Mathematical Foundations I: Probability and Inference

Stephanie Palmer sepalmer@uchicago.edu

BSD qBio^2 boot camp @ MBL

1 Probability Theory

Goal: This tutorial will cover basic probability theory, to give you a set of tools to model variability in biological data. You will also be able to understand and interpret common comparisons made between biological data and the behavior of standard stochastic processes, such as a Poisson process. We will derive equations for the mean and variance of a Poisson process, to build your intuition for these results instead of simply presenting these equations as facts to memorize. We will explore the limiting case where a Poisson distribution approaches a Gaussian distribution, another useful result of probability theory often used in biology.

1.1 Randomness in biology

This tutorial highlights an important and pervasive aspect of biological systems: stochasticity. (NB: 'stochasticity', 'variability', 'uncertainty', 'noise', and 'fluctuations' will all be used interchangeably here, though most of these terms have more technical and specific uses in other contexts.) Many of the variables that we observe in biological recordings fluctuate, sometimes because we cannot control all the states of the external and internal world of the organism, other times because thermal noise and other microscopic factors make the state of the biological system we interrogate inherently noisy. It is useful to model not only a median value for a fluctuating variable, but the full shape of its distribution of values.

For example, if we observe the firing of neurons in the brain to repeats of the same external stimulus, the precise times of spikes will vary between repeats. By fitting the statistics of this noise to models we deepen our understanding of the neural response.

In this tutorial, we will cover some basic concepts in probability theory, ending with some fundamental properties of entropy and information. In the Readings folder for this tutorial, you will find a review article by Tkačik and Bialek on information processing in biology. A special issue of the Journal of Statistical Physics (March 2016, v. 162(5)) is also dedicated to this topic.

To build your intuition about quantifying uncertainty, let's start with a toy problem I first encountered in David MacKay's lectures on information theory and inference.

1.2 Testing your intuition: the bent coin lottery

A biased coin is used to generate sequences of digits, 1 for heads, 0 for tails, in a lottery. The coin is tossed 25 times to determine the winning sequence. The probability of heads is 0.1. Tickets for the lottery cost $\$1$ and the prize is $\$10,000,000$.

Exercise 1.1

1. You are only allowed to purchase one ticket. Which ticket would you buy?

   Solution. The most likely ticket is the all-zeros ticket

   $$00000\,00000\,00000\,00000\,00000,$$

   with a probability of $P(000\ldots0) = (0.9)^{25} = 0.7178$.

   To compute the probability of any given sequence, you multiply together the probability of every coin flip, so

   $$P(000\ldots0) = P(0)P(0)P(0)\ldots P(0).$$

   How do we know this ticket has the highest probability? For any 0 that we replace by a 1, the probability correspondingly changes one of the 25 $P(0)$'s to $P(1)$, since $P(1)$ is 0.1 (which is less than $P(0) = 0.9$), then the resulting product must be less than the probability of all-zeros.

2. How many tickets to would you have to buy to cover every possible outcome?

   Solution. The total number of tickets is $2^{25} = 33,554,432$.

3. Is this lottery worth playing?

   Solution. At first glance, the prize money ($10,000,000) is less than the cost of buying all the tickets ($33,554,432), so we might think the lottery is not worth playing. However, not all tickets are equally likely, so we don't want to buy all the tickets. We only want to buy the most likely tickets. This doesn't guarantee that there are enough likely tickets to make lottery worthwhile, but we will show later that there are.

1.3 Binomial distribution

Each flip of a coin like this with probability, p, of heads is an example of a Bernoulli trial, the general term for an experiment with only two output states, success or failure. The number of heads in the sequence of independent coin flips generated by our lottery will follow a binomial distribution.

Exercise 1.2

1. Write down the probability of observing k heads in n coin flips, if the probability of heads is p.

   Solution. The probability of observing k heads in n coin flips is the same as the probability of observing **any** sequence of length n with k heads. We can find this probability by adding together the probabilities of every **particular** sequence of length n with k heads. That's a lot of terms to sum up. Luckily, the probability of a sequence of length n with k heads is $p^k(1-p)^{n-k}$, regardless of how the heads and tails are ordered. Why? As in Exercise 1.1, we find the probability of a particular sequence by multiplying the probabilities of each coin flip. Since the order of multiplication doesn't matter, the product is always $p^k(1-p)^{n-k}$.

   Now we just need to count how many sequences there are of length n with k heads, since the sum of their probabilities will equal the product of their count with $p^k(1-p)^{n-k}$. It turns out that counting sequences like this comes up a lot, so there is a notation for their count, $\binom{n}{k}$, pronounced "n choose

k," and known as a binomial coefficient. This binomial coefficient counts the number of ways you can choose k out of n objects. This is same as the number of different k-heads patterns because: we divide the total number of orderings of the sequence of coin flips $(n!)$ by the equivalence we define in the possible arrangements of $n-k$ tails, $(n-k)!$, and the k heads, $k!$. We're left with the just the unique placements of k heads in the sequence. It is defined as

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Then we have the probability of k heads in a sequence of n coin flips is

$$\binom{n}{k} p^k (1-p)^{n-k}$$

This is the binomial distribution. Rather than memorize this particular form, remember how to write it down as the product of intuitive terms. If we consider the limit of a very small p, we can relate the binomial distribution to the Poisson distribution.

