BE/Bi 103: Data Analysis in the Biological Sciences

Justin Bois

Caltech

Fall, 2018

Caltech

Donna and Benjamin M.
Rosen Bioengineering Center

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

1 Probability 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.

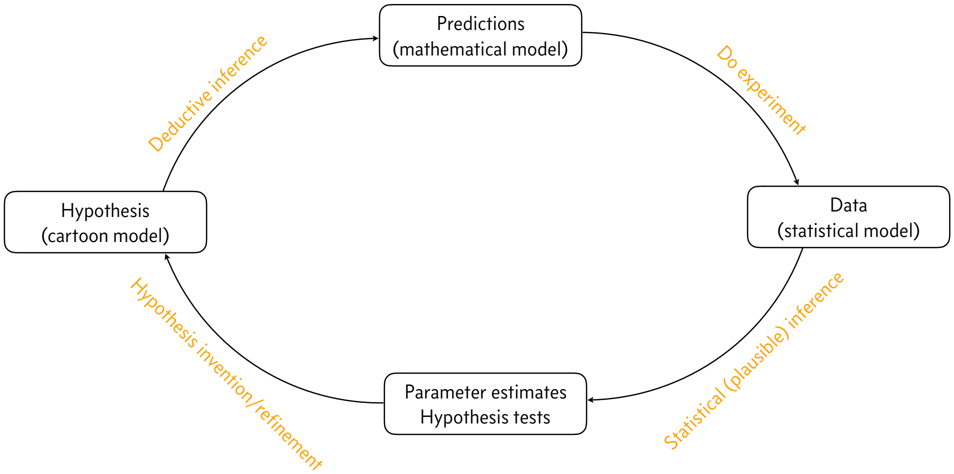

Figure 1: A sketch of the scientific process. Adapted from Fig. 1.1 of P. Gregory, *Bayesian Logical Data Analysis for the Physical Sciences*, Cambridge, 2005.

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 studies 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.

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.

So, we collected observations (we can call observations "data") to help us learn something about what generated the observations. That is, we are interested in the effect fracking has on earthquakes, and we collected information about earthquakes to try to infer the effects of fracking.

Because the connection between our observations and the process generating them is a weak syllogism, we cannot say anything with absolute certainty about the generative process. So, we need a way to quantify the uncertainty. 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 definition of probability

I will be a little formal[1] for a moment here as we construct this mathematical notion of probability. First, we need to define the world of possibilities. We denote by Ω a **sample space**, which is the set of all **outcomes** we could observe in a given experiment. We define an **event** A to be a subset of Ω ($A \subseteq \Omega$). Two events, A_i and A_j are **disjoint**, also called **mutually exclusive**, if $A_i \cap A_j = \emptyset$. That is to say that two events are disjoint if they do not overlap at all in the sample space; they do not share any outcomes. So, in common terms, the sample space Ω contains all possible outcomes of an experiment. An event A is a given outcome or set of outcomes. Two events are disjoint if they are totally different from each other.

We define the **probability of event** A to be $P(A)$, where P is a **probability function**. It maps the event A to a real number between zero and one. In order to be a probability, the function P must satisfy the following axioms.

1) The probability must be nonnegative; $P(A) \geq 0$ for all A.
2) The probability that an event was drawn from the entire sample space is one; $P(\Omega) = 1$.
3) The probability of the empty set is zero; $P(\emptyset) = 0$. Along with the previous axiom and the requirement that $P(A)$ range from zero to one, this essentially says that only events in the sample space are allowable outcomes.
4) If $A_1, A_2, \ldots$ are disjoint events, then

$$P\left(\bigcup_i A_i\right) = \sum_i P(A_i). \tag{1.1}$$

   This means that probability is additive. The probability of observing an event in the union of disjoint events is the sum of the probabilities of those events.

[1] But not too formal. For example, we are not discussing σ algebras, measurability, etc.

Putting together these axioms, we see that probability consists of positive real numbers that are distributed among the events of a sample space. The sum total of these real numbers over all of the sample space is one. So, a probability function and a sample space go hand-in-hand.

