Probability

Author: Cole Brookson **Date:** 29 August 2022

An Introduction (And Disclaimer. . .)

There are many excellent texts on probability and probability theory. There are numerous courses in several departments at the University of Toronto, and indeed pretty much any university, that can offer you the student a truly grounded view of what probability is and how to think about it. Indeed, probability itself can be interpreted in multiple ways, and multiple theories about foundational probability and how to interpret it exist. Most of the concepts in this section are unabashedly stolen from the excellent (and free!) text on probability by Blitzstein and Hwang which you can find here. This, however, is a short website page, within which we attempt to discuss some foundational components of probability that we think are important for students in EEB to know.

There's many reasons why we should study probability, but perhaps most simply, the biological world around us is full of randomness, and even in the processes we think we understand the best, uncertainty is abound. Without an understanding of probability, we don't have a framework with which to confront of understanding or lack thereof, and actually enable a robust discussion about how much randomness we actually understand, and how certain we are about things we think.

In this section we'll offer some definitions that will be useful to you in more formally discussing probability.

Key Terms:

1. Random Trials/Experiment
2. Events
3. Outcomes
4. Probability (Naive)
5. Sample Space
6. Population

Sample Spaces, Events, and Naive Probability

Probability is based on the mathematical concept of sets, from which we'll steal some jargon. Let's imagine an example to ground this.

Say there are 1000 spiders in a bag. One of those spiders happens to have a red dot on their thorax. You pull a single spider out of that bag, without looking. You could have either pulled A) a spider with no red dot, or B) the spider with the red dot. What is the probability that you have selected the spider with the red dot?

Above we've actually used all the key terms above in description. First, we've actually defined the sample space (the set of all possible outcomes of an experiment/trial) as being finite - the spiders in the bag that we can think of as a mathematical set. The population (related but different to a sample space) is the set of all units that some random process can select. Sample spaces can be finite, countably infinite, or uncountably infinite. Most biological applications, and all the applications we'll consider here, have finite sample spaces. That is to say, we can construct a set containing every single outcome of an experiment or random trial. In this example, each spider represents an outcome, and an event is some set of spiders. Side note: sets can have one or even no items in them. The events are what we would mathematically consider subsets of the sample space. So here we have two events of interest: The subset A) 999 spiders with no red dots, and B) 1 spider with a red dot.

Here, we have a naive version of a probability, wherein we can actually just define the probability of choosing the spider with the red dot as a simple proportion. To state this clearly, the probability of performing the experiment, and the B) being the outcome can be defined as:

$$P(\text{B}) = \frac{\text{number of outcomes favourable to B}}{\text{Total number of outcomes in the sample space } S}$$

which specifically means:

$$P(\mathrm{B}) = \frac{\text{number of spiders with red dots}}{\text{total number of spiders in } S}$$