1.4 Poisson distribution

The Poisson distribution describes the probability of finding k events in a fixed interval if we know the rate of occurrence of these events, λ, in that interval. In terms of the variables we have been working with for the bent coin lottery,

$$\lambda = p * n. \tag{1.1}$$

We are going to take the limit where p is very small and n is very large, but their product remains fixed.

Exercise 1.3

1. Derive an expression for the probability of observing k heads in n tosses in the limit of small p an large n.

   Solution. We just showed that the binomial distribution is given by

   $$P(k) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

   Equation 1.1 tells us $p = \lambda/n$, so

   $$P(k) = \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}.$$

   To make things easier, let's split that expression up and look at what happens to each part when we make n large. First, we'll take the first two terms:

   $$\frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k = \frac{n \cdot (n-1) \cdot (n-2) \ldots (n-k+1) \cdot (n-k) \cdot (n-k-1) \ldots 1}{k! \cdot (n-k) \cdot (n-k-1) \ldots 1} \left(\frac{\lambda}{n}\right)^k$$

This expansion makes it more apparent that the last $n-k$ terms of $n!$ are the same as $(n-k)!$, so we can simplify the above equation to

$$\begin{aligned}\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k &= \frac{n\cdot(n-1)\cdot(n-2)\ldots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k \\ &= \frac{n\cdot(n-1)\cdot(n-2)\ldots(n-k+1)}{k!\cdot\underbrace{n\cdot n\ldots n}_{\text{k times}}}\lambda^k \\ &= \frac{n}{n}\cdot\frac{n-1}{n}\cdots\frac{n-k+1}{n}\cdot\frac{\lambda^k}{k!}.\end{aligned}$$

As $n\to\infty$, all but the last fraction goes to one, so that the product is

$$\frac{n!}{k!(n-k)!}\left(\frac{\lambda}{n}\right)^k \xrightarrow[n\to\infty]{} \frac{\lambda^k}{k!}$$

Now let's look at the other half of the original equation,

$$\left(1-\frac{\lambda}{n}\right)^{n-k} = \left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k}.$$

The part with $-k$ in the exponent goes to one as n gets large. For the other part, we need to remember from calculus that $e^x = \lim_{n\to\infty}(1+\frac{x}{n})^n$, so

$$\left(1-\frac{\lambda}{n}\right)^{n-k} \xrightarrow[n\to\infty]{} e^{-\lambda}.$$

Combining both parts, we have

$$P(k) \xrightarrow[n\to\infty]{} \frac{\lambda^k}{k!}e^{-\lambda}$$

We have just written down the Poisson distribution. You will see this used as a model for biological variability again and again, either explicitly or implicitly. It is important to think about whether or not it is a good model for the system under study each time you come across it or are deciding to use it for your own research.

1.5 Interval between events

We can also write down the distribution of intervals between events in a Poisson process. This distribution has an exponential form.

Exercise 1.4

Derive an expression for the distribution of an interval, τ, between two events in a Poisson process with rate, λ, in this interval.

Solution.

Let's start by translating this problem back to the language of binomial distributions of sequences. Imagine we divide time into small, discrete bins of length Δt, such that the time interval of length τ is divided into n bins. We'll say the bins are so small that we can ignore the possibility of more than one event occurring in a single bin, so essentially we have $\Delta t \to 0$ or equivalently $n \to \infty$. Let r denote the rate per unit time of this process, so that the rate over the entire interval, τ is $\lambda = r\tau$. We now see the probability that we observe an interval τ between two events is the same as the probability that we observe n bins with no event followed by a bin with an event.

The probability that we observe n bins with no event is given by the Poisson distribution with $k = 0$ over τ, namely $P(0) = e^{-\lambda}$. The probability of one event in the last time bin is simply $r\Delta t$. Therefore we can write the probability of a τ-long silence followed by an event as

$$P(\text{no spike in } \tau\text{, then a spike in a subsequent } \Delta t) = r\Delta t e^{-r\tau}.$$

We usually express probabilities as density functions. Here, that's the probability per unit time, and yields

$$P(\tau) = \frac{P(\text{no spike in } \tau\text{, then a spike in a subsequent } \Delta t)}{\Delta t} = re^{-r\tau}.$$

1.6 Gaussian distribution

A Gaussian or 'normal' distribution (also called a bell-curve) of a variable x with mean μ and variance σ takes the form

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.2}$$

Exercise 1.5

1. As λ gets very large, show that the Poisson distribution approaches a Gaussian distribution with mean λ and variance λ.

