Supplementary material to:

Primer 3: Probability Theory

From:

A Biologist's Guide to Mathematical Modeling in Ecology and Evolution

S. P. Otto and T. Day (2005)

Princeton University Press

Supplementary Material P3.1: Scientific inference and Bayesian analysis

Scientists perform experiments and gather data to gain evidence for or against various hypotheses about how the world works. This sounds straightforward, but exactly how we use this data to make quantitative statements about various hypotheses requires a bit more care. In fact, there are two related, but philosophically distinct, approaches to scientific inference. To some extent these two approaches parallel the frequency interpretation and the subjective interpretation of probability. "Classical" statistical analysis is most closely allied with the frequency interpretation, whereas "Bayesian" analysis is allied with the subjective interpretation.

To quantify how the data from a particular experiment (or observation) informs us about the world under the classical approach, we first propose some hypothesis (usually referred to as the "null" hypothesis). We then imagine repeating the process of data collection (e.g., repeating an experiment) over and over many times assuming that the null hypothesis is true. Then we ask how likely it is that we would have obtained the data from our actual experiment, given that the null hypothesis is correct. In other words, what proportion of our imaginary experiments give rise to data like those that we collected assuming the null hypothesis is true? If this proportion is very small, then we reject the null hypothesis.

In many situations, however, it is difficult to imagine repeating the process of data collection. For example, if the data arise from observing all of the events that have happened (e.g., counting the number of bird species that have gone extinct on a particular island using archeological remains) then exactly what is meant by "repeating the process

of data collection"? Furthermore, we might wish to combine previous information about a hypothesis with such observations in order to be more confident about our conclusions. This is where Bayesian inference comes into play. These ideas are best illustrated by example.

Imagine that you are an apple farmer and manage an orchard of 100 trees. You observe a mutation on one branch of an apple tree that leads to particularly delicious apples, so you decide to replace one of your 100 trees with this mutant variety as a pilot project. You then have an orchard with 99 trees of the wild type strain and one tree with this new mutation. At this point, a disease spreads throughout your orchard and kills all 99 of the wild type trees, but the one mutant tree survives. The idea strikes you that perhaps your new apple line is resistant to the disease. How can we quantify the likelihood that this is the case, based on the observed data?

The classical statistical approach goes as follows. Suppose that the new mutation is not resistant to the disease – this is the null hypothesis. Then we ask, what is the probability of the observed outcome occurring if the null hypothesis were true? If this probability is very small (smaller than some pre-specified cutoff – typically 5% by convention), then we reject the null hypothesis and conclude that the new mutation must be resistant. In the present example, if the mutant is not actually resistant, then all trees are equally susceptible to the disease. Thus, given that only one tree survives, if you repeatedly sampled the population for a sole surviving tree at random, the probability that by chance the sole surviving tree is the mutant is just 1/100. In other words, if you could re-run the "experiment" over and over, then only 1 out of every 100 trials, or 1%, would result in the observed pattern if the null hypothesis were true. Because 1% is below the conventional cutoff of 5%, you conclude that the evidence supports the idea that the mutant tree is more resistant to the disease.

In trying to publish these results in a scientific journal, however, you might run into the following objection from a reviewer. The reviewer might argue that you have not taken into account the fact that we know mutations are rare, on the order of 10^{-6} per gene. This makes it extremely unlikely that the new line of apples happened to bear a new

mutation that provides resistance to the disease. How could you incorporate this prior information about the frequency of mutations into your calculations? This is where Bayesian analysis enters the picture.

