag{5.1}$$

where θ is the set of parameters. Solving the parameter estimation problem involves computing the posterior, which usually involves summarizing the posterior into a form that can be processed intuitively.

5.1 Adding models to the probabilities

When we write Bayes's theorem for the parameter estimation problem, implicit in the definition of the likelihood is the fact that we are using a specific statistical model. To be complete, especially in the context of model comparison, we should include which model we're using in the conditions of the probabilities. Let M_i denote a model i, and θ_i be the set of parameters associated with M_i.[11] Then, we have

$$g(\theta_i \mid D, M_i) = \frac{f(D \mid \theta_i, M_i)\, g(\theta_i \mid M_i)}{f(D \mid M_i)}. \tag{5.2}$$

This is a more explicit description of the probabilities associated with the parameter estimation problem.

5.2 Probabilities of models

Remember that Bayesian probability is a measure of the plausibility of any logical conjecture. So, we can talk about the probability of models being true. So, what is the probability that a model is true, given the observed data? Again, this is given by Bayes's theorem.

$$g(M_i \mid D) = \frac{f(D \mid M_i)\, g(M_i)}{f(D)}. \tag{5.3}$$

This is Bayes's theorem stated for the model comparison problem. Let's look at each term in turn.

[11] Do not be confused by the subscript here. The i does not signify the ith parameter of a set of parameters for a given mode. Here, it means that θ_i describes the set of parameters for model i.

- $g(M_i \mid D)$, as we said before, is the probability that model M_i is true given the measured data.
- $f(D)$ is a normalization constant for the posterior that is computed by marginalizing over all possible models

$$\sum_i g(M_i \mid D) = 1 \;\Rightarrow\; f(D) = \sum_i f(D \mid M_i)\, g(M_i). \tag{5.4}$$

- $g(M_i)$ is a measure of how plausible we thought model M_i is a priori, the prior probability for model M_i. For example, if a proposed model violates a physical conservation law, we know it is unlikely to be true even before we see the data. In practice, we typically assign equal probability to all models we have not ruled out prior to seeing the data.
- $f(D \mid M_i)$ is the likelihood of observing the data, given that model M_i is true.

As usual, we need to specify the likelihood and prior to assess the posterior probability of any given model. We already discussed how to specify the prior. We usually assume all models are equally likely. How about the likelihood? Well, glancing at equation (5.2), we see that the likelihood for the model comparison problem is the evidence for the parameter estimation problem! Because the posterior in the parameter estimation problem, $g(\theta_i \mid D, M_i)$, must be normalized, the evidence in the parameter estimation problem, and therefore also the likelihood in the model comparison problem, is given by

$$f(D \mid M_i) = \int \mathrm{d}\theta_i\, f(D \mid \theta_i, M_i)\, g(\theta_i \mid M_i). \tag{5.5}$$

So, if we can compute the likelihood and priors from the parameter estimation problem and can integrate their product, we have the likelihood for the model comparison problem.

5.3 Bayes factors and odds ratios

Computing the absolute probability of a model is difficult, since it would require considering all possible models, as is required to compute the normalization constant, $f(D)$. We therefore typically make pairwise comparisons between models. This comparison is called an **odds ratio**. It is the ratio of the probabilities of two models being true.

$$O_{ij} = \frac{g(M_i)}{g(M_j)} \left[\frac{f(D \mid M_i)}{f(D \mid M_j)}\right]. \tag{5.6}$$

The first factor in the product is the ratio of our prior knowledge of the truth of the models. If they are equally likely, this ratio is unity. The bracketed ratio is called the **Bayes factor**, which is the ratio of the evidences of the respective models.

Note that if we compute all of the odds ratios comparing a given model k to all others (and somehow did manage to consider all models that have nonzero probability), we can compute the posterior probability of model M_i as

$$g(M_i \mid D) = \frac{O_{ik}}{\sum_j O_{jk}}. \tag{5.7}$$

5.4 Approximate computation of the Bayes factor

Evaluating the integral in equation (5.5) to compute the Bayes factor is in general difficult. If the posterior is sharply peaked, we may compute this integral using the **Laplace approximation** in which we approximate the integral by the height of the peak times its width. In one dimension, this is

$$\begin{aligned} f(D \mid M_i) &= \int \mathrm{d}\theta_i f(D \mid \theta_i, M_i)\, g(\theta_i \mid M_i) \\ &\approx f(D \mid \theta_i^*, M_i)\, g(\theta_i^* \mid M_i) \sqrt{2\pi\,\sigma_i^2}, \end{aligned} \tag{5.8}$$

where θ_i^* is the MAP estimate, and σ_i^2 is the variance of the Gaussian approximation of the posterior. In n-dimensions, this is

$$g(D \mid M_i) = \int \mathrm{d}\theta_i f(D \mid \theta_i, M_i)\, g(\theta_i \mid M_i) \tag{5.9}$$

$$\approx f(D \mid \theta_i^*, M_i)\, g(\theta_i^* \mid M_i)\ (2\pi)^{|\theta_i|/2} \sqrt{\det \Sigma_i}, \tag{5.10}$$

where Σ_i is now the covariance matrix of the Gaussian approximation of the posterior under M_i. We have also denoted the number of parameters in M_i to be $|\theta_i|$. Note that we have already computed all of factors in the above product in the parameter estimation problem if we solved it by optimization. Therefore, we already have what we need to compute the (approximate) odds ratio.

5.5 The factors in the odds ratio

We can now write the approximate odds ratio as the product of three factors.

$$O_{ij} \approx \left(\frac{g(M_i)}{g(M_j)}\right) \left(\frac{f(D \mid \theta_i^*, M_i)}{f(D \mid \theta_j^*, M_j)}\right) \left(\frac{g(\theta_i^* \mid M_i)\ (2\pi)^{|\theta_i|/2} \sqrt{\det \Sigma_i}}{g(\theta_j^* \mid M_j)\ (2\pi)^{|\theta_j|/2} \sqrt{\det \Sigma_j}}\right). \tag{5.11}$$

- The first term represents the prior probability of the models. This is how plausible we thought the models were before the experiment.

- The second term is a measure of the goodness of fit. In other words, it comments on how probable the data are given the model and the MAP estimate.
- The third term is a ratio of **Occam factors**. An Occam factor is the ratio of the volume of parameter space accessible to the posterior to that of the prior. This is best seen by example. Consider a model M_1 with a single parameter where the parameter a that has a Uniform prior. Then,

$$\text{Occam factor} = \sqrt{2\pi} g(a^* \mid M_1)\sigma_1 = \frac{\sqrt{2\pi}\ \sigma_1}{a_{\max} - a_{\min}}. \tag{5.12}$$

  Remember, σ_1^2 is the variance of the Gaussian approximation of the posterior. So, the numerator here is the width of the posterior and the denominator is the width of the prior.

  Now, consider a model, M_2 with two parameters, b and c, each with Uniform priors. In this case, we have

$$g(b^*, c^* \mid M_j) = \frac{1}{b_{\max} - b_{\min}} \frac{1}{c_{\max} - c_{\min}}, \tag{5.13}$$

  and the Occam factor is

$$\text{Occam factor} = \frac{2\pi\sqrt{\det \Sigma_2}}{(b_{\max} - b_{\min})(c_{\max} - c_{\min})}. \tag{5.14}$$

  So, the volume of the parameter space accessible to the prior for model M_2 is larger than for M_1, so the part of the odds ratio is greater than one, favoring the model with fewer parameters. The ratio of Occam factors is then

$$\frac{\sigma_i}{\sqrt{2\pi\ \det \sigma_j^2}} (b_{\max} - b_{\min}). \tag{5.15}$$

  Comparing the Occam factors of the two models, we see that the more parameters you have, the bigger the denominator of the Occam factor is, making the Occam factor smaller. Furthermore, it is also often the case that complicated models with lots of parameters also have smaller determinants of the covariance because the multitude of parameters are "locked in" around the MAP estimate. Thus, we see where the Occam factor gets its name, since it penalizes more complicated models.[12]

This approximate calculation shows us everything that goes into the odds ratio. Any one factor can overwhelm the others:

- What we knew before
- How well the model fits the data
- How simple the model is

[12]Remember that Occam's razor states that among competing hypotheses, the one with fewest assumptions is preferred.

5.6 Example: Are two data sets from the same Gaussian distribution?

We will now look at an example. Say I do two sets of measurements of property x, a control and an experiment. We make n_c control measurements and n_e experiment measurements. We consider two models. M_1 says that both the control and the experiment are chosen from the same underlying Gaussian distribution with mean μ and variance σ. Model M_2 says that control and experiment come from different Gaussian distributions with means μ_c and μ_e. We wish to compare models M_1 and M_2. The odds ratio is

$$O_{12} = \frac{g(M_1)\, f(D_c, D_e \mid M_1)}{g(M_2)\, f(D_c, D_e \mid M_2)}, \tag{5.16}$$

where D_c denotes the data from the control experiment and D_e denotes the data from the experiment.