   Solution. To do this, we need to use moment generating functions (MGFs), which are an alternative (often more useful) way of describing a probability distribution. An MGF is the expectation of e^{tX}, where X is the random variable. We will show that the standardized Poisson variable approaches the standardized Gaussian variable as $\lambda \to \infty$. So, letting X be our Poisson random variable with mean and variance λ,

$$\begin{aligned}
E\left[e^{t\frac{X-\lambda}{\sqrt{\lambda}}}\right] &= \exp\left(-t\sqrt{\lambda}\right) E\left[\exp\left(\frac{tX}{\sqrt{\lambda}}\right)\right] \\
&= \exp\left(-t\sqrt{\lambda}\right) \sum_{k=0}^{\infty} P(k) \exp\left(\frac{tk}{\sqrt{\lambda}}\right) \\
&= \exp\left(-t\sqrt{\lambda}\right) \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} \exp\left(\frac{tk}{\sqrt{\lambda}}\right) \\
&= \exp\left(-t\sqrt{\lambda} - \lambda\right) \sum_{k=0}^{\infty} \frac{\left(\lambda \exp\left(\frac{t}{\sqrt{\lambda}}\right)\right)^k}{k!}
\end{aligned}$$

$$
\begin{aligned}
&= \exp\left(-t\sqrt{\lambda} - \lambda\right) \exp\left(\lambda \exp\left(\frac{t}{\sqrt{\lambda}}\right)\right) \\
&= \exp\left(\lambda \exp\left(\frac{t}{\sqrt{\lambda}}\right) - t\sqrt{\lambda} - \lambda\right) \\
&= \exp\left(-t\sqrt{\lambda} - \lambda + \lambda\left(1 + t\lambda^{-1/2} + \frac{t^2\lambda^{-1}}{2!} + \dots\right)\right) \\
&= \exp\left(-t\sqrt{\lambda} - \lambda + \lambda + t\lambda^{1/2} + \frac{t^2}{2!} + \frac{t^3\lambda^{-1/2}}{3!}\dots\right) \\
&= \exp\left(\frac{t^2}{2!} + \frac{t^3\lambda^{-1/2}}{3!}\dots\right) \\
&\xrightarrow[\lambda\to\infty]{} \exp\left(\frac{t^2}{2!}\right).
\end{aligned}
$$

This limit is the MGF of a standard Gaussian distribution! Therefore the Poisson distribution approaches the Gaussian distribution with the same mean and variance λ, for large λ.

Exercise 1.6

1. Derive the mean of a Poisson distribution with rate λ:

   Solution. If K is our random variable with a Poisson distribution,

   $$
   \begin{aligned}
   E[K] &= \sum_{k=0}^{\infty} kP(k) \\
   &= \sum_{k=0}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\
   &= 0 + e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\
   &= 0 + \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}
   \end{aligned}
   $$

   We now substitute $j = k - 1$ to shift the sum,

   $$
   \begin{aligned}
   &= \lambda e^{-\lambda}\sum_{j=0}^{\infty}\frac{\lambda^j}{j!} \\
   &= \lambda e^{-\lambda}e^{\lambda} \\
   &= \lambda
   \end{aligned}
   $$

   just as we wanted.

2. Derive the variance of a Poisson distribution with rate λ:

Solution.

$$\begin{aligned}
\sigma^2 &= E\left[K^2\right] - E\left[K\right]^2 \\
&= \sum_{k=0}^{\infty} k^2 P(k) - \lambda^2 \\
&= \sum_{k=0}^{\infty} k^2 \frac{\lambda^k}{k!} e^{-\lambda} - \lambda^2 \\
&= 0 + \lambda e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} - \lambda^2 \\
&= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \left((k-1) \frac{\lambda^{k-1}}{(k-1)!} + \frac{\lambda^{k-1}}{(k-1)!} \right) - \lambda^2 \\
&= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) - \lambda^2 \\
&= \lambda e^{-\lambda} \left(0 + \sum_{k=2}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) - \lambda^2 \\
&= \lambda e^{-\lambda} \left(\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) - \lambda^2
\end{aligned}$$

and we again substitute $j = k - 1$ and $l = k - 2$ to shift the sums,

$$\begin{aligned}
&= \lambda e^{-\lambda} \left(\lambda \sum_{l=0}^{\infty} \frac{\lambda^l}{l!} + \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \right) - \lambda^2 \\
&= \lambda e^{-\lambda} \left(\lambda e^{\lambda} + e^{\lambda} \right) - \lambda^2 \\
&= \lambda^2 + \lambda - \lambda^2 \\
&= \lambda
\end{aligned}$$

just as we wanted.

These quantities are often summarized as the ratio of the variance and the mean, or Fano Factor (FF). The FF for a Poisson process is clearly equal to one.