Bayesian analysis asks the following question: What is the probability that a hypothesis is true, given the data and any prior information that we have? The prior information is incorporated using Bayes' formula (Rule P3.7):

$$P(\text{hypothesis} \mid \text{data}) = \frac{P(\text{data} \mid \text{hypothesis})\, P(\text{hypothesis})}{P(\text{data})}, \qquad \text{(SP3.1.1)}$$

where P(hypothesis) describes the "prior" probability that the hypothesis is true. P(data) describes the probability of observing the data and must be calculated over the entire spectrum of possible hypotheses, using the Law of Total Probability (Rule P3.8):

$$P(\text{data}) = \sum_i P(\text{data} \mid \text{hypothesis} = i)\, P(\text{hypothesis} = i). \qquad \text{(SP3.1.2)}$$

In the present example we have two possible hypotheses: the null hypothesis that the new variety of apples is "not resistant", and the alternative hypothesis that it is "resistant". To incorporate our prior information about mutation rates, we could set the probability of resistance to a typical mutation rate, e.g., $P(\text{resistant}) = 10^{-6}$, so that $P(\text{not resistant}) = 1 - 10^{-6}$. The data that we have is that the only tree to survive the disease was the new line of apples ("new line survived"). The probability of observing this data can by calculated using the Law of Total Probability (Rule P3.8) as:

$$\begin{aligned} P(\text{new line survived}) = {} & P(\text{new line survived} \mid \text{not resistant})\, P(\text{not resistant}) \\ & + P(\text{new line survived} \mid \text{resistant})\, P(\text{resistant}) \end{aligned}. \qquad \text{(SP3.1.3)}$$

Under the hypothesis that the new variety is not resistant and given that one tree survived, the probability that the only surviving tree would be the new variety is $P(\text{new line survived} \mid \text{not resistant}) = 1/100$. Under the hypothesis that the new variety is resistant, the probability that it survived would be near one

$P(\text{new line survived} \mid \text{resistant}) \approx 1$. Putting all of these terms into (SP3.1.3) gives us the probability of the data:

$$P(\text{new line survived}) \approx (0.01)\left(1 - 10^{-6}\right) + (1)\left(10^{-6}\right) \approx 0.01000099. \quad \text{(SP3.1.4)}$$

At this point, we can use the above pieces of information in Bayes' formula (SP3.1.1) to calculate the probability that the new line is resistant given that it was the sole surviving tree:

$$P(\text{resistant} \mid \text{new line survived}) = \frac{P(\text{new line survived} \mid \text{resistant})\, P(\text{resistant})}{P(\text{new line survived})} = \frac{(1)\left(10^{-6}\right)}{0.01000099}, \quad \text{(SP3.1.5)}$$

which is approximately 0.0001. Conversely, the probability that the new line is *not* resistant given the data is 0.9999. Consequently, we would conclude that the new line is probably not resistant. Thus, accounting for the rarity of mutations paints a very different picture of whether the new line is resistant.

While the calculations in the apple example are straightforward, the philosophical issues are not, and it is tricky to know which approach is best. If resistance to the disease is economically critical, then it might well be worth following up and testing the new variety, because we know that this tree is more likely to be resistant than a randomly chosen tree. On the other hand, you would probably not be justified in claiming that the new variety is definitely resistant from this data alone, given how unlikely it is to have borne a new mutation.

The calculations in this example are relatively straightforward because there are only two possible hypotheses (resistant and not resistant). In many cases, there might be several, or even an infinite number of hypotheses. Suppose our prior information can be encapsulated by a *prior probability distribution*, $P(X = x)$, describing the probability that a random variable X (representing the hypothesized value of some entity, such as level of resistance) takes on the value x. We can then use (SP3.1.1) to describe the *posterior*