We will assume a prior that $g(M_i) = g(M_j)$. Then, we are left to compute $f(D_c, D_e \mid M_1)$ and $f(D_c, D_e \mid M_2)$. We can do this by approximate integration (see section 4.3.1 of Sivia). Note that we assume a uniform prior on σ, with $0 < \sigma < \sigma_{\max}$. We could also try the problem with a Jeffreys prior on σ, but I do not feel like doing the nasty integration. The result for the odds ratio is

$$O_{12} \approx \frac{\sigma_{\max}\,(\mu_{\max} - \mu_{\min})}{\pi\sqrt{2}}\; \frac{n_1\, n_2\, s^{2-n1-n2}}{(n_1 + n_2)\, s_1^{2-n_1}\, s_2^{2-n_2}}, \tag{5.17}$$

where

$$s^2 = \frac{1}{n_1 + n_2} \sum_{i \in D_1 \cup D_2} (x_i - \bar{x})^2, \tag{5.18}$$

$$s_1^2 = \frac{1}{n_1} \sum_{i \in D_1} (x_i - \bar{x}_1)^2, \tag{5.19}$$

$$s_2^2 = \frac{1}{n_2} \sum_{i \in D_2} (x_i - \bar{x}_2)^2, \tag{5.20}$$

with

$$\bar{x} = \frac{1}{n_1 + n_2} \sum_{i \in D_1 \cup D_2} x_i, \tag{5.21}$$

$$\bar{x}_1 = \frac{1}{n_1} \sum_{i \in D_1} x_i, \tag{5.22}$$

$$\bar{x}_2 = \frac{1}{n_2} \sum_{i \in D_2} x_i. \tag{5.23}$$

It seems that this question is often asked: does the experiment come from a different process than the control? My opinion is that in most situations, the answer is an obvious yes, and the more pertinent question is by how much they differ. Nonetheless, if we are asking the "if they are different" question, we can plug our data in and easily compute the odds ratio. Be careful, though. This, too, should not be a yes-or-no question. We should not really be asking *if* they come from different distributions, what are the odds that they do.

5.7 Caveats and motivation for information criteria

Gelman, et al., in their book *Bayesian Data Analysis* (3rd Ed, page 182), express some concern about the approach we have taken here. I quote them, emphasis theirs, bracketed comment mine.

> This fully Bayesian approach has some appeal but we generally do *not* recommend it because, in practice, the marginal likelihood [which we have been calling the evidence from the parameter estimation prior] is highly sensitive to aspects of the model that are typically assigned arbitrarily and are untestable from data.

These arbitrary and untestable aspects are typically the priors. We try to be uninformative, but they *must* be proper (meaning normalized) in order to do model comparison as we have done here. The Bayes factor is highly sensitive to the width of the priors.

In my opinion, this method of model comparison is often perfectly legitimate because the prior is part of the model and, while untestable from data, is not arbitrary. If constructed properly, the prior represents our knowledge before data acquisition and should therefore naturally be included in model comparison.

Whatever your position on this matter, it is still useful to have other metrics for assessing models.

5.8 Watanabe-Akaike Information Criterion (WAIC)

A good model is a predictive model. If we were to acquire more data under identical conditions, the parameters we derived from the posterior should be able to accurately predict what those new data would look like. It makes sense to assess a model on how well it can predict new data. Furthermore, the connection between predictive capabilities and the Bayes factor is clear if you think about what must be true of a predictive model. First, it must describe the data we have actually acquired well. The goodness-of-fit term in the Bayes factor covers this. Second, it must describe

new data well. If the parameters are such that it describes the data already collection very well, but cannot predict, the model is not good. This usually happens when the model has many parameters tailor-made for the data (such as fitting data with a higher-order polynomial). This is captured in the Occam factor. So, models with good Bayes factors are often predictive.

With that in mind, I will introduce a good metric for comparing models, the **Watanabe-Akaike Information Criterion**, also known as the **Widely Applicable Information Criterion** (WAIC). I will discuss it intuitively and not provide much rigor. For detailed descriptions of what follows, I recommend reading chapter 7 of Gelman, et al., *Bayesian Data Analysis, 3rd Ed.* and chapter 6 of McElreath, *Statistical Rethinking*.

In what follows, for notational convenience, I will drop explicit dependence of M_i, and also drop the subscripts from the parameter set θ_i. We define the predictive density of a single data point $x \in D$ as

$$\text{single point predictive density} = \int \mathrm{d}\theta \; f(x \mid \theta)\, g(\theta \mid D). \tag{5.24}$$

This is the likelihood for observing data point x, averaged over the posterior probability distribution of parameter values θ. We are therefore taking into account posterior information and using the likelihood to assess goodness-of-fit. We can take the product of each of the single point predictive densities in the data set and take the logarithm to get the **log pointwise predictive density**, or **lppd**,

$$\begin{aligned} \text{lppd} &= \ln \left(\prod_{x \in D} \int \mathrm{d}\theta \; f(x \mid \theta)\, g(\theta \mid D) \right) \\ &= \sum_{x \in D} \ln \left(\int \mathrm{d}\theta \; f(x \mid \theta)\, g(\theta \mid D) \right). \end{aligned} \tag{5.25}$$

This gives a metric of how well the model manages to predict the observed data. Put succinctly, the lppd is the sum of the logarithm of the average likelihood of each observation in a data set.

This metric is biased toward complicated models, so we add a correction. We compute the **effective number of parameters**, p_{WAIC} as

$$p_{\text{WAIC}} = \sum_{i \in D} \text{variance}(\ln f(x \mid D)), \tag{5.26}$$

where the variance is computed over the posterior. Written out, this is

$$\text{variance}(\ln f(x \mid D)) = \int \mathrm{d}\theta \; g(\theta \mid D)\, (\ln f(x \mid D))^2$$

$$- \left(\int \mathrm{d}\theta \, g(\theta \mid D) \, \ln f(x \mid D) \right)^2 . \tag{5.27}$$

This parameter p_{WAIC}, can be thought of as the number of unconstrained parameters in a model. Parameters that are influences only by the prior contribute little to p_{WAIC}, while those that are determined mostly by the data contribute more.

The WAIC is then

$$\text{WAIC} = -2(\text{lppd} - p_{\text{WAIC}}). \tag{5.28}$$

The factor of -2 is there for historical reasons to enable comparisons to the Akaike Information Criterion (AIC) and the Deviance Information Criterion (DIC). These two information criteria are also widely used, but have assumptions about Gaussianity, and in the case of the AIC, also flat priors. The WAIC is a better choice.

Computing the WAIC is difficult, unless, of course, you managed to get MCMC samples! Given a set of S MCMC samples of the parameters θ (where $\theta^{(s)}$ is the sth sample), the lppd may be calculated as

$$\text{lppd} = \sum_{x \in D} \ln \left(\frac{1}{S} \sum_{s=1}^{S} f(x \mid \theta^{(s)}) \right) . \tag{5.29}$$

Another beautiful example of how sampling converts integrals into sums. Similarly we can compute p_{WAIC} from samples.

$$p_{\text{WAIC}} = \sum_{x \in D} \frac{1}{S-1} \sum_{s=1}^{S} \left(\log f(x \mid \theta^{(s)}) - q(x) \right)^2 , \tag{5.30}$$

where

$$q(x) = \frac{1}{S} \sum_{s=1}^{S} \ln f(x \mid \theta^{(s)}). \tag{5.31}$$

While you can compute the WAIC from your MCMC samples, PyMC3 has a built-in function to do it.

For an intuitive description of the WAIC, you may think of it as an estimate of the negative log likelihood of new data.[13] That is, it is an estimate of how badly the model would perform with new data. So, the lower the WAIC, the better the model.

[13] Stated precisely, the WAIC is an estimate of the out-of-sample deviance. "Out-of-sample" just means data that is yet to come. I did not want to go through the trouble of defining deviance.

5.9 The Akaike weights

The value of a WAIC by itself does not tell us anything. Only comparison of two or more WAICs makes sense. Recalling that the WAIC is a measure of a log likelihood, if we exponentiate it, we get something proportional to a probability. If we have two models, M_i and M_j, the **Akaike weight** of model i is

$$w_i = \frac{\exp\left[-\frac{1}{2}\,\text{WAIC}_i\right]}{\exp\left[-\frac{1}{2}\,\text{WAIC}_i\right] + \exp\left[-\frac{1}{2}\,\text{WAIC}_j\right]}. \tag{5.32}$$

This weight may be interpreted as an estimate of the probability that M_i will make the best predictions of new data.[14] We can generalize this to multiple models.

$$w_i = \frac{\exp\left[-\frac{1}{2}\,\text{WAIC}_i\right]}{\sum_j \exp\left[-\frac{1}{2}\,\text{WAIC}_j\right]}. \tag{5.33}$$

We can compute a quantity analogous to the Bayesian odds ratio,

$$\frac{w_i}{w_j} = \exp\left[-\frac{1}{2}(\text{WAIC}_i - \text{WAIC}_j)\right]. \tag{5.34}$$

5.10 Computing odds ratios and information criteria

You may have noticed that computing the WAIC almost always required performing an MCMC calculation In the approximate calculation of the odds ratio, I only used MAP information that could be found by optimization. This, however, is approximate, and has all the perils associated with posteriors that are strongly non-Gaussian. There are information criteria that can be computed from MAP estimates as well. These also have dangers associated with them.