1.3 Interpretations of probability

Before we go on to talk more about probability, it will help to be thinking about how we can apply it to understand measured data. To do that, we need to think about how probability is interpreted. Note that these are *interpretations* of probability, not definitions. We have already defined probability, and both of the two dominant interpretations below are valid.

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.[2] So in the frequentist view, we are using probability to understand how the results of an experiment might vary from repetition to repetition.

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, not just a random variable.

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 uncertainty. 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 (they are not random variables; they cannot vary meaningfully from experiment to experiment), and we can only study how repeated experiments will convert the real parameter value to an observation.

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

We will use some frequentist approaches in class, especially when we study *non-parametric* methods, but we will generally focus on Bayesian analysis.

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

The *sum rule*, which may be derived from the axioms defining probability, says that the probability of *all* events must add to unity. Let A^c be all events *except* A, called the **complement** of A. Then, the sum rule states that

$$P(A) + P(A^c) = 1. \tag{1.2}$$

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

$$\textbf{(sum rule)} \qquad P(A \mid B) + P(A^c \mid B) = 1, \tag{1.3}$$

for any B.

The *product rule* states that

$$P(A, B) = P(A \mid B)\, P(B), \tag{1.4}$$

where $P(A, B)$ is the probability of both A *and* B happening. (It could be written as $P(A \cap B)$.) 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.

$$\textbf{(product rule)} \qquad P(A, B \mid C) = P(A \mid B, C)\, P(B \mid C), \tag{1.5}$$

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 and B.

$$A = \text{hypothesis (or parameter value), } \theta, \tag{1.6}$$

$$B = \text{Measured data set, } y. \tag{1.7}$$

So, we may be interested in the probability of obtaining a data set y given some set of parameters θ. In other words, we want to learn about $P(y \mid \theta)$.

To go a bit further, let's rewrite the product rule using our data set y and parameter θ.

$$P(y, \theta) = P(\theta \mid y)\, P(y). \tag{1.8}$$

Ahoy! The quantity $P(\theta \mid y)$ is exactly what we want from our statistical inference. This describes probability for values of a parameter, given measurements.

But wait a minute. The parameter θ is not something that can vary meaningfully from experiment to experiment; it is not a random variable. So, in the frequentist picture, we cannot assign a probability to it. That is, $P(\theta \mid y)$ and $P(y, \theta)$ do not make any sense. So, in the frequentist perspective, we can really only analyze $P(y \mid \theta)$.

Nonetheless, we proceed assuming we take a Bayesian interpretation of probability and discuss how we might get a useful expression for $P(\theta \mid y)$.

1.6 Bayes's Theorem

Note that because "and" is commutative, $P(y, \theta) = P(\theta, y)$. We apply the product rule to both sides of this seemingly trivial equality.

$$P(\theta \mid y)\, P(y) = P(\theta, y) = P(y, \theta) = P(y \mid \theta)\, P(\theta). \tag{1.9}$$

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

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

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.

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. We just chose to have θ take a meaning of a parameter value in the above example, but Bayes's theorem holds in general for any events that can be assigned a probability.

The quantities on the right hand side of Bayes's theorem 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, $P(\theta)$. As probability is a measure of plausibility, or how believable a hypothesis is. This represents the plausibility about hypothesis or parameter set θ given everything we know *before* we did the experiment to get the data.

The likelihood. The likelihood, $P(y \mid \theta)$, describes how likely it is to acquire the observed data, *given the hypothesis or parameter value* θ. 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.[3]

The posterior probability. This is what we are after. How plausible is the hypothesis or parameter value, given that we have measured some 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.

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. Let θ_i be a particular possible value of a parameter or hypothesis. Then,

$$\begin{aligned} 1 &= P(\theta_j \mid y) + P(\theta_j^c \mid y) \\ &= P(\theta_j \mid y) + \sum_{i \neq j} P(\theta_i \mid y) \\ &= \sum_i P(\theta_i \mid y). \end{aligned} \tag{1.12}$$

Now, Bayes's theorem gives us an expression for $P(\theta_i \mid y)$, so we can compute the sum.