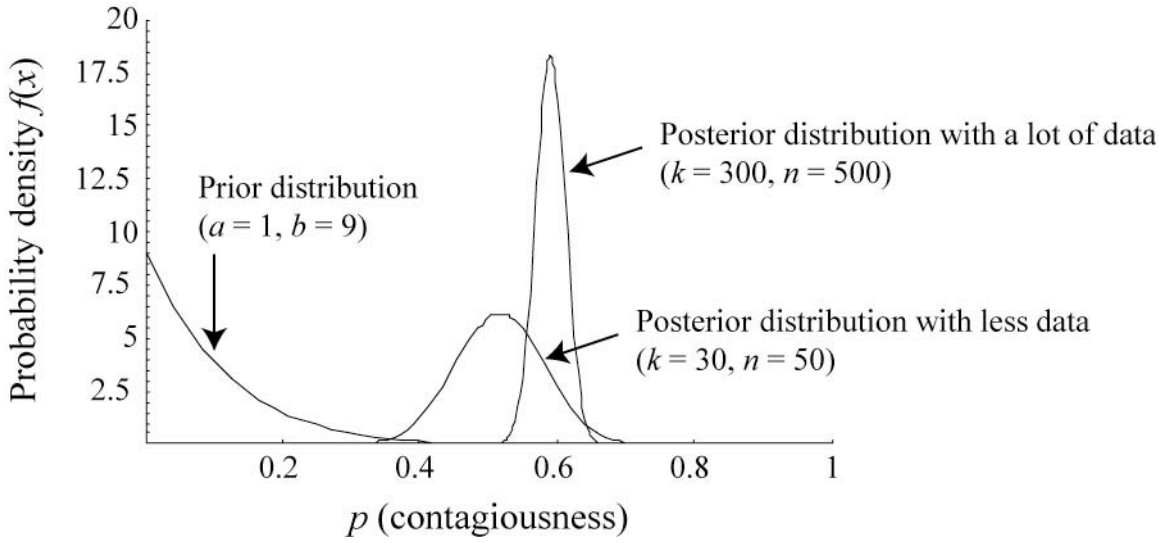
probability distribution , $P(X = x \mid \text{data})$, that the random variable takes on value x given that certain data have been observed:

$$P(X = x \mid \text{data}) = \frac{P(\text{data} \mid X = x)\, P(X = x)}{P(\text{data})}, \qquad \text{(SP3.1.6)}$$

where $P(\text{data})$ is calculated from Rule (P3.8) but by integrating over all values, x, that X can take rather than discrete summation.

As an example, the beta distribution (Definition P3.12) has been used as a prior distribution to describe an individual's risk of contracting a disease after exposure (e.g., Pfeiffer *et al*. 2004). If data are subsequently collected that follow a binomial distribution (e.g., the number of individuals in a study group that actually contract the disease), then the posterior distribution will be a beta distribution as well. The fact that the prior and posterior distributions are both beta distributions when the data are binomially distributed makes it a natural choice as a prior distribution for probabilities (Figure SP3.1.1).

Figure SP3.1.1: The beta distribution and Bayesian inference. Let's use the beta distribution to estimate the contagiousness, *p*, of a new strain of influenza. You assume, initially, that the probability that an exposed individual gets infected is likely to be around 10%, as observed for the previous strain. To account for your uncertainty, you choose a beta distribution as a prior probability distribution, $f(p)$, with parameters a_0 = 1 and b_0 = 9. You then study *n* individuals known to have been exposed to the virus, among whom *k* individuals become infected (say, k/n = 60%). The posterior probability distribution describing the contagiousness of the virus is then given by the prior probability for *p* (a beta distribution) multiplied by the probability that *k* individuals out of *n* become infected given *p* (a binomial distribution), normalized so that the posterior distribution for *p* integrates to one (see SP3.1.6). The result is a beta distribution with parameters $a = a_0 + k$ and $b = b_0 + n - k$. The posterior distribution shifts toward the observed proportion, 60%, and exhibits less variation when the dataset is large, indicating greater confidence in the estimate of *p*.



For further information on Bayesian analyses in ecology and systematics, consult The Ecological Detective (Hilborn and Mangel 1997) and Inferring Phylogenies (Felsenstein 2004).

Exercise SP3.1.1:

(a) Imagine that the apple farmer had two trees that grew from the delicious apples and that both trees survived the disease. Modify (SP3.1.5) to determine the probability that the new apple line is resistant. How many trees would there have to be in order for the farmer to infer that there is a 95% chance that the new apple line is resistant given the prior information about mutation rates? [Answer[i]]

[i] ANSWER:

(a) If both trees survive (data is "both survived"), we can rewrite (SP3.1.5) as:

$$P(\text{resistant} \mid \text{both survived}) = \frac{P(\text{both survived} \mid \text{resistant})P(\text{resistant})}{P(\text{both survived})}$$