So, how do you compute the odds ratio (via Bayes factor) from MCMC? We can use a technique called parallel-tempering Markov chain Monte Carlo (PTMCMC) to exactly compute the odds ratio. As you likely have guessed, this is computationally intensive, but effective. We will learn about this in an auxiliary lesson.

[14] This interpretation is common, but not entirely agreed upon.

6 Frequentist methods

We have taken a Bayesian approach to data analysis in this class. So far, the main motivation for doing so is that I think the approach is more intuitive. We often think of probability as a measure of plausibility, so a Bayesian approach jibes with our natural mode of thinking. Further, the mathematical and statistical models are explicit, as is all knowledge we have prior to data acquisition. The Bayesian approach, in my opinion, therefore reflects intuition and is therefore more digestible and easier to interpret.

Nonetheless, frequentist methods are in wide use in the biological sciences. They are not more or less valid than Bayesian methods, but, as I said, can be a bit harder to interpret. Importantly, as we will soon see, they can very very useful, and easily implemented, in **nonparametric inference**, which is statistical inference where no model is assumed; conclusions are drawn from the data alone. In fact, most of our use of frequentist statistics will be in the nonparametric context. But first, we will discuss some parametric estimators from frequentist statistics.

6.1 The frequentist interpretation of probability

In the tutorials this week, we will do parameter estimation and hypothesis testing using the frequentist definition of probability. As a reminder, in the frequentist definition of probability, the probability P(A) represents a long-run frequency over a large number of identical repetitions of an experiment. Much like our strategies thus far in the class have been to start by writing Bayes's theorem, for our frequentist studies, we will directly apply this definition of probability again and again, using our computers to "repeat" experiments many time and tally the frequencies of what we see.

The approach we will take is heavily inspired by Allen Downey's wonderful book, *Think Stats* and from Larry Wasserman's All of Statistics. You may also want to watch this great 25-minute talk by Jake VanderPlas, where he discusses the differences between Bayesian and frequentist approaches.

6.2 The plug-in principle

In Bayesian inference, we tried to find the most probable value of a parameter. That is, we tried to find the parameter values at the MAP, or maximum a posteriori probability. We then characterized the posterior distribution to get a credible region for the parameter we were estimating. We will discuss the frequentist analog to the credible region, the **confidence interval** in a moment. For now, let's think about how to get an estimate for a parameter value, given the data.

While what we are about to do is general, for now it is useful to have in your mind a concrete example. Imagine we have a data set that is a set of repeated measurements, such as the repeated measurements of the Dorsal gradient width we studied from the Stathopoulos lab. We have a model in mind: the data are generated from a Gaussian distribution. This means there are two parameters to estimate, the mean μ and the variance σ.

To set up how we will estimate these parameters directly from data, we need to make some definitions first. Let $F(x)$ be the cumulative distribution function (CDF) for the distribution. Remember that the probability density function (PDF), $f(x)$, is related to the CDF by

$$f(x) = \frac{\mathrm{d}F}{\mathrm{d}x}. \tag{6.1}$$

For a Gaussian distribution,

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, \mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \tag{6.2}$$

which defines our two parameters μ and σ.

A **statistical functional** is a functional of the CDF, $T(F)$. A parameter θ of a probability distribution can be defined from a functional, $\theta = T(F)$. For example, the mean, variance, and median are all statistical functionals.

$$\mu = \int_{-\infty}^{\infty} \mathrm{d}x\, x f(x) = \int_{-\infty}^{\infty} \mathrm{d}F(x)\, x, \tag{6.3}$$

$$\sigma^2 = \int_{-\infty}^{\infty} \mathrm{d}x\, (x-\mu)^2 f(x) = \int_{-\infty}^{\infty} \mathrm{d}F(x)\, (x-\mu)^2, \tag{6.4}$$

$$\text{median} = F^{-1}(1/2). \tag{6.5}$$

Now, say we made a set of n measurements, $\{x_1, x_2, \ldots x_n\}$. You can this of this as a set of Dorsal gradient widths if you want to have an example in your mind. We define the **empirical cumulative distribution function**, $\hat{F}(x)$ from our data as

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x), \tag{6.6}$$

with

$$I(x_i \leq x) = \begin{cases} 1 & x_i \leq x \\ 0 & x_i > x. \end{cases} \tag{6.7}$$

We saw this functional form of the ECDF in our first homework. We can then also define an **empirical distribution function**, $\hat{f}(x)$ as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i), \tag{6.8}$$

where $\delta(x)$ is the Dirac delta function. To get this, we essentially just took the derivative of the ECDF.

So, we have now defined an empirical distribution that is dependent only on the data. We now define a **plug-in estimate** of a parameter θ as

$$\hat{\theta} = T(\hat{F}). \tag{6.9}$$

In other words, to get a plug-in estimate a parameter θ, we need only to compute the functional using the empirical distribution. That is, we simply "plug in" the empirical CDF for the actual CDF.

The plug-in estimate for the median is easy to calculate.

$$\widehat{\text{median}} = \hat{F}^{-1}(1/2), \tag{6.10}$$

or the middle-ranked data point. The plug-in estimate for the mean or variance, seem at face to be a bit more difficult to calculate, but the following general theorem will help. Consider a functional of the form of an expectation value, $r(x)$.

$$\begin{aligned}\int \mathrm{d}\hat{F}(x)\, r(x) = \int \mathrm{d}x\, r(x)\hat{f}(x) &= \int \mathrm{d}x\, r(x) \left[\frac{1}{n}\sum_{i=1}^{n} \delta(x - x_i)\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} \int \mathrm{d}x\, r(x)\, \delta(x - x_i) = \frac{1}{n}\sum_{i=1}^{n} r(x_i).\end{aligned} \tag{6.11}$$

This means that the plug-in estimate for an expectation value of a distribution is the mean of the observed values themselves. The plug-in estimate of the mean, which has $r(x) = x$, is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \equiv \bar{x}, \tag{6.12}$$

where we have defined $\bar{x}$ as the traditional sample mean, which we have just shown is the plug-in estimate. This plug-in estimate is implemented in the `np.mean()` function. The plug-in estimate for the variance is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2. \tag{6.13}$$

This plug-in estimate is implemented in the `np.var()` function.

We can compute plug-in estimates for more complicated parameters as well. For example, for a bivariate distribution, the correlation between the two variables, x and y, is defined with

$$r(x) = \frac{(x - \mu_x)(y - \mu_y)}{\sigma_x \sigma_y}, \tag{6.14}$$

and the plug-in estimate is

$$\hat{\rho} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(\sum_i (x_i - \bar{x})^2)\,(\sum_i (y_i - \bar{y})^2)}}. \tag{6.15}$$

6.3 Bias

The **bias** of an estimate is the difference between the expectation value of the estimate and value of the parameter.

$$\text{bias}_F(\hat{\theta}, \theta) = \langle \hat{\theta} \rangle - \theta = \int \mathrm{d}x\, \hat{\theta} f(x) - T(F). \tag{6.16}$$

We often want a small bias because we want to choose estimates that give us back the parameters we expect.

Let's consider a Gaussian distribution. Our plug-in estimate for the mean is

$$\hat{\mu} = \bar{x}. \tag{6.17}$$

In order to compute the the expectation value of $\hat{\mu}$ for a Gaussian distribution, it is useful to know that

$$\langle x \rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, x\, \mathrm{e}^{-(x-\mu)^2/2\sigma^2} = \mu\,. \tag{6.18}$$

Then, we have

$$\langle \hat{\mu} \rangle = \langle \bar{x} \rangle = \frac{1}{n} \left\langle \sum_i x_i \right\rangle = \frac{1}{n} \sum_i \langle x_i \rangle = \langle x \rangle = \mu\,, \tag{6.19}$$

so the bias in the plug-in estimate for the mean is zero. It is said to be **unbiased**.