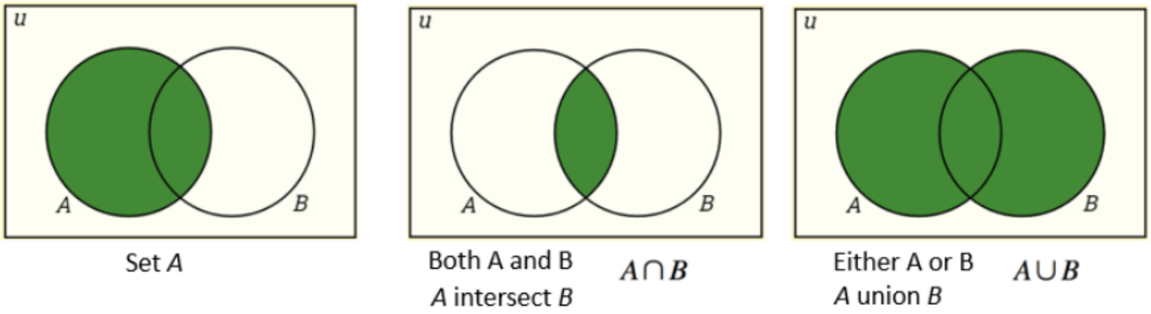
Or

$$P(\mathrm{B}) = \frac{1}{1000}$$
$$P(\mathrm{B}) = 0.001$$

Now, to add more conditions, we're assuming here that when we put our hand into the bag to select a spider, there are absolutely no defining features that make one spider more or less likely to be chosen. However, there is a case where some spiders are bigger, or perhaps less afraid of the hand coming into the bag, and these traits may make some spiders more or less likely to be chosen.

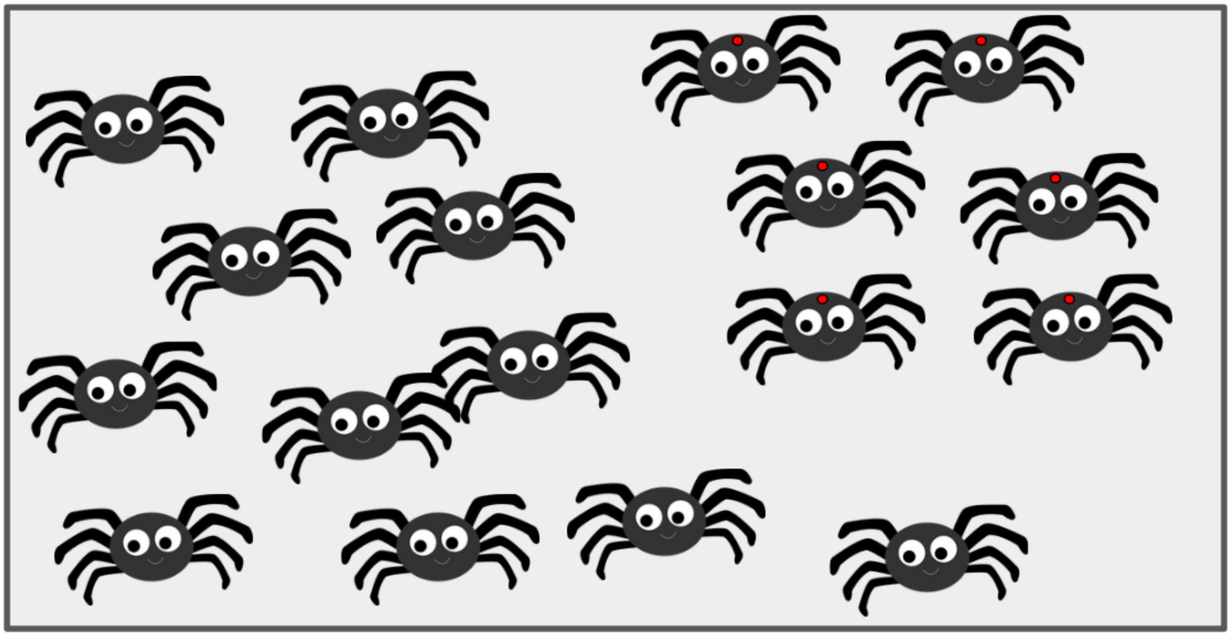
This is easy enough to replicate in R, as it's simply a division calculation:

SIDE BOX: Quickly, let's define two (only two!) operations from set theory to help us out. We'll define the sample space S as a set, and our two events as subsets A and B of S. We can think of the part of S encapsulated by both subset A and subset B as being the *intersection* of A and B, and we can separately think about the part of the sample space S that is encapsulated by either A or B as the *union* of A and B. We write the intersection of A and B as $A \cap B$ and the union of A and B as $A \cup B$. Refer to this image for a visual idea of this:



Two Events

For now, let's just consider that there are only two events we're interested in, A and B. The outcome of A is selecting a spider with no red dot, and the outcome of B is selecting a spider with a red dot. To make things more interesting, now let's assume there are 234 spiders with red dots (and therefore 766 without red dots). We can visualize this by imagining peeking inside our bag of spiders and seeing something like the following:



