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.

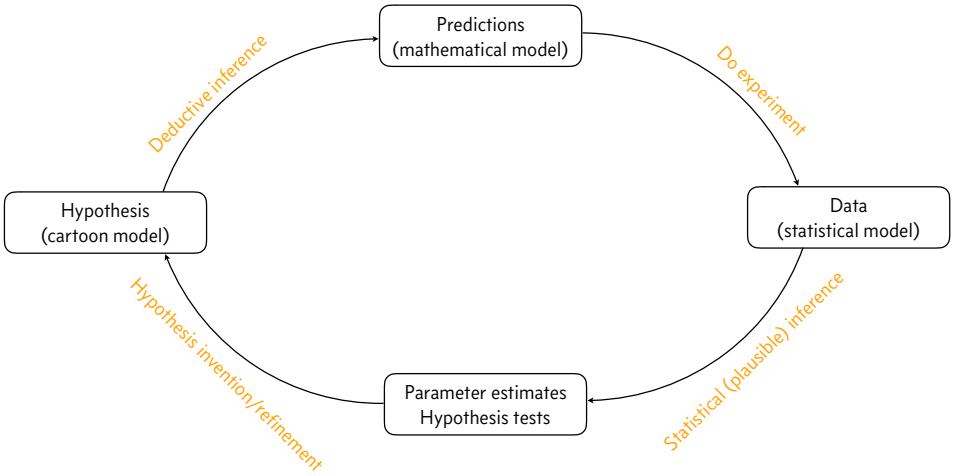

Figure 1: A sketch of the scientific process. Adapted from Fig. 1.1 of P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge, 2005.

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.