To compute the bias of the plug-in estimate for the variance, it is useful to know that

$$\langle x^2 \rangle = \int_{-\infty}^{\infty} \mathrm{d}x\, x^2\, \mathrm{e}^{-(x-\mu)^2/2\sigma^2} = \sigma^2 + \mu^2, \tag{6.20}$$

so

$$\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2. \tag{6.21}$$

So, the expectation value of the plug-in estimate is

$$\left\langle \hat{\sigma}^2 \right\rangle = \left\langle \frac{1}{n} \sum_i x_i^2 \right\rangle - \left\langle \bar{x}^2 \right\rangle = \frac{1}{n} \sum_i \left\langle x_i^2 \right\rangle - \left\langle \bar{x}^2 \right\rangle = \mu^2 + \sigma^2 - \left\langle \bar{x}^2 \right\rangle. \tag{6.22}$$

We now need to compute $\left\langle \bar{x}^2 \right\rangle$, which is a little trickier. We will use the fact that the measurements are independent, so $\langle x_i x_j \rangle = \langle x_i \rangle \langle x_j \rangle$ for $i \neq j$.

$$\begin{aligned}
\left\langle \bar{x}^2 \right\rangle &= \left\langle \left(\frac{1}{n} \sum_i x_i \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \left(\sum_i x_i \right)^2 \right\rangle = \frac{1}{n^2} \left\langle \sum_i x_i^2 + 2 \sum_i \sum_{j>i} x_i x_j \right\rangle \\
&= \frac{1}{n^2} \left(\sum_i \left\langle x_i^2 \right\rangle + 2 \sum_i \sum_{j>i} \langle x_i x_j \rangle \right) = \frac{1}{n^2} \left(n(\sigma^2 + \mu^2) + 2 \sum_i \sum_{j>i} \langle x_i \rangle \langle x_j \rangle \right) \\
&= \frac{1}{n^2} \left(n(\sigma^2 + \mu^2) + n(n-1) \langle x \rangle^2 \right) = \frac{1}{n^2} \left(n\sigma^2 + n^2 \mu^2 \right) = \frac{\sigma^2}{n} + \mu^2.
\end{aligned} \tag{6.23}$$

Thus, we have

$$\left\langle \hat{\sigma}^2 \right\rangle = \left(1 - \frac{1}{n} \right) \sigma^2. \tag{6.24}$$

Therefore, the bias is

$$\text{bias} = -\frac{\sigma^2}{n} \tag{6.25}$$

An unbiased estimator would instead be

$$\frac{n}{n-1} \hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2. \tag{6.26}$$

Note that in the none of the above analysis depended on $F(x)$ being the CDF of a Gaussian distribution. For any distribution, we define the property of the distribution known as the mean as $\langle x \rangle$ and that known as the variance as $\langle x^2 \rangle - \langle x \rangle^2$.

Comparison to Bayesian treatment. To compare this parameter estimate to a Bayesian treatment, we will consider a Gaussian likelihood in a Jeffreys prior on σ. Recalling Lecture 2, we found that in this case we got $\bar{x}$ as our most probable value of μ, meaning this is the value of μ at the MAP. The most probable value of σ^2 was $\hat{\sigma}^2$. But wait a minute! We just found that was a biased estimator. What gives?

The answer is that we are considering the *maximally probable* values and not the expectation value of the posterior. Recall that the posterior for estimating the parameters of a Gaussian distribution is

$$P(\mu, \sigma \mid \{x_i\}, I) \propto \frac{e^{-n\hat{\sigma}^2/2\sigma^2}}{\sigma^{n+1}} \exp\left[\frac{n(\mu - \bar{x})^2}{2\sigma^2}\right]. \tag{6.27}$$

After some gnarly integration to compute the normalization constant and the expectation values of μ and σ^2 from this posterior, we get

$$\langle \mu \rangle = \bar{x} \tag{6.28}$$

$$\langle \sigma^2 \rangle = \frac{n}{n-1}\hat{s}^2, \tag{6.29}$$

the same as the unbiased frequentist estimators. Note that $\langle \sigma^2 \rangle \neq \langle \sigma \rangle^2$. Remember, in frequentist statistics, we are not computing a posterior distribution describing the parameters. There is no such thing as the "probability of a parameter value" in frequentist probability. A parameter has a value, and that's that. We report a frequentist estimate for the parameter value based on the expectation values of the assumed underlying distribution. We just showed that, at least for a Gaussian, the expectation value of the posterior gives the unbiased frequentist estimate and the MAP gives the plug-in estimate.

Justification of using plug-in estimates. Despite the apparent bias in the plug-in estimate for the variance, we will normally just use plug-in estimates going forward. (We will use the hat, e.g. $\hat{\theta}$, to denote an estimate, which can be either a plug-in estimate or not.) Note that the bootstrap procedures we lay out in what follows do not *need* to use plug-in estimates, but we will use them for convenience. Why do this? First, the bias is typically small. We just saw that the biased and unbiased estimators of the variance differ by a factor of $n/(n-1)$, which is negligible for large n. In fact, plug-in estimates tend to have much smaller error than the confidence intervals for the parameter estimate, which we will discuss in a moment. Finally, we saw when connecting to the Bayesian estimates that the expectation value is not necessarily always what we want to describe; sometimes (though certainly not always, perhaps even seldom) the MAP is preferred. In this sense, attempting to minimize bias is somewhat arbitrary.

6.4 Bootstrap confidence intervals

The frequentist analog to a Bayesian credible region is a **confidence interval**. Remember, with the frequentist interpretation of probability, we cannot assign a probability to a parameter value. A parameter has one value, and that's that. We can only describe the long-term frequency of observing results about random variables. So, we can define a 95% confidence interval as follows.

> If an experiment is repeated over and over again, the estimate I compute for a parameter, $\hat{\theta}$, will lie between the bounds of the 95% confidence interval for 95% of the experiments.

While this is a correct definition of a confidence interval, some statisticians prefer another. To quote Larry Wasserman,

> [The above definition] is correct but useless since we rarely repeat the same experiment over and over. A better interpretation is this: On day 1, you collect data and construct a 95 percent confidence interval for a parameter θ_1. On day 2, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_2. On day 3, you collect new data and construct a 95 percent confidence interval for an unrelated parameter θ_3. You continue this way constructing confidence intervals for a sequence of unrelated parameters $\theta_1, \theta_2, \ldots$. Then 95 percent of your intervals will trap the true parameter value. There us no need to introduce the idea of repeating the same experiment over and over.

In other words, the confidence interval describes the construction of the confidence interval itself. 95% of the time, it will contain the true (unknown) parameter value. Wasserman's description contains a reference to the *true* parameter value, so if you are going to talk about the true parameter value, his description is useful. However, the first definition of the confidence interval is quite useful if you want to think about how repeated experiments will end up.

We will use the first definition in thinking about how to construct a confidence interval. To construct the confidence interval, then, we will repeat the experiment over and over again, each time computing $\hat{\theta}$. We will then generate an ECDF of our $\hat{\theta}$ values, and report the 2.5th and 97.5th percentile to get our 95% confidence interval. But wait, how will we repeat the experiment so many times?

Remember that the data come from a probability distribution with CDF $F(x)$. Doing an experiment where we make n measurements amounts to drawing n numbers out of $F(x)$[15]. So, we could draw out of $F(x)$ over and over again. The problem

[15] We're being loose with language here. We're drawing out of the distribution that has CDF $F(x)$, but we're saying "draw out of F" for short.

is, we do now know what $F(x)$ is. However, we do have an empirical estimate for $F(x)$, namely $\hat{F}(x)$. So, we could draw n samples out of $\hat{F}(x)$, compute $\hat{\theta}$ from these samples, and repeat. This procedure is called **bootstrapping**.

To get the terminology down, a **bootstrap sample**, $\mathbf{x}^*$, is a set of n x values drawn from $\hat{F}(x)$. A **bootstrap replicate** is the estimate $\hat{\theta}^*$ obtained from the bootstrap sample $\mathbf{x}^*$. To generate a bootstrap sample, consider an array of measured values $\mathbf{x}$. We draw n values out of this array, *with replacement*. This is equivalent to sampling out of $\hat{F}(x)$.

So, the recipe for generating a bootstrap confidence interval is as follows.

1) Generate B independent bootstrap samples. Each one is generated by drawing n values out of the data array with replacement.
2) Compute $\hat{\theta}$ for each bootstrap sample to get the bootstrap replicates.
3) The $100(1-\alpha)$ percent confidence interval consists of the percentiles $100\,\alpha/2$ and $100(1-\alpha/2)$ of the bootstrap replicates.

This procedure works for any estimate $\hat{\theta}$, be it the mean, median, variance, skewness, kurtosis, or any other esoteric thing you can think of. Note that we use the empirical distribution, so there is never any assumption of an underlying "true" distribution. Thus, we are doing *nonparametric inference* on what we would expect for parameters coming out of unknown distributions; we only know the data. We will not discuss Bayesian nonparameterics, but they are generally not nearly as straightforward.[16] In this way, frequentist procedures are often useful in the nonparametric context.

There are plenty of subtleties and improvements to this procedure, but this is most of the story. We will discuss bootstrap confidence intervals for regression parameters in the tutorials, but we have already covered the main idea.