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

[3] 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.

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

$$P(y) = \sum_i P(y \mid \theta_i)\, P(\theta_i). \tag{1.14}$$

Using the definition of conditional probability, we also have

$$P(y) = \sum_i P(y, \theta_i) \tag{1.15}$$

This process of eliminating a variable (in this case the hypotheses θ_i) in the joint distribution by summing is called **marginalization**. This will prove useful in finding the probability distribution of a single parameter among many, as you will show in your homework.

1.8 Probability distributions

So far we have talked about probability of events, and we have in mind measurements and, in the Bayesian case, parameter values as the events. We have a bit of a problem, though, if the sample space consists of real numbers, which we often encounter in our experiments and modeling. The probability of getting a single real value is identically zero. This is my motivation for introducing **probability distributions**, but the concept is more general and has much more utility than just dealing with sample spaces containing real numbers. Importantly, probability distributions provide the link between outcomes in the sample space to probability. Probability distributions describe both **discrete** quantities (like integers) and **continuous** quantities (like real numbers).

Though we cannot assign a nonzero the probability for an outcome from a sample space containing all of the real numbers, we can assign a probability that the outcome is less than some real number. Notationally, we write this as

$$P(\text{having outcome that is } \le y) = F(y). \tag{1.16}$$

The function $F(y)$, which returns a probability, is called a **cumulative distribution function** (CDF), or just **distribution function**. It contains all of the information we need to know about how probability is assigned to y. A CDF for a Gaussian distribution (which we will discuss in coming weeks) is shown in Fig. 2a.

Related to the CDF for a continuous quantity is the **probability density function**, or PDF. The PDF is given by the derivative of the CDF,

$$f(y) = \frac{\mathrm{d}F(y)}{\mathrm{d}y}. \tag{1.17}$$

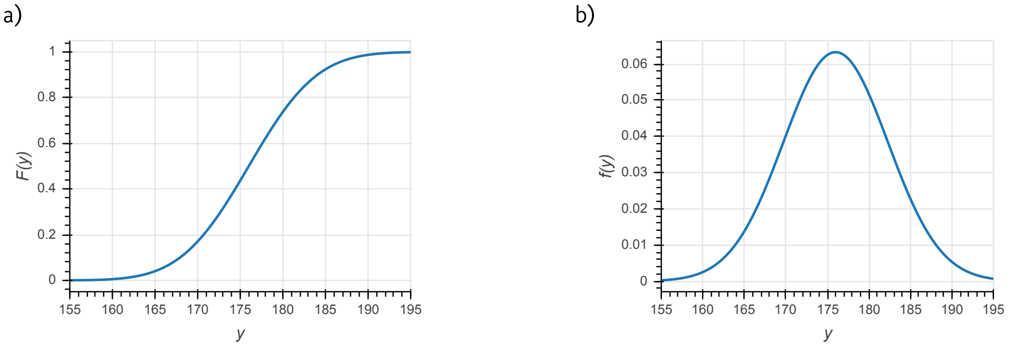

Figure 2: a) The cumulative distribution function for a Gaussian distribution that could describe, for example, the heights of men in centimeters in a given country. b) The corresponding probability distribution function.

Note that $f(y)$ is *not* the probability of outcome y. Rather, the probability that of outcome y lying between y_0 and y_1 is

$$P(y_0 \le y \le y_1) = F(y_1) - F(y_0) = \int_{y_0}^{y_1} \mathrm{d}y\, f(y). \tag{1.18}$$

Conversely, for a discrete quantity, we have a **probability mass function**, or PMF,

$$f(x) = P(x). \tag{1.19}$$

The PMF is a probability, unlike the PDF. An example of a CDF and a PMF for a discrete distribution are shown in Fig. 3. In this example, n is the outcome of the roll of a fair die ($n \in \{1, 2, 3, 4, 5, 6\}$).