Exercise 1.7

1. Does observing an FF $= 1$ in data mean that the underlying stochastic process is a Poisson process?

   Solution. It amounts to evidence that your underlying process might be Poisson, but one should be careful to look a bit deeper at one's data. Do you have enough data to determine that the FF is not different from 1? One might also want to verify that other characteristics of the biological process follow the Poisson predictions, not just this single characteristic. You might also measure the interval distribution between events, for example.

1.7 Winning the bent coin lottery

Now that we have all of these distributions at our fingertips, let's return to the bent coin lottery.

Exercise 1.8

1. Derive how many tickets you need to buy to guarantee yourself a 99% chance of winning.

   Solution. Let's use the Gaussian approximation for the probability of getting k heads out of n coin flips. Recall that the probability of one head is $p = 0.1$ and the number of coin flips is $n = 25$. Then we can calculate the mean number of heads is $np = 2.5$, and the variance is $np(1-p) = 2.25$, which gives a standard deviation of $\sigma = \sqrt{2.25} = 1.5$.

   The nature of a Gaussian distribution is that 95% of the distribution is within two standard deviations of the mean. That means that if we get all the lottery tickets less than two standard deviations above the mean, we will have at least a 95% chance of winning. In fact, we will have a higher chance of winning because the left tail of the distribution is heavier than a Gaussian, since the probability of $k = 0$ is so high. That means we want to buy up to 5.5, which doesn't really make sense since you can't have half a head. Instead, rounding up we guess that if you buy all tickets with 6 or fewer heads, we will have $\sim 99\%$ chance of winning.

   Now we can count how many tickets we need to buy according to our approximation:

   $$\binom{25}{0} + \binom{25}{1} + \binom{25}{2} + \ldots + \binom{25}{6} = 1 + 25 + 300 + \ldots 177,100$$

   That's way less than $10,000,000$, so at a dollar per ticket the lottery is definitely worth playing!

One of the new concepts we introduced in this derivation is a particularly fine quantification of uncertainty: entropy.

1.8 Entropy

Measuring uncertainty in a system allows you to attach a single quantity that describes the array of states you observe as its output. This concept of entropy derives its name from analogies to the Boltzmann entropy in statistical physics. Claude Shannon coined this term in his classic paper from 1948 on information theory. Shannon was working on communication systems at the time and was interested in developing a theory that would describe how to capture everything a transmitter produced and the maximal amount a receiver could reliably reproduce, given some noisy channel in between. Information theory has found many uses in biology. In this section, we aim to demystify Shannon's formula for entropy and give you some conceptual tools that will help you critically assess articles that use information theory in biology.

Shannon was interested in a measure, H, of a set of possible events with probabilities $p_1, p_2, \ldots p_i$, that would quantify uncertainty. He determined that it should obey a few simple rules to be considered a good measure. Namely, from Shannon's original paper:

- **continuity** H should be continuous in p_i.
- **monotonicity** If all p_i are equal, i.e. the distribution over n total states is uniform, such that $p_i = \frac{1}{n}$, H should increase monotonically with n.
- **branching** If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

1.8.1 Uniqueness

In this space, write out our derivation, using these three axioms, of the uniqueness of the entropy function, $H = -\sum_i p_i \log p_i$.

Solution. First, let's allow the entropy, H, of our distribution of n equally likely outcomes to be denoted $A(n)$. Our task is to show that the entropy function we defined above is the only reasonable function that describes the uncertainty in this distribution, and obeys the above three axioms. You have to take those at face value here. Please do read Shannon's lovely paper if you would like to dig into this a bit deeper.

Now, let's consider m choices, from each of s equally likely states, such that $A(s^m) = mA(s)$. Now, we want to find m such that

$$s^m \leq t^n < s^{m+1}.$$

Taking the logarithm of both sides we obtain

$$\begin{aligned} \log(s^m) &\leq \log(t^n) < \log(s^{m+1}) \\ m\log(s) &\leq n\log(t) < (m+1)\log(s) \end{aligned}$$

Next we divide through by $n\log(s)$

$$\frac{m}{n} \leq \frac{\log(t)}{\log(s)} < \frac{m}{n} + \frac{1}{n}$$

We take n arbitrarily large, so we can make the rightmost side arbitrarily small. Let's call that small number ϵ. Now we have that

$$\left| \frac{m}{n} - \frac{\log(t)}{\log(s)} \right| < \epsilon$$

Let's now invoke the fact that we want $A(n)$ to be monotonic, meaning that as n increases, so too should $A(n)$. Let's apply $A(\cdot)$ to the inequalities we just worked with

$$\begin{aligned} A(s^m) &\leq A(t^n) < A(s^{m+1}) \\ mA(s) &\leq nA(t) < (m+1)A(s) \\ \frac{m}{n} &\leq \frac{A(t)}{A(s)} < \frac{m}{n} + \frac{1}{n} \\ \left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| &< \epsilon \end{aligned}$$