6.5 Hypothesis tests

The frequentist analog to model comparison is hypothesis testing. But we should be careful, it is an analog, but *most definitety not the same thing*. It is important to note that frequentist hypothesis testing is different from Bayesian model comparison in that in frequentist hypothesis tests, we will only consider how probable it is to get the observed data under a specific hypothesis, often called the **null hypothesis**. It is

[16] But Bayesian nonparametrics is a fascinating and useful field. The basic idea is that you have infinite dimensional priors over models and proceed with Bayesian inference from there. A new book on the subject, *Fundamentals of Nonparametric Baysian Inference*, by Ghosal and van der Vaart, is a good, complete reference.

just a name for the hypothesis you are testing. We will not assess other hypotheses nor compare them. Remember that the probability of a hypothesis being true is not something that makes any sense to a frequentist.

A frequentist hypothesis test consists of these steps.

1) Clearly state the null hypothesis.
2) Define a **test statistic**, a scalar value that you can compute from data. Compute it directly from your measured data.
3) *Simulate* data acquisition for the scenario where the null hypothesis is true. Do this many times, computing and storing the value of the test statistic each time.
4) The fraction of simulations for which the test statistic is at least as extreme as the test statistic computed from the measured data is called the **p-value**, which is what you report.

We need to be clear on our definition here. The p-value is the probability of observing a test statistic being at least as extreme as what was measured if the null hypothesis is true. It is exactly that, and nothing else. It is not the probability that the null hypothesis is true.

Importantly, **a hypothesis test is defined by the null hypothesis, the test statistic, and what it means to be at least as extreme.** That uniquely defines the hypothesis test you are doing. All of the named hypothesis tests, like the Student-t test, the Mann-Whitney U-test, Welch's test, etc., describe a specific hypothesis with a specific test statistic, with a specific definition of what it means to be at least as extreme (e.g., one-tailed or two-tailed). I can never remember what these are, nor do I encourage you to; you can always look them up. Rather, you should just clearly write out what your test is in terms of the hypothesis, test statistic, and definition of extreme.

Now, the real trick to doing a hypothesis test is step 3, in which you simulate the data acquisition assuming the null hypothesis was true. I will demonstrate two hypothesis tests and how we can simulate them. For both examples, we will consider the commonly encountered problem of performing the same measurements under two different conditions, control and test. You might have in mind the example of Dorsal gradient widths for wild type Dorsal versus those of the Dorsal-Venus construct.

Test and control come from the same distribution. Here, the null hypothesis is that the distribution F of the control measurements is the same as that G of the test, or $F = G$. To simulate this, we can do a **permutation test**. Say we have n measurements from control and m measurements from test. We then concatenate

arrays of the control and test measurements to get a single array with $n + m$ entries. We then randomly scramble the order of the entries (this is implemented in `np.random.permuation()`). We take the first n to be labeled "control" and the last m to be labeled "test." In this way, we are simulating the null hypothesis: whether or not a sample is test or control makes no difference.

For this case, we might define our test statistic to be difference of means, or difference of medians. These can be computed from the two data sets and are a scalar value.

Test and control have the same mean. The null hypothesis here is exactly as I have stated, and nothing more. To simulate this, we shift the data sets so that they have the same mean. In other words, if the control data are $\mathbf{x}$ and the test data are $\mathbf{y}$, then we define the mean of all measurements to be

$$\bar{z} = \frac{n\bar{x} + m\bar{y}}{n+m}. \tag{6.30}$$

Then, we define

$$x_{\text{shift},i} = x_i - \bar{x} + \bar{z}, \tag{6.31}$$

$$y_{\text{shift},i} = y_i - \bar{y} + \bar{z}. \tag{6.32}$$

$$\tag{6.33}$$

Now, the data sets $\mathbf{x}_{\text{shift}}$ and $\mathbf{y}_{\text{shift}}$ have the same mean, but everything else about them is the same as $\mathbf{x}$ and $\mathbf{y}$, respectively.

To simulate the null hypothesis, then, we draw bootstrap samples from $\mathbf{x}_{\text{shift}}$ and $\mathbf{y}_{\text{shift}}$ and compute the test statistic from the bootstrap samples, over and over again.

In both of these cases, no assumptions were made about the underlying distributions. Only the empirical distributions were used; these are nonparametric hypothesis tests.

6.5.1 Interpretation of the p-value

If the p-value is small, the effect is said to be **statistically significant**. But what is small? I strongly discourage a bright line p-value used to deem a result statistically significant or not. You computed the p-value, it has a specific meaning; you should report it. I do not see a need to convert a computed value, the p-value, into a Boolean, True/False on whether or not we attach the word "significant" to the result.

The question the p-value addresses is rarely the question we want to ask. For example, say we are doing a test of the null hypothesis that two sets of measurements

have the same mean. In most cases, which of the following questions are we interested in asking:

1) How different are the means of the two samples?
2) Would we say there is a statistically significant difference of the means of the two samples? Or, more precisely, what is the probability of observing a difference in means of the two samples at least as large as the the observed difference in means, if the two samples in fact have the same mean?

The second question is convoluted and often of little scientific interest. I would say that the first question is much more relevant. To put it in perspective, say we made trillions of measurements of two different samples and their mean differs by one part per million. This difference, though tiny, would still give a low p-value, and therefore often be deemed "statistically significant." But, ultimately, it is the size of the difference, or the *effect size* we care about.

6.5.2 What is with all those names?

You have no doubt heard of many named frequentist hypothesis tests, like the Student-t test, Welch's t-test, the Mann-Whitney U-test, and countless others. What is with all of those names? It helps to think more generally about how frequentist hypothesis testing is usually done.

To do a frequentist hypothesis test, people unfortunately do not do what I laid out above, but typically follow the following prescription (borrowing heavily from the treatment in Gregory's excellent book).

1) Choose a null hypothesis. This is the hypothesis you want to test the truth of.
2) Choose a suitable test statistic that can be computed from measurements *and* has a predictable distribution. For the example of two sets of repeated measurements, we can choose as our statistic

$$T = \frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{S_D\sqrt{n_1^{-1} + n_2^{-1}}},$$

$$\text{where } S_D^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

$$\text{with } S_1^2 = \frac{1}{n_1 - 1}\sum_{i \in D_1}(x_i - \bar{x}_1)^2, \tag{6.34}$$

and S_2^2 similarly defined. The T statistic is the difference of the difference of the observed means and the difference of the true means, weighted by the

spread in the data. This is a reasonable statistic for determining something about means from data. This is the appropriate statistic when σ_1 and σ_2 are both unknown but assumed to be equal. (When they are assumed to be unequal, you need to adjust the statistic you use. This test is called Welch's t-test.) It can be derived that this statistic has the Student-t distribution,

$$P(t) = \frac{1}{\sqrt{\pi\nu}} \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \left(\frac{t^2}{\nu}\right)\right)^{-\frac{\nu+1}{2}}, \tag{6.35}$$

$$\text{where } \nu = n_1 + n_2 - 2. \tag{6.36}$$

3) Evaluate the statistic from measured data. In the case of the Student-t test, we compute T.

4) Plot $P(t)$. The area under the curve where $t > T$ is the p-value, the probability that we would observe our data under the null hypothesis. Reject the null hypothesis if this is small.

As you can see from the above prescription, item 2 can be tricky. Coming up with test statistics *that also have a distribution that we can write down* is difficult. When such a test statistic is found, the test usually gets a name. The main reason for doing things this way is that most hypothesis tests were developed before computers, so we couldn't just bootstrap our way through hypothesis tests. (The bootstrap was invented by Brad Efron in 1979.) Conversely, in the approach we have taken, sometimes referred to as "hacker stats," we can invent any test statistic we want, and we can test is by numerically "repeating" the experiment, in accordance with the frequentist interpretation of probability.

So, I would encourage you not to get caught up in names. If someone reports a p-value with a name, simply look up the things you need to define the p-values (the hypothesis being tested, the test statistic, and what it means to be as extreme), and that will give you an understanding of what is going on with the test.

That said, many of the tests with names have analytical forms and can be rapidly computed. Most are included in the `scipy.stats` module. I have chosen to present a method of hypothesis testing that is intuitive with the frequentist interpretation of probability front and center. It also allows you to design your own tests that fit a null hypothesis that you are interested in that might not be "off-the-shelf."

6.5.3 Warnings about hypothesis tests

There are many.

1) An effect being statistically significant does not mean the effect is significant in practice or even important. It only means exactly what it is defined to mean:

an effect is unlikely to have happened by chance under the null hypothesis. Far more important is the **effect size**.

