Quantitative Understanding in Biology
1.7 Bayesian Methods

Jason Banfelder

September 30th, 2021

1 Introduction

So far, most of the methods we've looked at fall under the heading of classical, or 'frequentist' statistics. In this school of thought, one designs an experiment and, for the most part, uses only the collected data to draw inferences. One of the strengths of this approach is that it is objective; your biases and pre-conceived notions do not play a role in the inference process (or at least they shouldn't). There is, however, another school of thought that says that we **should** incorporate prior knowledge or beliefs into our analyses. After all, we are all trying very hard to understand complex biological systems; why should we hamstring ourselves by ignoring most of what we know when we are interpreting our data? The Bayesians use a framework for incorporating prior beliefs into analyses.

In this section, we'll understand Bayes' rule (a useful tool in its own right), and then have an introductory look at contemporary Bayesian methods.

2 Conditional Probability

Imagine we choose a student at random from the class, and want to make a statement about that student's height. If we assume that heights of students are normally distributed, and we have an estimate of the mean and standard deviation of students' heights (from data that we collected in the past), we can use this information to compute the 95% CI for the heights of sampled students, or compute the probability that the student sampled will be over 6 feet tall. We'll label the event that the student is at least 6 feet tall T; the probability of this happening is written as $P(T)$.

Now imagine that after we choose a student, but before you measure her height, you are told that the student is female. Since we know that sex and height are related, this would

influence your computation of the probability that the student is over six feet tall. If we label the event that a selected student is female as F, then we denote the probability that a selected student is at least six feet tall, given that she is female, as $P(T|F)$. This is called a conditional probability. In general, you might expect that $P(T) > P(T|F)$, because you know that women tend to be shorter than men (can you think of a case where this may not be true?).

The interplay between T and F goes both ways. If you select a student and are told that the student is over six feet tall, then you'll probably want to lower the probability that the selected student is a female. In general, you'd say $P(F|T) < P(F)$.

3 Joint Probability and Bayes' Rule

The probability that two events occur together is called their joint probability. Continuing our example, we'd denote the probability of choosing a female over six feet tall from a group of students as $P(T, F)$. You may see this written as $P(T \cap F)$. You can also write it as $P(F, T)$ or as $P(F \cap T)$.

The probability of a joint event is related to conditional events. We write:

$$P(T, F) = P(F) \cdot P(T|F) \tag{1}$$

This says that the probability of choosing a tall female is equal to the probability of choosing a female, times the probability that a female is tall. But we can think about this the other way as well, and write:

$$P(T, F) = P(T) \cdot P(F|T) \tag{2}$$

which says that the probability of choosing a tall female is equal to the probability of choosing a tall person, times the probability that a tall person is a female.

We can combine the two equations above and write

$$P(T|F) = \frac{P(F|T) \cdot P(T)}{P(F)} \tag{3}$$

This is true for any two events, and when written in terms of generic events A and B, we have Bayes' rule:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{4}$$

4 An Application of Bayes' Rule

Applying Bayes' Rule is pretty straightforward, although it can often lead to some counterintuitive results. Let's work out a canonical example, based on these three assumptions:

- 1% of women in a given population have breast cancer.
- If a woman has breast cancer, there is a 90% chance that a particular diagnostic test will return a positive result.
- If a woman does not have breast cancer, there is a 10% chance that this diagnostic test will return a positive result.

We wish to compute the probability that a woman with a positive test result actually has cancer. Looking at what we know, the test looks pretty reasonable: it has a 10% false positive rate and a 10% false negative rate, so your intuition may tell you that the probability that your positive-testing patient actually has cancer is pretty high.

We'll denote having cancer as C, not having cancer as H, getting a positive test result as '+', and getting a negative test result as '-'. Using this nomenclature, we are being asked to compute $P(C|+)$.

Bayes' Rule tells us that this must be

$$P(C|+) = \frac{P(+|C) \cdot P(C)}{P(+)} \tag{5}$$

We know that $P(+|C) = 0.9$, and we know that $P(C) = 0.01$. The tricky part is computing the denominator, $P(+)$. In this case, there are only two ways to get a positive result from our diagnostic: true positives, and false positives. We can write:

$$P(+) = P(\text{true positive}) + P(\text{false positive}) \tag{6}$$

$$P(+) = P(+|C) \cdot P(C) + P(+|H) \cdot P(H) \tag{7}$$

$$P(+) = 0.9 \cdot 0.01 + 0.1 \cdot 0.99 \tag{8}$$

$$P(+) = 0.009 + 0.099 \tag{9}$$

$$P(+) = 0.108 \tag{10}$$

We should stop and think about what these results mean. We see here that nearly 11% of our tests will return a positive result, even though the overall cancer rate is only 1%. The worked out calculation shows you that the overwhelming majority (about 10 to 1) of positive results are actually false positives!