So now we arrive at our first result

$$A(t) = K\log(t)$$

The variable K sets our choice of scale and it must be positive to satisfy monotonicity. Now let's get back to the probabilities of these n states. The probability of any particular state, n_i, is just

$$p_i = \frac{n_i}{\sum_i n_i}$$

meaning it's just the number of times we observed state i divided by the total number of times we observed any state. This means we're working with so-called commensurable probabilities. The p_i's can all be partioned into commensurate parts. We're now going to invoke the branching axiom. We are going to break the total $\sum n_i$ possibilities into n possibilities with probabilities p_i that are each divided into n_i equally probably choices. The branching axiom tells us how to relate these two

$$\begin{aligned} K \log\left(\sum n_i\right) &= H(p_1, p_2, \ldots, p_n) + K \sum p_i \log(n_i) \\ H &= K\left(\sum p_i \log\left(\sum n_i\right) - \sum p_i \log n_i\right) \\ H &= -K \sum p_i \log\left(\frac{n_i}{\sum n_i}\right) \\ H &= -K \sum p_i \log\left(p_i\right) \end{aligned}$$

That's the entropy we're all familiar with! To wrap things up, we need to return to our assumption about commensurate probabilities. Invoking continuity, we can say that the p_i may be approximated by rational numbers and we'll arrive at this same expression for the entropy.

1.9 Generating samples of a stochastic process

When modeling biological systems, it is often necessary to generate sequences from a Poisson or other stochastic process. We did this to generate our draws from the bent coin lottery. An introduction to simulating stochastic processes can be found in the Readings folder for this tutorial.

1.10 Markov processes

One feature of the stochastic processes we have been considering today is that they are independent. A flip of the coin doesn't depend on the flip before, or any of the other previous flips. In biological systems, what came before often influences a fluctuating quantity. For example, having spiked, a neuron is unable to spike for a millisecond or two. Modeling this type of variability falls requires using stochastic processes that have an explicit history dependence. Markov processes depend only on the previous time step, in generating the current state. Part of the introduction to point processes in the Readings folder covers Markov processes.

2 Inference

Goal: This section covers basic concepts in inference and will introduce the Bayesian and frequentist perspectives on the interpretation of data. We will resolve and discuss two logic puzzles to illustrate the differences in these approaches. We will define tools for incorporating prior knowledge into an estimate of the probability of an outcome of an experiment.

2.1 What is inference?

Inference is the act of drawing conclusions from data, usually by making some assumptions about the structure of the data. This involves selecting a model that describes how the data were generated and then drawing some conclusions (inferences) about this model, given the sampled data. Scientists in the machine learning community sometimes use the term 'inference' to describe the particular process of finding the time-evolving, unobserved or 'hidden' states in their models. More generally, scientists use the term inference to refer to the act of fitting statistical models to data. These can be fully parametric or non-parametric or mixed. Our goal in this tutorial is to familiarize ourselves with Bayesian and frequentist approaches to data, highlighting which approach is more useful given the problem we are trying to solve.

2.2 A frightening diagnosis

Let's start with a simple problem to build our intuition about how to make inferences from data.

Exercise 2.1

Hester is given a test for a terrible disease, a horrible, skin-decaying, eye-bleeding, zombification-type disease. The result of this test can be only positive (indicating presence of the disease) or negative. The test gives accurate positive results for 95% of those tested who have the disease, and accurate negative results for 95% of those tested who do not have the disease. About 0.5% of people in Hester's demographic have the disease. The test returns a positive result for Hester.

1. What is the probability that poor Hester has the disease?

   Solution. First, it is important to notice that the probability Hester has the disease is not 95%. Rather, someone who has the disease will test positively 95% of the time. For notational purposes, let's give each of these events a name. We'll call a positive test result $+$ and a negative result $-$, and we'll denote actually having the disease by D and not having the disease by N. Then we have $P(+ \mid D) = 95\%$, which in words means "The probability of a positive test result given that the patient has the disease is 95%."

   In order to evaluate the probability that Hester has the disease, we need to take into account our prior knowledge about the probability that Hester has the disease.

   To see why, imagine that we bring in $1,000$ people at random, all just like Hester, and test them all. Given the accuracy of the test, even if none of them has the disease, we would expect to get 50 positive results. On average, only 5 of them (from Hester's demographic) will actually have the disease. That's a big difference! We basically need to account for our expectation that a lot of people like Hester could get *false positive* results from this test.

   Now imagine you're the doctor who administers these tests. The first time you get a positive result, maybe you run screaming from the room. Zombies are, after all, quite terrifying. But now suppose

you test 1,000 random people every day for a month (probably prudent in the midst of a zombie apocalypse). By the end of the month, you might look a little more closely at someone who tests positive, but you don't think they're particularly likely to have the disease. Every day, about fifty people test positive, and every day, less than one in ten of those people actually turns out to be a zombie.