If the new apple line is resistant, then the probability that the two trees survived is still one, so the numerator remains unchanged. The denominator, however, needs to be updated to give the total probability that both trees survive: P(both survived) = P(both survived | not resistant) P(not resistant) + P(both survived | resistant) P(resistant) = $(0.01)^2(1-10^{-6}) + (1)^2(10^{-6})$. In squaring the probability of survival, we assume that the probability of survival of two susceptible trees is independent (Rule P3.4). Thus, the overall probability that the new line is resistant given that both trees survived is $\frac{(1)^2(10^{-6})}{(0.01)^2(1-10^{-6}) + (1)^2(10^{-6})} \approx 0.01$. To infer that there is a 95% chance that the new apple line is resistant, you would have to have observed x trees of the new variety survive the disease, where $\frac{(1)^x(10^{-6})}{(0.01)^x(1-10^{-6}) + (1)^x(10^{-6})}$ must equal 0.95. Solving for x, we get $x = 3.6$, indicating that at least four trees that all survived the disease would be necessary to infer that the new apple variety is resistant.

Supplementary Material P3.2: The sum of Poisson random variables

The property that the sum of different Poisson random variables is itself Poisson distributed is extremely important, because it allows us to describe a mixture of different Poisson variables without knowing the details of each distribution separately. Let's prove this property, starting with two underlying types of events, for example, two types of mutations (e.g., transitions and transversions), each following a Poisson distribution with its own expected number of events, μ_1 and μ_2. Our goal is to prove that the probability of observing a total of k events is given by the Poisson distribution.

To observe a total of k events, there must have been some number of the first type of event (say j) and the remaining $k - j$ events must have been of the second type. The probability of observing a total of k events is thus given by the following sum:

$$\begin{aligned} P\left(X_{total} = k\right) &= \sum_{j=0}^{k} P\left(X_1 = j\right) P\left(X_2 = k - \mathrm{j}\right) \\ &= \sum_{j=0}^{k} \frac{e^{-\mu_1}\mu_1^{\,j}}{j!} \frac{e^{-\mu_2}\mu_2^{\,k-j}}{(k-j)!} \end{aligned} \quad . \qquad \text{(SP3.2.1)}$$

Mathematical software packages (like *Mathematica*) include tables of known sums and can help evaluate a sum like this. Alternatively, such sums can be solved by massaging them into the form of known sums. If we factor out $e^{-(\mu_1+\mu_2)}$ from (SP3.2.1), we are left with various factorial and power terms, which can be rewritten in terms of the binomial distribution (see Definition P3.4):

$$\frac{e^{-(\mu_1+\mu_2)}(\mu_1+\mu_2)^k}{k!} \underbrace{\left(\sum_{j=0}^{k} \frac{k!}{j!(k-j)!}\ p^j (1-p)^{k-j} \right)}_{\text{Binomial distribution}}$$

where $p = \dfrac{\mu_1}{(\mu_1+\mu_2)}$. The term in parenthesis sums over an entire binomial distribution (describing the probability of j events in k trials) and must equal one (Rule P3.10). Thus,

the probability distribution for the sum of two independent random variables that are Poisson distributed becomes:

$$P\left(X_{total} = k\right) = \frac{e^{-(\mu_1+\mu_2)}\left(\mu_1 + \mu_2\right)^k}{k!}, \qquad \text{(SP3.2.2)}$$

which is just the Poisson distribution for a process with an expected number of events equal to the sum, $\mu = \mu_1 + \mu_2$. This fact can be proven more easily using moment generating functions (see Exercise A5.2 in Appendix 5).

With *n* random variables, each following a Poisson distribution with mean μ_1, …, μ_n, respectively, we can apply the above proof to any pair of these random variables, reducing the number of random variables by one. Following this process repeatedly, we conclude that the sum of the *n* random variables will follow a Poisson distribution with an expected number of events equal to the sum, $\mu = \mu_1 + \ldots + \mu_2$.

References:

Felsenstein, J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Mass.

Hilborn, R., and M. Mangel. 1997. The Ecological Detective. Princeton University Press, Princeton, NJ.

Pfeiffer, R. M., S. Mbulaiteye, and E. Engels. 2004. A model to estimate risk of infection with human herpesvirus 8 associated with transfusion from cross-sectional data. Biometrics 60:249-56.