To complete our calculation, we now have:

$$P(C|+) = \frac{P(+|C) \cdot P(C)}{P(+)} = \frac{0.9 \cdot 0.01}{0.108} = 8.3\% \tag{11}$$

So even in the face of a positive result, the probability that our hypothetical patient has breast cancer is quite low. Is this what your intuition told you when we started?

5 Bayesian Methods

Contemporary Bayesian methods take Bayes' Rule, and apply it to estimating model parameters. In some sense, you can think of the process as a framework for model fitting. In the process, we have the opportunity to incorporate prior knowledge or beliefs.

Let's imagine that we flip a coin 20 times, and observe 13 heads. We wish to estimate the true probability that the coin, over a large series of tosses, will come up heads. We'll call this probability the coin's bias (note that a bias of 0.5 is actually a completely unbiased, or fair, coin). You (hopefully!) recall that you can do this with a binomial test.

```
binom.test(13, 20)
```

```
##
##  Exact binomial test
##
## data:  13 and 20
## number of successes = 13, number of trials = 20, p-value =
## 0.2632
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.4078115 0.8460908
## sample estimates:
## probability of success
##                   0.65
```

According to this frequentist approach, your best guess is that the bias is 65%, and the 95% CI of the bias ranges from 41% up to 85%. Another way of thinking about this is that, given only the data you have (13 out of 20 tosses turned up heads), the most

likely probability for getting a head is 0.65, and the values between 0.41 and 0.85 are all plausible.

Let's see how we might get at this result, thinking like a Bayesian. Bayes' Rule is written in terms of two events (e.g., being tall and being female, or having cancer and testing positive for cancer). When working problems like this, the two 'events' are typically of different kinds: one relates to the model (M), and the other to the data (D).

The data events are easy to understand. In our case, we observed 13 heads out of 20 flips. We know how to compute the probability of any particular data outcome (13 heads, 10 heads, or no heads, as examples), given a particular assumption (or model) about the underlying properties of the coin. If we assume that the coin is fair, then the probability of observing 13 heads is given by:

```
dbinom(13, size = 20, prob = 0.5)
```

```
## [1] 0.07392883
```

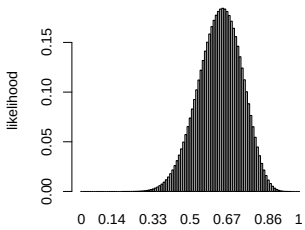
If we assume that the coin has a bias of 0.25, then the probability of observing the data that we collected is

```
dbinom(13, size = 20, prob = 0.25)
```

```
## [1] 0.0001541923
```

We might attack our problem by computing the probability of getting the data that we observed at various values of the coin's bias, and see what the maximum is. There are an infinite number of possible biases, but we'll be content with checking out 101 possibilities.

```
coin.bias <- seq(from = 0, to = 1, by = 0.01)
likelihood <- dbinom(13, 20, prob = coin.bias)
barplot(likelihood, names.arg = coin.bias, ylab = "likelihood")
```



Unsurprisingly, the maximum likelihood occurs where the bias is 0.65. The plot also gives

you a sense of what range of values for the bias is credible. Intuitively, you might conclude that a bias of 0.5 is plausible, but a bias of 0.25 is not.

It is worth noting here that the graph above is not (yet) a formal probability distribution function; the area under the curve is not necessarily one. Accordingly, we've labeled the y-axis as likelihood, and not probability density.

In the above analysis, we explored the universe of possible coin biases in a relatively objective way. We didn't make any assumptions about where the coin being investigated came from. We might imagine that, before we started, there was a collection of 101 coins, each with a different bias (0.00 – a two tailed coin, 0.01 – a very heavily weighted coin, ...), and that we chose one at random and tested that one by flipping it 20 times. Using this contrived scenario, we can now think about what the model (M) probabilities are. When we pick a coin at random, but before we test it by repeatedly flipping and observing, we know that there is an equal probability of the bias being any of the possible values (i.e., the probability distribution of the coin's bias resembles a uniform distribution). There are 101 possible models to choose from, and the probability of each one is $\frac{1}{101} = 0.0099$.

We can now use Bayes' Rule to compute the probabilities of each of the 101 coins being the one that we chose, given the data that we observed. Let's compute the probability that we chose the perfectly fair coin (we'll call this $M_{0.50}$)

$$P(M_{0.50}|D_{13}) = \frac{P(D_{13}|M_{0.50}) \cdot P(M_{0.50})}{P(D_{13})} \tag{12}$$

Now, $P(D_{13}|M_{0.50})$ is the probability of observing 13 heads when flipping a fair coin 20 times. We've already calculated that to be 0.0739. Next, we have $P(M_{0.50}) = 0.0099$, because we assume all of the 101 biases are equally likely.