If we imagine all the spiders present as the set S defining the sample space, it's possible to coarsely replicate the Venn diagram above to make an estimate of what the probabilities of both the union and the intersection might be:

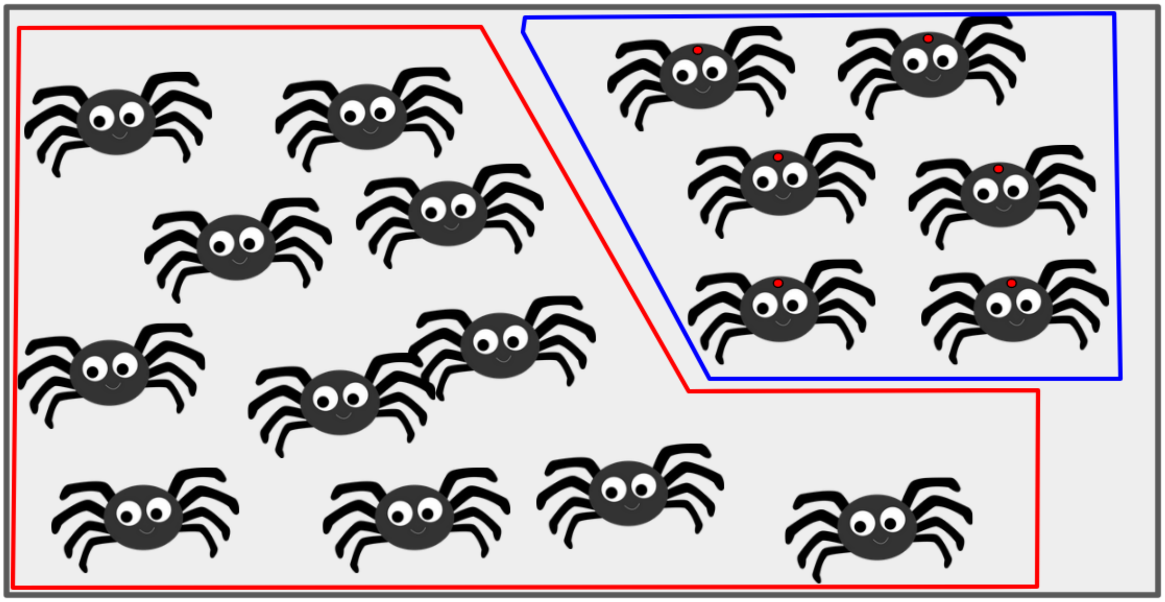


Figure 1: spiders

We can now see clearly that all of the spiders (i.e. all of S) is contained within the two event subsets A and B. Further, there are actually no spiders that fall within both subsets.

What we have defined here are two *mutually exclusive* events. With only one experiment, the outcome cannot satisfy both events. Formally, we say that the probability of both A and B is zero:

$$P(\mathrm{A} \cap \mathrm{B}) = 0$$

But also, since there are are no spiders that fall outside of A or B, the probability of either one event OR the other occurring with one experiment is actually 1:

$$P(\mathrm{A} \cup \mathrm{B}) = 0$$

Naive Vs. Standard Probability

The reason to differentiate between these two definitions is one of caution. We defined above the naive probability, which has two incredibly important assumptions:

1. The sample space is finite
2. Each outcome is equally likely

This clearly serves us well for our current example, and in fact from here on we will refer to the naive probability simply as probability, but it's useful to recognize that as soon as we care to make more interesting conditions surrounding our sample space, we must adjust our definition of probability (we won't cover that here).

