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?

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

3. Is this lottery worth playing?

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

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

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,

(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

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

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.

Exercise 1.4

Show that in the limit of large λ, the Poisson distribution is well-approximated by a Gaussian distribution..

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 1.5

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

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

(1.2)

Exercise 1.6

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

You have now derived the form of a very useful distribution and have learned how to evaluate a Gaussian integral along the way.

Exercise 1.7

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

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

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 1.8

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

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 1.9

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

1.8 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.

1.9 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. 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?

To combine all of the information given carefully, we will use a simple identity dubbed Bayes' Rule to write out the relevant conditional probability distributions. Follow the derivation supplied at the board and write it down here:

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. Should we stay or switch?

A simpler scenario, the so-called necktie paradox, may help clarify your thinking about 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. 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. Paradoxically, both men seem to have an advantage in betting.

2. Can you describe the flaw in this logic?

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

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

5. What would a Bayesian do?

The two-envelope problem shows us how a knee-jerk approach leads us astray, but a Bayesian 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?

2. What would a Bayesian do?

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

(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:

(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

(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:

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

(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.