In fact, this gives you good intuition for the real probability: Once you see a patient with a positive test result, you expect there's slightly less than a 10% chance they are a zombie. (Slightly less because, given the expected 5 zombies, about 55 people will test positive, 5 of them being zombies.)

To account for the prior probability, we need to use Bayes' Rule

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+)}$$

How do we understand this equation? The left hand side is precisely the probability we are looking for: Given a positive result, what is the chance of having the disease? The numerator on the right is composed of the probabilities we have been given. Intuitively, first you ask what is the chance that you have the disease overall, multiplied by the chance that the test will give you that answer when taken. We next want to normalize by the probability that you get a positive result at all and that's the denominator.

For the denominator, we only know the probability of a positive diagnosis given a disease state, so we find $P(+)$ by summing up (or "integrating out") the different ways of arriving at that test result, meaning $P(+) = P(+ \mid D)P(D) + P(+ \mid N)P(N)$. Then we can write Bayes' rule as

$$\begin{aligned} P(D \mid +) &= \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid N)P(N)} \\ &= \frac{0.95 \cdot 0.005}{0.95 \cdot 0.005 + 0.05 \cdot 0.995} \\ &\approx 0.087. \end{aligned}$$

Therefore the probability that Hester has the disease is actually only about 8.7%.

2.3 Bayesian versus Frequentist

2.3.1 Priors and posteriors

In the previous derivation, we used Bayes' Rule to help us construct the correct estimate of Hester's probability of disease. The prior incorporated our knowledge of the risk for the disease in Hester's demographic, before we had knowledge of her test result. In this case, the prior greatly modified our estimate of whether or not Hester had the disease. The probability we calculated is called the posterior. It measures the probability of disease given Hester's test result. Moving from a prior to a posterior value by incorporating data is called a Bayes update. Moving from a prior distribution to a posterior distribution is a true Bayesian step.

2.3.2 Bayesian view of data

In a Bayesian framework, the parameters that one is trying to estimate are characterized by probability distributions that one has beliefs about characterized by prior distributions. A Bayesian uses data to answer

the question: Which parameters are most likely? The Bayesian view of the world is, in some sense, very abstract. There are no real fixed values of parameters in the world, only distributions. Bayesians must often perform averages over distributions of parameters to arrive at estimates. Because of this, it is sometimes joked that Bayesians spend their lives doing integrals.

2.3.3 Frequentist view of data

In the frequentist view of the world, there are true underlying physical parameters that have specific values, which we are trying to estimate from random samples of data. Frequentists set parameters with the data. Frequentists often 'average over the data' while Bayesians 'average over parameter distributions'.

We will now explore two classic problems that illustrate the differences between these approaches and will build your intuition about when to use each approach.

2.3.4 Two-envelope paradox

We begin by laying out the problem, which was first described over 50 years ago by Maurice Kraitchik.

Exercise 2.2

Two identical envelopes are prepared. One contains a quantity of money, x, while the other contains twice as much, $2x$. You pick an envelope, but before opening it or otherwise gaining any information about its contents, you are asked if you would like to switch envelopes or keep the one you have.

1. Let's describe the apparent paradox in the two-envelope problem: (copy from the board)

   Solution. If x denotes the quantity in our envelope, then we can calculate the expected value of switching as

   $$\frac{1}{2}\left(\frac{1}{2}x\right) + \frac{1}{2}\left(2x\right) = 1.25x > x.$$

2. Should we stay or switch?

   Solution. Intuitively, it seems like switching can't possibly help, but the expected value described above seems pretty convincing. Many of you were pretty smart about this and decided not to switch. Let's explore exactly why the above argument is wrong, though.

3. A simpler scenario, the so-called necktie paradox, may help clarify your thinking about the flawed logic in this problem. Two men are at a Father's Day party and both have been given ties by their children. They argue over which tie was more expensive. They propose a bet. They will consult their children and find out whose is pricier. The one with the more expensive tie has to give it to the other man. Only the man who thinks his kids might have gipped him should take the bet, right? Let's see: Each dad's reasoning goes like follows: I have a 50/50 change of winning. If I bet and lose, I lose the value of my tie. If I bet and win, I gain more than the value of my tie. It seems like betting (against your children's good will) is a good thing! Also, paradoxically, both men seem to have an advantage in betting. Can you describe the flaw in this logic?

   Solution. Let's make a table of possible outcomes to clarify the situation.

Dad 1	Dad 2	Dad 1 Gain
\$20	\$40	+\$40
\$40	\$20	−\$40

Even though each dad only either "loses their own tie or gains one of greater value," the expected gain in any particular situation (say, with the value of the cheaper tie fixed to \$20 as in the table) is zero dollars.

4. In the two-envelope paradox, describe a similar string of flawed logic:

   Solution. Let's make another table!