Probability Distributions

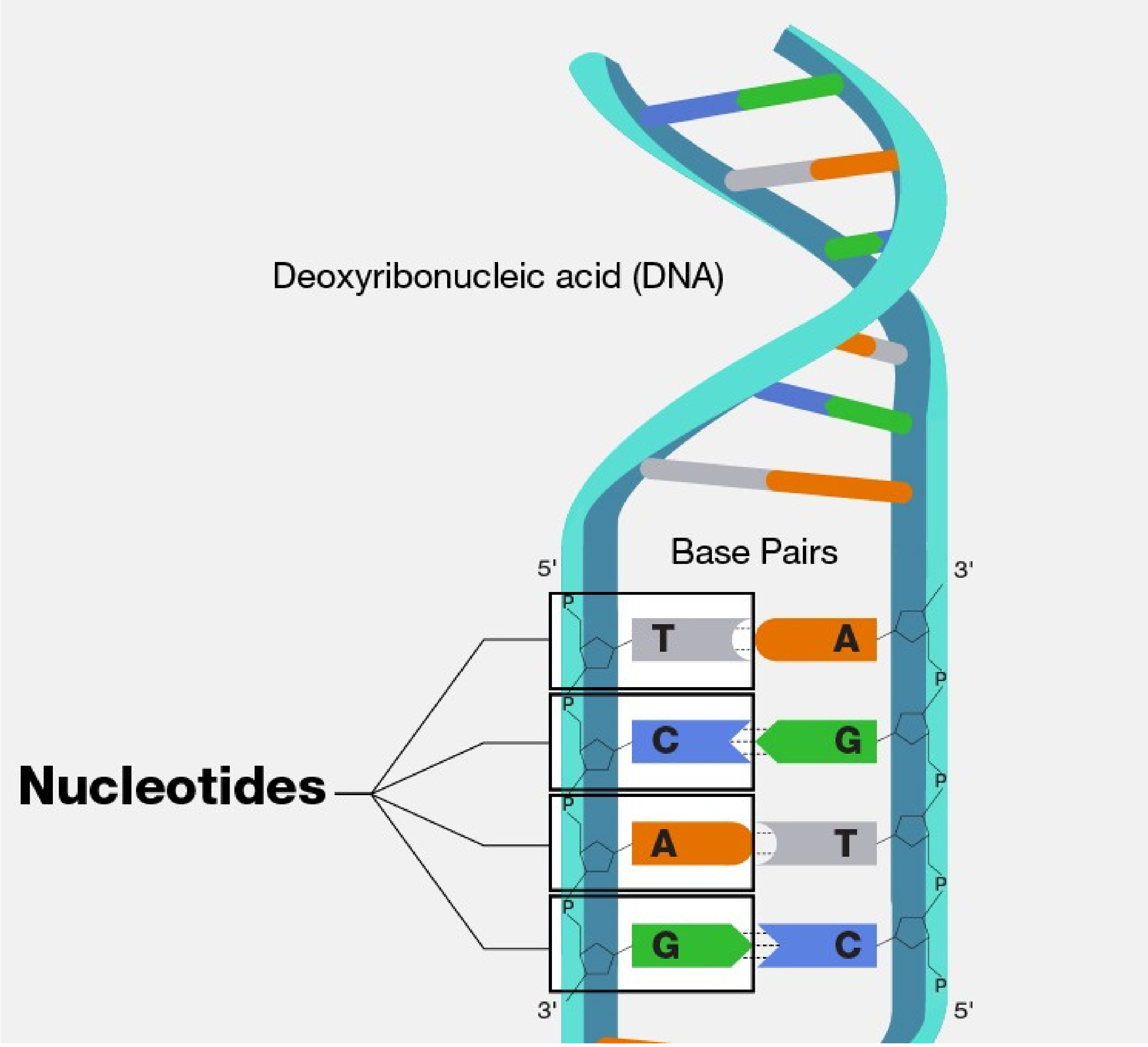
Our above example was primarily focused on an example where two possible events covered the entire sample space. A commonly used example in the same vein would be flipping a coin - there are only two possibly outcomes: heads or tails. However, this is usually not the case. While it's possible we may only be interested in a handful of the events possible, it's rare that those encompass the entire sample space.

