Components of a Bayesian model

Building a simple Bayesian model

1. Data story: motivate the model by narrating how the data might arise. In more technical terms: what is the data-generating process? Can we identify a relevent model with associated probability distribution function(s) that could plausibly give rise to the data?
2. Update: educate your model by feeding it the data
3. Evaluate: supervise, critique, and revise your model

Data story: globe tossing

How much of the earth's surface is covered in water? To answer this question, you have at your disposal a perfect beach ball replica of the earth. You toss the ball up high, with a bit of spin, catch and point to a spot - and determine whether you pointed at land or ocean. In the first nine tosses, you observe that:

W L W W W L W L W

How did these observations arise?

1. The true proportion of water covering the globe is p
2. A single toss of the globe has a probability p of producing a water (W) observation, and a $1 - p$ probability of producing a Land (L) observation
3. Each toss of the globe is independent of the others

Divide the world (*variables*) into observed and unobserved:
Observed: W and L; N = W + L
Unobserved: p

We are interested in *estimating* p using our observations. To do this, we will rely on probability theory.

Probability theory

We use probability theory as a tool for counting:

1. The number of ways each conjecture could produce an observation
2. The accumulated number of ways each conjecture could produce the entire dataset
3. The initial plausibility of each conjectured cause of the data

More formally, we build a statistical model with these ideas in mind, but call them:

1. A likelihood (which counts the number of ways each *conjecture* - i.e., potential parameter value - could produce a single observation)
2. The total likelihood of *all* the observations, given a conjecture (assuming independence, you can get the total likelihood by calculating the product of the individual likelihoods)
3. A prior probability distribution

We are interested in evaluating this expression:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}$$
$$Pr(\theta|D) \propto Pr(D|\theta) \times Pr(\theta)$$

where θ represents the parameter(s) of our model, which I sometimes also write as H. The $\propto$ symbol highlights the fact that we have left out the denominator to the equation (which is Bayes rule). We don't concern ourselves with it too much here, but the denominator is critical to make sure that the result is a *bona fide* probability distribution that sums or integrates to 1. More on those details later.

Likelihood

We use mathematical functions to count up the number of ways each conjecture could produce the dataset. These functions are called *probability distribution functions*, and when they are applied to data (i.e., an observed variable), they are often called *likelihood* functions.

Bayesians consider the data D to be fixed after observation, whereas the parameters θ are random variables.

Frequentists are the opposite: D is random and θ is fixed. The likelihood, in a frequentist sense, is written as $\mathcal{L}(\theta|y) = \Pi_{i=1}^{n} f_Y(y_i|\theta)$.

For the globe tossing data story, the probability distribution function we'll use to calculate the likelihood follows directly from the data story. With practice, you will be able to identify plausible probability distribution functions for many different situations.

The binomial distribution

For a discrete random variable Z, the number of 'successes' (z) over a given number (N) of trials is governed by the probability (p) of success:

$$F_Z(z \mid N, p) = \binom{N}{z} p^z (1-p)^{N-z}$$
$$\binom{N}{z} = \frac{N!}{z!(N-z)!}$$

where $\binom{N}{z}$ can be read aloud as 'N choose z'. It is also known as the binomial coefficient.

In our globe tossing example, what is z, N?

Let's get a bit of intuition for the binomial distribution. Consider our specific sequence:

W L W W W L W L W

The probability of this sequence, assuming independence of individual tosses, is the product of Bernoulli probabilities, $\Pi_{i=1}^{n} p^z (1-p)^{1-z}$. In our globe tossing example, a Bernoulli trial is the outcome of a single toss (i.e., experiment): either a water (W = 1) or land (L = 0).

This sequence resulted in 6 W's and 3 L's. But there are many other sequences that would result in the same set of observed outcomes. The $\binom{N}{z}$ counts the number of ways of allocating z W's in N tosses, without double counting equivalent sequences.

If we have z W's in N tosses, we can allocate the first W into any of the N slots. The second W can go into any of the remaining $N-1$ slots.The third W can go into any of the remaining $N-2$ slots. And so on until, the zth W goes into the remaining $N-(z-1)$ slot. Multiplying these possibilities means that there are $N \times (N-1) \times (N-2) \times ... \times (N-(z-1))$ ways of allocating z W's to N tosses of the globe. This statement is algebraically equivalent to $N!/(N-z)!$, where ! denotes factorial. To remove the duplicates, we divide the numerator by the number of ways z W's go into z slots, i.e., $z!$. In case you need a refresher on combinatorics, visit: https://www.mathsisfun.com/combinatorics/combinations-permutations.html.

Globe tossing example

We now apply the binomial distribution to our problem. Let's start by assuming that there is equal probability of Water and Land (i.e., $p = 0.5$).