My envelope	The OTHER envelope	Switching gain
\$20	\$40	+\$20
\$40	\$20	−\$20

   In this way, we see that fixing hypothetical values of the envelopes resolves the paradox. Many math, statistics, and even philosophy papers have been written when you try to pin down precisely what it means to make assumptions about the cash that might be in the envelopes. There are even scenarios when switching always makes *real* sense. We covered that a little in class but we can actually sew this up by avoiding that question and without too much more work. Let's press on!

5. What would someone versed in probability theory do?

   Solution. This is one of the simple resolutions to the two envelopes paradox: Let's label the two envelopes A and B. We've been handed A to start. What we want to do is compute the amount of money we expect to find in each of the envelopes. Note that this is not the actual amount in the envelope, it's just what we expect to be there if we were allowed to play the game many many times with the same dollar amounts in A and B. Let's start with envelope B, the one we get by switching. We now want to compute how much money we expect to find in B given the two scenarios that might occur, namely that there's more money in A, i.e. $A > B$, or there's less money in A, $A < B$. These two scenarios are equally likely since we have no information about which envelope contains more money a priori. Let's write that out:

$$E(B) = E(B|A < B)P(A < B) + E(B|A > B)P(A > B)$$

   Now let's substitute in the fact that we now the two envelopes contain some amount x, and another amount $2x$. That means that the larger amount is twice the smaller, and the smaller amount is half the larger (no magic here, just plain arithmetic). So, if $A < B$, we expect that envelope B contains twice the value in A. If $A > B$, we expect B to contain half of what's in A. These two scenarios are equally likely. Let's write that in:

$$\begin{aligned} E(B) &= E(2A|A < B)\frac{1}{2} + E\left(\frac{1}{2}A \,\middle|\, A > B\right)\frac{1}{2} \\ &= 2E(A|A < B)\frac{1}{2} + \frac{1}{2}E(A|A > B)\frac{1}{2}. \end{aligned}$$

   Now let's assume the amount in envelope A, the one we started with, is some value x, then we have

$$E(B) = x + \frac{1}{4}2x = \frac{3}{2}x.$$

We could go through the whole procedure for envelope A and arrive at exactly the same answer, meaning that the amount we expect to find in each envelope is exactly equal and there is not point to switching. The amount we expect in each envelope is weirdly more than the amount we hypothesize to exist in the envelope we're given first on a particular trial, but our calculation is sound.

The logic puzzle about how our expectations concerning what could possibly be in the envelopes shapes our strategy, and what we're actually trying to estimate in this puzzle, takes you down a rabbit hole we won't explore here. If you'd like to know more, have a look at some of the articles written recently about the two envelopes problem

The two-envelope problem shows us how a knee-jerk approach leads us astray, but a probability theory approach, explicitly calculating conditional probabilities, resolves the apparent paradox.

2.3.5 Lindley's paradox

We now turn to another apparent paradox, that will show us how Bayesian and frequentist approaches can arrive at opposing conclusions. Lindley's paradox is about model comparison between H_0, our null, and H_1, given some data, x. It exists when a frequentist rejects the null hypothesis, H_0, but a Bayesian favors H_0 over H_1.

Exercise 2.3

A classic example of Lindley's paradox applied to estimating the boy/girl birth ratio in a population. In the city Bayfreak, $49,581$ boys and $48,870$ girls were born in the last three years. Assume that the number of male births is a binomial variable with parameter, θ. We wish to test whether $\theta = 0.5$ or some other value.

1. What would a frequentist do?

   Solution. One frequentist approach would be to approximate the number of boy births by a Gaussian distribution, and ask if it is likely that the observed number of boys came from the Gaussian distribution you would expect with $\theta = 0.5$.

   The hypothesized mean is the total number of births multiplied by the hypothesized ratio, so $\mu = N\theta = (49,581 + 48,870) * 0.5 = 49,225.5$. Since this Gaussian is an approximation to a binomial, we can calculate the hypothesized variance as $\sigma^2 = N\theta(1-\theta) = (49,581 + 48,870) * 0.5 * 0.5 = 24,612.75$.

   With these values, we can calculate how likely it is that the number of boys born, x, would be greater than what we observed, x_{obs}, assuming x is distributed as a Gaussian with mean μ and variance σ^2.

$$\begin{aligned} P(x \geq x_{\text{obs}}; \mu, \sigma^2) &= \int_{49,581}^{98,451} \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(u-\mu)^2}{2\sigma^2}} \\ &= 0.0117 \end{aligned}$$

   which means $p < 0.05$, so at the 0.05 significance level, a frequentist would reject the hypothesis that $\theta = 0.5$.