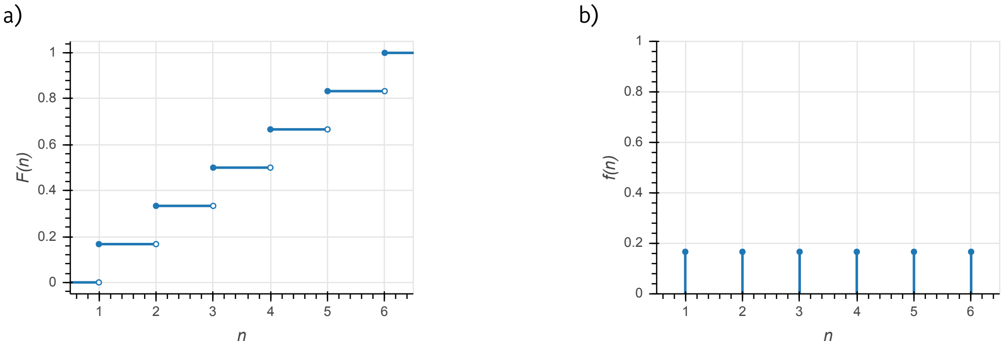

Figure 3: a) The cumulative distribution function for the outcome of a fair dice roll. b) The corresponding probability mass function.

1.9 Joint and conditional distributions and Bayes's theorem for PDFs

We have defined a PDF as $f(x)$, that is, describing a single variable x. We can have **joint distributions** with a PDF $f(x, y)$. The joint CDF is not well-defined, so we restrict our discussions of CDFs to the univariate case.

We may also have **conditional distributions** that have PDF $f(x \mid y)$. This is interpreted similarly to conditional probabilities we have already seen. $f(x \mid y)$ is the probability density function for x, *given* y. As similar relation between joint and conditional PDFs holds as in the case of joint and conditional probabilities.

$$f(x \mid y) = \frac{f(x, y)}{f(y)}. \tag{1.20}$$

That this holds is not at all obvious. One immediate issues is that we are conditioning on an event y that has zero probability. We will not carefully derive why this holds, but state it without proof.

As a consequence, Bayes's theorem also holds for PDFs, as it does for probabilities.[4]

$$f(\theta \mid y) = \frac{f(y \mid \theta) f(\theta)}{f(y)}. \tag{1.21}$$

Notationally in this course, we will use f to describe a PDF or PMF of a random variable. and g to describe the PMF or PDF of a parameter or other logical conjecture that is not measured data or a random variable. For example, $f(y)$ is the PDF for a continuous measured quantity and $g(\theta)$ is the PDF for a parameter value. So, Bayes's theorem is

$$g(\theta \mid y) = \frac{f(y \mid \theta)\, g(\theta)}{f(y)}. \tag{1.22}$$

Finally, we can marginalize probability distribution functions to get **marginalized PDFs**.

$$f(x) = \int \mathrm{d}y\, f(x, y) = \int \mathrm{d}y\, f(x \mid y) f(y). \tag{1.23}$$

In the case of a discrete distribution, we can compute marginal a marginal PMF.

$$f(x) = \sum_i f(x, y_i) = \sum_i f(x \mid y_i) f(y_i). \tag{1.24}$$

[4] This is very subtle. Jayne's book, *Probability: The Logic of Science*, Cambridge University Press, 2003, for more one these subtleties.

1.10 Statistical modeling

As scientists, we often have in mind a **generative process** by which the data we measure are produced. For example, we might expect the optical density of a solution of *E. coli* in LB media to grow exponentially over time, with some small measurement error. To model this, we specify a probability distribution to describe the measurements. We can then use the data and statistical inference to learn something about the parameters in the 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 "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. In the bacterial growth example, the mathematical model is that they grow exponentially.

Statistical model. A statistical model goes a step beyond the mathematical model and uses a probability distribution to describe any measurement error, or stochastic noise in the system being measured. This essentially means specifying $f(y \mid \theta)$ (and $g(\theta)$ in the Bayesian case). The statistical models we will use are **generative** in that the encompass the cartoon and mathematical models and any noise to use probability to describe how the data are generated.