2) The p-value is **not** the probability that the null hypothesis is true. It is the probability of observing the test statistic being at least as extreme as what was measured if the null hypothesis is true. I.e., if H_0 is the null hypothesis,

$$\text{p-value} = P(\text{test stat at least as extreme as observed} \mid H_0). \tag{6.37}$$

It is not the probability that the null hypothesis is true given that the test statistic was at least as extreme as the data.

$$\text{p-value} \neq P(H_0 \mid \text{test stat at least as extreme as observed}). \tag{6.38}$$

We often actually want the probability that the null hypothesis is true, and the p-value is often erroneously interpreted as this to great peril.

3) Null hypothesis significance testing does not say anything about alternative hypotheses. Rejection of the null hypothesis does not mean acceptance of any other hypotheses.

4) P-values are not very reproducible, as we will see in the tutorials when we do "dance of the p-values."

5) Rejecting a null hypothesis is also kind of odd, considering you computed

$$P(\text{test stat at least as extreme as observed} \mid H_0). \tag{6.39}$$

This does not really describe the probability that the hypothesis is true. This, along with point 4, means that the p-value better be *really* low for you to reject the null hypothesis.

6) Throughout the literature, you will see null hypothesis testing when the null hypothesis is not relevant at all. People compute p-values because that's what they are supposed to do. The Dorsal gradient might be an example: of course the gradients will be different; we have made a big perturbation. We slapped a giant glowing barrel onto the Dorsal protein. Again, it gets to the point that **effect size** is waaaaay more important than a null hypothesis significance test.

Given all these problems with p-values, I generally advocate for their abandonment. I am not the only one. They seldom answer the question scientists are asking and lead to great confusion.

7 Introduction to images

This lecture was presented as a Jupyter notebook, available here.

8 Parallel tempering MCMC

In this lecture, we will discuss parallel tempering Markov chain Monte Carlo (PTMCMC). This technique allows for effective sampling of multimodal distributions and it avoids getting trapped on local maxima of the posterior. Perhaps even more importantly, it allows us to perform model selection.

8.1 The basic idea

Recall that the posterior distribution we seek to sample in the model selection problem is

$$g(\theta_i \mid D, M_i) \propto g(\theta_i \mid M_i) f(D \mid \theta_i, M_i). \tag{8.1}$$

Now, we define

$$g_{\text{hot}}(\theta_i \mid D, M_i, \beta) = \frac{1}{Z_i(\beta)}\, g(\theta_i \mid M_i)\, [f(D \mid \theta_i, M_i)]^{\beta} \tag{8.2}$$

$$= \frac{1}{Z_i(\beta)}\, g(\theta_i \mid M_i) \exp\left[\beta \ln f(D \mid \theta_i, M_i)\right]. \tag{8.3}$$

Here, $\beta \in (0, 1]$ is an "inverse temperature" in analogy to statistical mechanics, where the negative log likelihood, $-\ln f(D \mid \theta_i, M_i)$, is an energy. Keeping with the analogy, the normalization constant $Z_i(\beta)$, given by

$$Z_i(\beta) = \int \mathrm{d}\theta_i\, g(\theta_i \mid M_i)\, [f(D \mid \theta_i, M_i)]^{\beta}\,, \tag{8.4}$$

is called a **partition function**. We will call the distribution $g_{\text{hot}}(\theta_i \mid D, M_i, \beta)$ a **hot posterior** because it is the posterior with a high temperature.

If β is close to zero (the "high temperature" limit), we are just sampling the prior. If $\beta = 1$, we are sampling our target posterior, the so-called "cold distribution." So, lowering β has the effect of flattening the posterior distribution. Therefore, walkers at higher temperature (lower β) are not trapped at local maxima. By occasionally swapping walkers from adjacent temperatures, we can effectively sample a broader swath of parameter space.

In practice, we choose a set of β's with $\beta = [\beta_0, \beta_1, \ldots, \beta_m]$, with $\beta_{i+1} < \beta_i$ and $\beta_0 = 1$. We propose a swap roughly every n_s steps and accept it based on criteria that guarantees the posterior is a stationary distribution of the transition kernel. To do this in practice, we choose a uniform random number on $[0, 1]$ every iteration and propose a swap when this random number is less than $1/n_s$. When we do propose a

swap, we randomly pick a temperature β_j from $\{\beta_1, \beta_2, \ldots \beta_m\}$. We then compute

$$r = \min\left(1, \frac{g_{\text{hot}}(\theta_{i,j} \mid D, M_i, \beta_{j-1})}{g_{\text{hot}}(\theta_{i,j-1} \mid D, M_i, \beta_{j-1})} \frac{g_{\text{hot}}(\theta_{i,j-1} \mid D, M_i, \beta_j)}{g_{\text{hot}}(\theta_{i,j} \mid D, M_i, \beta_j)}\right). \tag{8.5}$$

Here, we have defined $\theta_{i,j}$ as the value of parameter i for a walker at temperature β_j. Note that this calculation does not require calculation of any partition functions; the $Z_i(\beta)$ cancel out in the expression for r. We then draw another uniform random number on $[0, 1]$ and accept the swap is that number if less than r.

This useful technique is implemented the package ptemcee (pronouced tem-see; the p is silent). Conveniently, it automatically chooses reasonable values of β and swapping rate, though you can specify these as well. It also has a bit more sophistication that what I have described here, using adaptive parallel tempering.

8.2 Model selection with PTMCMC

We will now do some clever ticks to see how we can use PTMCMC to do model comparison without making the approximations we have thus far. In fact, we do not necessarily need parallel tempering with swapping; we only need samples of $g_{\text{hot}}(\theta_i \mid D, M_i, \beta)$ for various values of β. Recall the statement of Bayes's theorem for the model comparison problem, equation (5.3).

$$g(M_i \mid D) = \frac{f(D \mid M_i)\, g(M_i)}{f(D)}. \tag{8.6}$$

The likelihood in the model selection problem is given by the evidence, a.k.a. fully marginalized likelihood, from the parameter estimation problem, as we derived in equation (5.5). Thus,

$$g(M_i \mid D) = \frac{g(M_i)}{f(D)} \left[\int \mathrm{d}\theta_i\, g(\theta_i \mid M_i) f(D \mid \theta_i, M_i)\right]. \tag{8.7}$$

We recognize the bracketed term as $Z_i(1)$. Our goal is to calculate this quantity.

Now, we're going to do a usual trick in statistical mechanics: we will differentiate the log of the partition function (analogous to the derivative of a free energy).

$$\begin{aligned}
\frac{\partial}{\partial\beta} \ln Z_i(\beta) &= \frac{1}{Z_i(\beta)} \frac{\partial Z_i}{\partial\beta} \\
&= \frac{1}{Z_i(\beta)} \int \mathrm{d}\theta_i \frac{\partial}{\partial\beta} \exp\left[\ln g(\theta_i \mid M_i) + \beta \ln f(D \mid \theta_i, M_i)\right] \\
&= \frac{1}{Z_i(\beta)} \int \mathrm{d}\theta_i \ln f(D \mid \theta_i, M_i) \exp\left[\ln g(\theta_i \mid M_i) + \beta \ln f(D \mid \theta_i, M_i)\right]
\end{aligned}$$

$$= \frac{1}{Z_i(\beta)} \int \mathrm{d}\theta_i \, \ln f(D \mid \theta_i, M_i) \, g(\theta_i \mid M_i) \, [f(D \mid \theta_i, M_i)]^{\beta} \, . \tag{8.8}$$

We recognize this as the average of the log likelihood $\ln f(D \mid \theta_i, M_i)$ over the distribution $g_{\mathrm{hot}}(\theta_i \mid D, M_i, \beta)$. We denote this as

$$\frac{\partial}{\partial \beta} \ln Z_i(\beta) = \langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)} \, . \tag{8.9}$$

Note that this average if done for each specific value of β we are considering, and that the derivative of the log partition function is thus a function of β. Now, we can integrate both sizes of this equation to give

$$\begin{aligned} \int_0^1 \mathrm{d}\beta \, \frac{\partial}{\partial \beta} \ln Z_i(\beta) &= \ln Z_i(1) - \ln Z_i(0) \\ &= \int_0^1 \mathrm{d}\beta \, \langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)} \, . \end{aligned} \tag{8.10}$$

Now, if the prior is normalized, as it should be,

$$Z_i(0) = \int \mathrm{d}\theta_i \, g(\theta_i \mid M_i) = 1, \tag{8.11}$$

which means $\ln Z_i(0) = 0$. Thus, we get a fully marginalized likelihood of

$$\begin{aligned} \ln Z_i(1) &= \int \mathrm{d}\theta_i f(D \mid \theta_i, M_i) \, g(\theta_i \mid M_i) \\ &= \int_0^1 \mathrm{d}\beta \, \langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)} \, . \end{aligned} \tag{8.12}$$

Fortunately, if we have done PTMCMC, we have sampled out of the distribution $g_{\mathrm{hot}}(\theta_i \mid D, M_i, \beta)$ for various values of β. We can then compute the integrand in the above equation for each β at which we sampled.