2. What would a Bayesian do?

Solution. A Bayesian would, quite naturally, use Bayes Rule. Let's denote

$$H_0 : \theta = 0.5 \qquad H_1 : \theta \neq 0.5$$

so that we can write

$$P(H_0 \mid x_{\text{obs}}) = \frac{P(x_{\text{obs}} \mid H_0)P(H_0)}{P(x_{\text{obs}} \mid H_0)P(H_0) + P(x_{\text{obs}} \mid H_1)P(H_1)}.$$

We can assume both hypotheses are equally likely. Then it remains to find the probability of the observed number of boys, x_{obs}, under each hypothesis.

The probability under the null hypothesis is simply the probability of a binomial with $k = 49,581$ and $n = 98,451$,

$$P(x_{\text{obs}} \mid H_0) = \binom{n}{k}\theta^k(1-\theta)^{n-k} = 1.95 \times 10^{-4}.$$

The probability under the alternate hypothesis is a little bit more complicated. It is actually the average over the probabilities for all possible θ,

$$P(x_{\text{obs}} \mid H_1) = \int_0^1 \binom{n}{k}\theta^k(1-\theta)^{n-k}p(\theta)\,d\theta.$$

This forces us to choose a prior on θ, denoted by $p(\theta)$, and, lacking any better information, we might choose a uniform prior, so $p(\theta) \equiv 1$ for $0 \leq \theta \leq 1$. The result of this integral is 1.02×10^{-5}. Plugging these values into Bayes rule, we find $P(H_0 \mid x_{\text{obs}}) \approx 0.95$, so we fail to reject H_0.

Lindley's paradox teaches us that if you have information, use it, but if you do not, don't make an overly diffuse prior. If you do, you will need a lot of data to overcome it.

2.4 Regression

The term regression describes a model class for performing inference. The 'linear' part of linear regression describes the structure of the relationship between the data and the parameters in the model. If your function is linear in the fitted parameters, it is linear regression, even if the model being fit is, say, a polynomial where the parameters are the coefficients in front of each term. Let us label our sampled data, x, and our the parameters we would like to fit, λ. We will now cover some of the most popular methods for estimating λ.

2.5 Maximum Likelihood (ML) inference

In the maximum likelihood framework, we seek to find the λ that maximize

$$P(x|\lambda), \tag{2.1}$$

the likelihood that the data, x, were drawn from a distribution defined by the parameters, λ. There is an important distinction between a probability distribution and a likelihood, though they may written in the same form. Take this conditional probability, $P(x|\lambda)$, as an example. If the λ are fixed numbers, then this is just a probability distribution over x. If the λ are the variables and the data are fixed, then $P(x|\lambda)$ is a likelihood. This likelihood is not a probability distribution over the parameters, it is a probability of the data given the parameters.

A few other notes: It is often useful to work with the logarithm of a function, and maximizing a function or its logarithm are equivalent. Within the ML framework, one can add a Bayesian prior that the parameters are most likely zero, amounting to a type of L1 penalty.

2.6 Maximum A Posteriori (MAP) inference

In contrast, maximum a posteriori inference seeks to maximize the conditional probability of the parameters given the data:

$$P(\lambda|x) \tag{2.2}$$

We use Bayes' Rule to express this quantity in terms of things we can measure. Expressing $P(\lambda|x)$ in this way, we have

$$P(\lambda|x) = \frac{P(x|\lambda)P(\lambda)}{P(x)}. \tag{2.3}$$

The $P(x)$ are the same in both the ML and MAP frameworks. What MAP inference has added to the mix is the prior on the parameters, $P(\lambda)$. This addition might lead you to infer that MAP inference is Bayesian, however MAP inference is a point estimator (i.e. the output of the procedure is a set of numbers that are the fit parameters) while the output of a Bayesian estimator would be characteristics of a distribution of parameters.

Exercise 2.4

1. Show how ML and MAP are related when the prior is uniform:

   Solution. Say the prior is uniform so that $P(\lambda) \equiv c$. Then we can see from Bayes rule

   $$P(\lambda \mid x) = \frac{P(x \mid \lambda)P(\lambda)}{P(x)} = \frac{P(x \mid \lambda)c}{P(x)}$$

   and since $P(x)$ doesn't depend on λ, we then have $P(\lambda \mid x) \propto P(x \mid \lambda)$.

2.7 Bayes estimator

A Bayesian estimator seeks to minimize the 'risk' generated by a loss function, $L(\lambda, \lambda^{\text{est}})$, in forming an estimate of the parameters, λ^{est}, given that the true parameters are λ. This risk is

$$\int d\lambda L(\lambda, \lambda^{\text{est}})P(\lambda|x), \tag{2.4}$$

where x is, again, the data. The mean-squared error is a commonly used loss function.

2.8 Further reading

In this tutorial, we have focused on big concepts rather than particular methods for model analysis, generation, and data sampling. Some tutorials on the techniques that you should familiarize yourself with are included in the Readings section of this tutorial and cover: Hidden Markov models, ROC analysis, quantile-quantile or QQ plots, and the Markov chain Monte Carlo (MCMC) sampling technique.