We'll do this 'by hand' in `R`.

```
library(tidyverse)
library(rethinking)
```

We start by simplifying the binomial coefficient (though this is not necessary, it will help demystify it):

$$
\begin{aligned}
F_Z(z \mid N, p) =& \frac{N!}{z!(N-z)!} p^z (1-p)^{N-z} \\
=& \frac{9!}{6!(9-6)!} p^z (1-p)^{N-z} \\
=& \frac{9!}{6!(9-6)!} p^z (1-p)^{N-z} \\
=& \frac{9 \times 8 \times 7}{3!} p^z (1-p)^{N-z}
\end{aligned}
$$

```
p <- 0.5
N <- 9
z <- 6
bin_coef <- (9*8*7) / (factorial(3))
bin_coef * (p ^ z) * ((1 - p)^(N - z))
```

```
[1] 0.1640625
```

We can check our answer using the function `dbinom`:

```
dbinom(x = z, size = N, p = p)
```

```
[1] 0.1640625
```

The `d` in `dbinom` stands for 'density', which is short for probability density. What matters here is that this function returns the probability of observing 6 W's out of 9 globe tosses, given $p = 0.5$. Later, we'll evaluate the likelihood for other values of p. (we usually refer to distribution functions for *discrete* outcomes, like the counts in a binomial distribution, as probability *mass* functions; the term 'probability *density* function' typically is reserved for *continuous* outcomes; but I suspect this is just a function naming convention in `R`).

Priors

For every parameter (unobserved random variable) in a Bayesian model, we must provide an initial probability distribution - otherwise known as the *prior*. Priors must be given careful thought, and so-called 'uninformative' or 'flat' priors are rarely the best choice, and sometimes can have unintended consequences that result in misguided inferences. The choice of prior, like the many other choices made in a modeling exercise, should be evaluated and revised with care.

Nevertheless, for this exercise, we'll use a uniform (i.e., flat) prior that gives equal probability to all possible values of p from 0 to 1.

A model is born

We now express our Bayesian machine mathematically:

$$
\begin{aligned}
z &\sim \text{Binomial}(N, p) \\
p &\sim \text{Uniform}(0, 1)
\end{aligned}
$$

where z represents the number of times water was observed.

So how do we actually go about estimating p using our observations? For this, we need some fundamental rules of probability, and in particular, we need Bayes theorem. We will derive Bayes theorem in a separate lesson on the rules of probability.

Bayes theorem

Here is Bayes theorem, where θ represents the unobserved parameter(s) of our model, and d represents the observed data:

$$\Pr(\theta \mid d) = \frac{\Pr(d \mid \theta)\ \Pr(\theta)}{\Pr(d)}$$

And this expression states what we want to estimate: the posterior probability of our model, conditional on the data that we have observed (i.e., $\Pr(\theta \mid d)$).

So, applying this to our globe tossing problem, we get the following expression:

$$\Pr(p \mid z, N) = \frac{\Pr(z, N \mid p)\ \Pr(p)}{\Pr(z, N)}$$

$$\text{posterior} \ = \frac{\text{likelihood} \times \text{prior}}{\text{average likelihood}}$$

The likelihood can also be called the 'probability of the data' (McElreath switched from the first edition to the second, probably to distinguish between Frequentist and Bayesian likelihood).

The job of the denominator is to normalize the numerator so that the posterior probability distribution sums / integrates to 1. To accomplish this, we need the *law of total probability* (to be defined elsewhere).

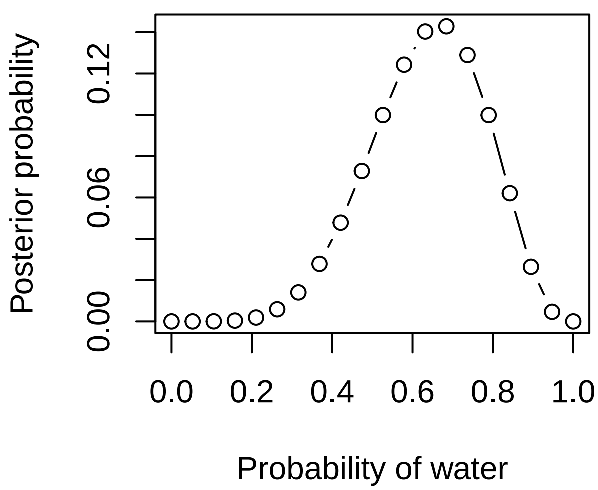
Using grid approximation to calculate the posterior

```r
# define grid
p_grid <- seq(0, 1, length.out = 20)
# define prior
prior <- rep(1, 20)
# compute likelihood at each value in grid
likelihood <- dbinom(x = 6, size = 9, prob = p_grid)
# compute product of likelihood and prior
unstd_posterior <- likelihood * prior
# standardize the prior, so it sums to 1
posterior <- unstd_posterior / sum(unstd_posterior)
sum(posterior)
```

```
[1] 1
```

Let's plot this:

```r
plot(p_grid, posterior, type = "b",
     xlab = "Probability of water", ylab = "Posterior probability")
```



Attribution

These notes are based on chapter 2 of *Statistical Rethinking* (McElreath 2020).

McElreath, Richard. 2020. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. CRC Press.