$$\langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)} = \frac{1}{n_{\mathrm{samples}}} \sum_{\mathrm{samples}} \ln f(D \mid \theta_i, M_i). \tag{8.13}$$

We just have to compute the log likelihood (*not* the hot log-likelihood) for each MCMC sample for a given inverse temperature β, and we have all we need. We then perform numerical quadrature across the values of β that we sampled to get the integral. We therefore can compute the odds ratio of two models M_i and M_j as

$$O_{ij} = \frac{g(M_i \mid I)}{g(M_j \mid I)} \frac{Z_i(1)}{Z_j(1)} \tag{8.14}$$

$$= \frac{g(M_i \mid I)}{g(M_j \mid I)} \exp \left[\frac{\int_0^1 \mathrm{d}\beta \ \langle \ln f(D \mid \theta_i, M_i) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)}}{\int_0^1 \mathrm{d}\beta \ \langle \ln f(D \mid \theta_j, M_j) \rangle_{g_{\mathrm{hot}}(\theta_i | D, M_i, \beta)}} \right],$$

where the last ratio is computed via numerical quadrature on results computed directly from our PTMCMC traces using equation (8.13). Note that we have made no approximations at all in the model. The calculation is only approximate to the extent that the PTMCMC sampler takes a finite number of samples and numerical quadrature is not exact.

9 Hierarchical models

In this lecture, we will investigate **hierarchical models**, in which some model parameters are dependent on others in specific ways. This is best learned by example.

In homework problem 5.2, we studied reversals under exposure to blue light in *C. elegans* with Channelrhodopsin in two different neurons. Let's consider one of the strains which contains a Channelrhodopsin in the ASH sensory neuron. The experiment was performed three times by the students of Bi 1x. In 2015, we found that 9 out of 35 worms reversed under exposure to blue light. In 2016, 12 out of 35 reversed. In 2017, 18 out of 54 reversed.

Considering for a moment only the 2015 experiment, we can use this measurement to estimate the probability p of reversal. We modeled the likelihood of reversal with a Binomial likelihood. Taking a uniform prior on p, we derived that the posterior probability of reversal given r our of n trials showed reversals was

$$g(p \mid r, n) = \begin{cases} \dfrac{(n+1)!}{(n-r)!r!}\, p^r (1-p)^{n-r} & 0 \le p \le 1 \\ 0 & \text{otherwise.} \end{cases} \tag{9.1}$$

We did the experiment again in 2016, getting $r = 12$ and $n = 35$, and in 2017 with $r = 18$ and $n = 54$. Actually, we could imagine doing the experiment over and over again, say k times, each time getting a value of r and n. Conditions may change from experiment to experiment. For example, we may have different lighting set-ups, slight differences in the strain of worms we're using, etc. We are left with some choices on how to model the data.

9.1 Pooled data: identical parameters

We could pool all of the data together. In other words, let's say we measure r_1 out of n_1 reversals in the first set of experiments, r_2 out of n_2 reversals in the second set, etc., up to k total experiments. We could pool all of the data together to get

$$\begin{aligned} r &= \sum_{i=1}^{k} r_i \\ \text{out of } n &= \sum_{i=1}^{k} n_i \text{ reversals.} \end{aligned} \tag{9.2}$$

We then compute our posterior as in equation (9.1). Here, the assumption is that the result in each experiment are governed by *identical parameters*. That is to say that we assume $p_1 = p_2 = \cdots = p_k = p$.

This is similar to what we did in section 1.9, in which we looked at how a single hypothesis (or parameter value) is informed by more data.

9.2 Independent parameters

As an alternative, we could instead say that the parameters in each experiment are totally independent of each other. In this case, we assume that $p_1, p_2, \ldots, p_k$ are all independent of each other. The likelihoods and priors all multiply and the posterior probability is

$$g(\mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \prod_{i=1}^{k} \frac{(n_i + 1)!}{(n_i - r_i)! r_i!} p_i^{r_i} (1 - p_i)^{n_i - r_i}, \tag{9.3}$$

where $\mathbf{p} = \{p_1, p_2, \ldots p_k\}$, with $\mathbf{n}$ and $\mathbf{r}$ similarly defined, and the posterior is understood to be zero if any the p_i's fall out of the interval $[0, 1]$.

When we make this assumption, we often report a value of p that is given by the mean of the p_i's with some error bar.

9.3 Best of both worlds: a hierarchical model

Each of these extremes have their advantages. We are often trying to estimate a parameter that is more universal than our experiments, e.g., something that describes worms with Channelrhodopsin in the ASH neuron generally. So, pooling the experiments makes sense. On the other hand, we have reason to assume that there is going to be a different value of p in different experiments, as biological systems are highly variable, not to mention measurement variations. So, how can we capture both of these effects?

We can consider a model in which there is a "master" reversal probability, which we will call q, and the values of p_i may vary from this q according to some probability distribution, $g(p_i \mid q)$. So now, we have parameters $p_1, p_2, \ldots, p_k$ and q. So, the posterior can be written using Bayes's theorem,

$$g(q, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid q, \mathbf{p})\, g(q, \mathbf{p})}{f(\mathbf{n}, \mathbf{r})}. \tag{9.4}$$

Note, though, that the observed values of r do not depend directly on q, only on $\mathbf{p}$. In other words, the observations are only *indirectly* dependent on q. So, we can write $f(\mathbf{r}, \mathbf{n} \mid q, \mathbf{p}) = f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})$. Thus, we have

$$g(q, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})\, g(q, \mathbf{p})}{f(\mathbf{n}, \mathbf{r})}. \tag{9.5}$$

Next, we can rewrite the prior using the definition of conditional probability.

$$g(q, \mathbf{p}) = g(\mathbf{p} \mid q)\, g(q). \tag{9.6}$$

Substituting this back into our expression for the posterior, we have

$$g(q, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})\, g(\mathbf{p} \mid q)\, g(q)}{f(\mathbf{n}, \mathbf{r})}. \tag{9.7}$$

Now, if we read off the numerator of this equation, we see a chain of dependencies. The experimental results $\mathbf{r}$ depend on parameters $\mathbf{p}$. Parameters $\mathbf{p}$ depend on *hyperparameter* q. Hyperparameter q then has some **hyperprior** distribution. Any model that can be written as a chain of dependencies like this is called a **hierarchical model**, and the parameters that do not *directly* influence the data are called **hyperparameters**.

So, the hierarchical model captures both the experiment-to-experiment variability, as well as the master regulator of outcomes. Note that the product $g(\mathbf{p} \mid q)\, g(q)$ comprises the prior, as it is independent of the observed data.

9.4 Exchangeability

The conditional probability, $g(\mathbf{p} \mid q)$, can take any reasonable form. In the case where we have no reason to believe that we can distinguish any one p_i from another prior to the experiment, then the label "i" applied to the experiment may be exchanged with the label of any other experiment. I.e., $g(p_1, p_2, \ldots, p_k \mid q)$ is invariant to permutations of the indices. Parameters behaving this way are said to be **exchangeable**. A common (simple) exchangeable distribution is

$$g(\mathbf{p} \mid q) = \prod_{i=1}^{k} g(p_i \mid q), \tag{9.8}$$

which means that each of the parameters is an independent sample out of a distribution $g(p_i \mid q)$, which we often take to be the same for all i. This is reasonable to do in the worm reversal example.

9.5 Choice of the conditional distribution

We need to specify our prior, which for this hierarchical model means that we have to specify the conditional distribution, $g(p_i \mid q)$, as well as $g(q)$ For the latter, we will take it to be uniform on $[0, 1]$. This is equivalent to taking it to be a Beta distribution with $\alpha = \beta = 1$. The Beta distribution is a good choice in this case, as it is a probability distribution of probabilities. For the conditional distribution $g(p_i \mid q)$, we also assume it is Beta-distributed.

The Beta distribution is typically written as

$$g(p \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \tag{9.9}$$

where it is parametrized by positive constants α and β. The Beta distribution has mean and **concentration**, respectively, of

$$q = \frac{\alpha}{\alpha+\beta}, \tag{9.10}$$

$$\kappa = \alpha + \beta. \tag{9.11}$$

The concentration κ is a measure of how sharp the distribution is. The bigger κ is, the most sharply peaked the distribution is.