To think about the probability of a variety of events occurring, we can differentiate between discrete probabilities and continuous probabilities.

Discrete Probability Distributions

As the name implies, discrete probability distributions are distributions that describe the probabilities of each possible event in the sample space.

For example, let's consider the basic building blocks of DNA, the nucleotide bases adenine (A), cytosine (C), guanine (G) and thymine (T). Imagine we were able to "zoom in" on some random part of our own DNA, we would see something like this:



If we zoom in on a single base, we can only "see" one of A, C, G, or T. So, our sample space S is made up of the four mutually exclusive events, which is seeing either A, C, G, or T. If we assume that bases are distributed with even frequency, we can imagine that the probability of each would be the same:

$$P(\mathrm{A}) = 0.25$$
$$P(\mathrm{C}) = 0.25$$
$$P(\mathrm{G}) = 0.25$$
$$P(\mathrm{T}) = 0.25$$

which would result in a probability distribution that may look like this:

This is intuitive. Any way that the distribution is partitioned, the sum of probabilities of each event must sum to 1.0 without exception.

The most common discrete probability distributions are the Poisson and Bernoulli distributions.

Mass Functions Discrete distributions typically are described by their probability mass functions, which give the probabilities of a discrete random variable being exactly equal to some values. These probabilites must always sum to 1.0 and have a general form of

$$p_X(x) = P(X = x)$$

. It's helpful to know how to plot probability distributions, as it's visually easiest to see when different parameters result in different probability distributions when we can compare parameter values.

The probability mass function for the Poisson distribution takes the following form:

$$p(x; \lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

And as there's only one parameter (λ), this parameter simply indicates the "shape" of the function, or how many average events happen in a given time interval. Conveniently in R, there are mass functions for most common distributions. Let's take a look at ?rpois and we see that we can plot the probability densities with the dpois function. Let's choose three different parameter values and see how that changes our probability density. In addition, it's actually much easier to make plots of discrete distributions in base R so we'll do that:

Now note here that we've plotted this as these vertical lines, instead of some continuous curve. Why? Well, recall this is a discrete probability distribution. There are no probabilities in the continuous space between integer values. So to plot this as a curve would actually be incorrect. That's why it's better to be explicit in this fashion and plot them as such here. Other λ values:

We can see how the mass function changes given our changing values for the λ.

Continuous Probability Distribution

Again as the name implies, between two variables which are both continuous, an infinite number of other values are possible. To think about these distributions, we often think of probability densities.

It's useful to be able to plot probability densities for a handful of common continuous distributions. We'll plot probability densities for the Gaussian, Exponential, and Gamma distributions.

Density Functions Conveniently in R, there are density functions for most common distributions. In fact, if we take a look at ?rnorm, we see a helpful set of functions to do with the normal distribution. We have functions for the density dnorm, the distribution function pnorm, the quantile function qnorm and a random deviate sample rnorm. This exists for other distributions too (HINT: try ?rgamma)

Let's plot the density of a few normal distributions showing how a difference in parameters gives different probability density curves. As a good reminder, to recall what forms the probability density functions for some given distribution look like, the NIST/SEMATCH e-handbook of Statistical Methods has a great set of descriptions on probability distributions. Let's refresh our minds about the Normal Distribution. The probability density function for the normal distribution is

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}$$

This looks a bit complicated, but in reality, there's only two parameters! The one's we're used to for the normal distribtuion, μ the mean, and σ the standard deviation. Similarly to how we did with the Poisson distribution above, let's see how changing these variables will change our probability density curves.

Let's first see how changing the σ will affect the curves. Again, base plotting actually has an advantage here:

Now let's add some other standard deviations:

We can also hold the standard deviation constant and plot different mean values. Let's try that now:

This likely looks much as we may have anticipated. We are simply shifting the plot left or right depending on the mean.