Because the Beta distribution has two parameters, we cannot just parametrize the model with q. We would have to use q and κ or alternatively α and β. So, our expression for the posterior is

$$g(\alpha, \beta, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})\, g(\alpha, \beta) \prod_{i=1}^{k} g(p_i \mid \alpha, \beta)}{f(\mathbf{n}, \mathbf{r})}. \tag{9.12}$$

Alternatively, we could parametrize the model in terms of q and κ, giving

$$g(q, \kappa, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) = \frac{f(\mathbf{r}, \mathbf{n} \mid \mathbf{p})\, g(q, \kappa) \prod_{i=1}^{k} g(p_i \mid q, \kappa)}{f(\mathbf{n}, \mathbf{r})}. \tag{9.13}$$

Note that if we do choose to parametrize our model with q and κ, we can convert back to α and β using

$$\alpha = q\kappa \tag{9.14}$$

$$\beta = (1-q)\kappa. \tag{9.15}$$

9.6 Choice of prior

As already stated, the likelihood is Binomial, with

$$f(\mathbf{r}, \mathbf{n} \mid \mathbf{p}) = \prod_{i=1}^{k} f(r_i, n_i \mid p_i) = \prod_{i=1}^{k} \frac{n_i!}{r_i!(n_i - r_i)!} p_i^{r_i}(1-p_i)^{n_i - r_i}, \tag{9.16}$$

and $g(p_i \mid \alpha, \beta)$ is Beta distributed, with

$$g(p_i \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1}(1-p_i)^{\beta-1}. \tag{9.17}$$

We are now left to specify the hyperprior, $g(\alpha, \beta)$. We might choose to specify the prior in terms of q and κ, since these seem at face to be more intuitive. We can take a Uniform prior for q and a Jeffreys prior for κ, as we often do. That is, $g(q, \kappa) \propto 1/\kappa$. Applying the change of variables formula, we have

$$g(\alpha, \beta) \propto \left| \begin{matrix} \frac{\partial q}{\partial \alpha} & \frac{\partial q}{\partial \beta} \\ \frac{\partial \kappa}{\partial \alpha} & \frac{\partial \kappa}{\partial \beta} \end{matrix} \right| \frac{1}{\alpha + \beta} = \left| \begin{matrix} \frac{\beta}{(\alpha+\beta)^2} & -\frac{\alpha}{(\alpha+\beta)^2} \\ 1 & 1 \end{matrix} \right| \frac{1}{\alpha + \beta} = \frac{1}{(\alpha + \beta)^2}. \tag{9.18}$$

So, a uniform prior for q and a Jeffreys prior for κ results in a uniform prior in α and β, defined on $\alpha, \beta \in (0, \text{inf})$. If we use this Uniform prior, we have

$$\begin{aligned} g(\alpha, \beta, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) &\propto \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{\alpha-1}(1 - p_i)^{\beta-1} p_i^{r_i} (1 - p_i)^{n_i - r_i} \\ &\propto \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{r_i+\alpha-1} (1 - p_i)^{n_i - r_i + \beta - 1}. \end{aligned} \tag{9.19}$$

We can integrate the right hand side over $p_1, p_2, \ldots$ to get the marginalized posterior for the hyperparameters α and β. We can do the integral by inspection, noting that $p_i^{r_i+\alpha-1}(1 - p_i)^{n_i - r_i + \beta - 1}$ is the same functional form of an unnormalized Beta distribution, so we must have

$$\int_0^1 \mathrm{d}p_i \, p_i^{r_i+\alpha-1}(1 - p_i)^{n_i - r_i + \beta - 1} = \frac{\Gamma(r_i + \alpha)\Gamma(n_i - r_i + \beta)}{\Gamma(n_i + \alpha + \beta)}. \tag{9.20}$$

So, the unnormalized marginalized posterior is

$$g(\alpha, \beta \mid \mathbf{r}, \mathbf{n}) \propto \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(r_i + \alpha)\Gamma(n_i - r_i + \beta)}{\Gamma(n_i + \alpha + \beta)}. \tag{9.21}$$

There is a problem with this posterior: it is improper. That is to say that it is unnormalizable. This can be seen by using the reciprocal relation for gamma functions, $x\Gamma(x) = \Gamma(x + 1)$ to re-write the marginalized posterior.

$$\begin{aligned} g(\alpha, \beta \mid \mathbf{r}, \mathbf{n}) &\propto \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\left(\prod_{m=0}^{r_i-1}(\alpha + m)\right)\left(\prod_{m=0}^{n_i-r_i-1}(\beta + m)\right)}{\prod_{m=0}^{n-1}(\alpha + \beta + m)} \\ &= \frac{1}{(\alpha + \beta)^2} \prod_{i=1}^{k} \frac{\mathcal{O}(\alpha^{r_i})\,\mathcal{O}(\beta^{n_i-r_i})}{\mathcal{O}\left((\alpha + \beta)^{n_i}\right)} \end{aligned}$$

$$= \frac{1}{(\alpha+\beta)^2} \prod_{i=1}^{k} \mathcal{O}\left(\left(\frac{\alpha}{\alpha+\beta}\right)^{r_i}\right) \mathcal{O}\left(\left(\frac{\beta}{\alpha+\beta}\right)^{n_i-r_i}\right). \tag{9.22}$$

Since we have $q = \alpha/(\alpha+\beta) = (1+\beta/\alpha)^{-1}$ and must lie between zero and one, we can consider a limit of large α and β with the ratio α/β fixed at some constant, finite value. Then, for large α and β, the product term in the expression for the unnormalized marginal posterior is constant. Therefore, the integral

$$\int_0^\infty \mathrm{d}\alpha \int_0^\infty \mathrm{d}\beta \; g(\alpha, \beta \mid \mathbf{r}, \mathbf{n}) \tag{9.23}$$

diverges because the integral over $(\alpha+\beta)^{-2}$ diverges. This gives an improper *posterior*, which is not acceptable.

It turns out that this problem occurs generally in hierarchical models. The variance of a Beta distribution is approximately proportional to κ^{-1}, especially at large α and β. By choosing a Jeffreys prior for the variance, we are choosing a Uniform prior for the log of the variance. When we do this with hierarchical models, that is choose a Uniform prior for the log of the variance of a hyperprior for exchangeable parameters, we get an improper posterior.

So, it is often tricky to be truly uninformative with your hyperpriors. For the present example, we will instead choose a Uniform prior in the standard deviation, so that $\kappa^{-1/2}$ has a Uniform prior; $g(q, \kappa^{-1/2}) = \text{constant}$. If we do this, we have

$$g(\alpha, \beta) \propto \begin{vmatrix} \frac{\partial q}{\partial \alpha} & \frac{\partial q}{\partial \beta} \\ \frac{\partial \sqrt{\kappa}}{\partial \alpha} & \frac{\partial \sqrt{\kappa}}{\partial \beta} \end{vmatrix} = \begin{vmatrix} \frac{\beta}{(\alpha+\beta)^2} & -\frac{\alpha}{(\alpha+\beta)^2} \\ -\frac{1}{2(\alpha+\beta)^{3/2}} & -\frac{1}{2(\alpha+\beta)^{3/2}} \end{vmatrix} \propto \frac{1}{(\alpha+\beta)^{5/2}}. \tag{9.24}$$

With this prior, we have an unnormalized posterior of

$$g(\alpha, \beta, \mathbf{p} \mid \mathbf{r}, \mathbf{n}) \propto \frac{1}{(\alpha+\beta)^{5/2}} \prod_{i=1}^{k} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p_i^{r_i+\alpha-1} (1-p_i)^{n_i-r_i+\beta-1}. \tag{9.25}$$

This is a proper posterior, which you can prove with similar arguments as we made to show that the first posterior we considered was *im*proper.

9.7 Implementation

In some cases, we can do some gnarly integration and work out analytical results for the posterior of a hierarchical model. This usually involves choosing conjugate priors. Most often, though, we need to resort to numerical methods, MCMC as usual being the most powerful. To see the worm reversal problem solved with a hierarchical model, see the implementation here.

9.8 Generalization

The worm reversal problem is easily generalized. You can imagine having more levels of the hierarchy. This is just more steps in the chain of dependencies that are factored in the prior. For general parameters θ and hyperparameters ϕ, we have

$$g(\theta, \phi \mid D) = \frac{f(D \mid \theta)\, g(\theta \mid \phi)\, P(\phi)}{f(D)} \tag{9.26}$$

for a two-level hierarchical model. For a three-level hierarchical model, we can consider hyperparameters ξ that depend on ϕ, giving

$$g(\theta, \phi, \xi \mid D) = \frac{f(D \mid \theta)\, g(\theta \mid \phi)\, g(\phi \mid \xi)\, g(\xi)}{f(D)}, \tag{9.27}$$

and so on for four, five, etc., level hierarchical models. As we have seen in the course, the work is all in coming up with the models for the likelihood $f(D \mid \theta)$, and prior, $g(\theta \mid \phi)\, g(\phi)$, in this case for a two-level hierarchical model. For coming up with the conditional portion of the prior, $g(\theta \mid \phi)$, we often assume a Gaussian distribution because this often describes experiment-to-experiment variability. (The Beta distribution we used in our example is approximately Gaussian and has the convenient feature that it is defined on the interval $[0, 1]$.) Bayes's theorem gives you the posterior, and it is then "just" a matter of computing it by sampling from it.