The hard part is that pesky denominator. The probability of picking one of our 101 coins, then flipping it 20 times and observing 13 heads, is...

$$P(D_{13}) = P(D_{13}|M_{0.00}) \cdot P(M_{0.00}) + P(D_{13}|M_{0.01}) \cdot P(M_{0.01}) + P(D_{13}|M_{0.02}) \cdot P(M_{0.02}) + \ldots \tag{13}$$

This sum can be computed easily:

```
(p.d13 <- sum(dbinom(13, 20, coin.bias) * (1 / 101)))
```

```
## [1] 0.04714757
```

We now have...

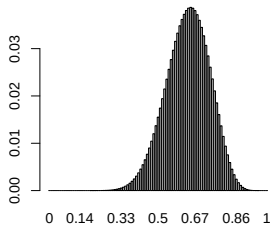
$$P(M_{0.50}|D_{13}) = \frac{P(D_{13}|M_{0.50}) \cdot P(M_{0.50})}{P(D_{13})} = \frac{0.0739 \cdot 0.0099}{0.04715} = 0.0155 \tag{14}$$

We can go ahead and compute $P(M_x|D_{13})$ for all 101 possible values of x, and plot that. Note that the denominator is always the same, for all of the x values.

```
posterior.probability <- dbinom(13, 20, coin.bias) * (1 / 101) / p.d13
sum(posterior.probability)
```

```
## [1] 1
```

```
barplot(posterior.probability, names.arg = coin.bias)
```



Now this is a *bona fide* probability distribution function. We've confirmed that we got everything right by making sure that all of the probabilities summed to one. We can use this distribution to pull out a 95% CI for the underlying coin bias. This is a bit trickier than you might imagine, but if you eyeball it, you'll see that this is consistent with the 95% CI that came from the binomial test.

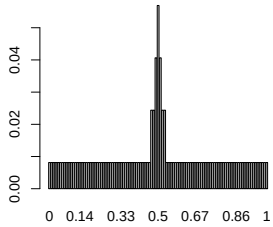
6 Prior Probabilities

Phew! That was a long way to go; you might be asking yourself why we'd ever want to follow this somewhat convoluted line of reasoning when typing `binom.test` works just fine.

The answer is that this analysis made one assumption – that all coin biases were equally likely – that you in fact didn't have to make. You could have modeled any other assumptions or beliefs that you wanted to.

For example, if your classmate was using a coin to decide who buys dinner tonight, you might want to factor in your prior beliefs about how honest he is, as well as how fair a coin you think he has access to. You might use a prior distribution (that's what we call $P(M_x)$) to capture your belief that the coin is most likely very close to fair.


```
prior.probability <- numeric(101)
prior.probability[0:101] <- 1
prior.probability[48:54] <- 3
prior.probability[50:52] <- 5
prior.probability[51] <- 7
# Normalize; since it is a PDF, sum must be 1.0
prior.probability <- prior.probability / (sum(prior.probability))
barplot(prior.probability, names.arg = coin.bias)
```



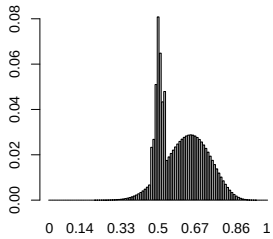
The denominator is now...

```
(p.d13 <- sum(dbinom(13, 20, coin.bias) * prior.probability))
```

```
## [1] 0.05205051
```

...and the posterior probability distribution is...

```
posterior.probability <-
  dbinom(13, 20, coin.bias) * prior.probability / p.d13
barplot(posterior.probability, names.arg = coin.bias)
```

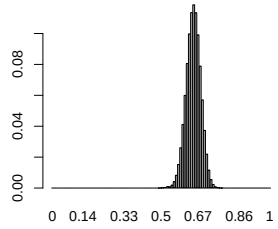


You can see here that the posterior probability distribution reflects a complex interplay

between the prior and the data. While we could compute a 95% CI for the coin's bias from this distribution, that interval would not capture the whole of what we can infer; in particular that the distribution of credible values is bimodal.

Let's see what happens when we use the same prior, but collect more evidence. We'll imagine that we've flipped the coin 200 times, and observed 130 heads. This is the same ratio as before.

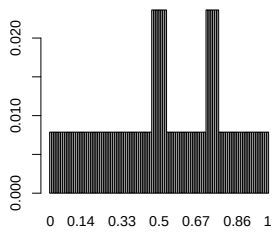
```
p.d130 <- sum(dbinom(130, 200, coin.bias) * prior.probability)
posterior.probability <-
  dbinom(130, 200, coin.bias) * prior.probability / p.d130
barplot(posterior.probability, names.arg = coin.bias)
```



Here you can see that the evidence provided by the data is overwhelming the prior.

As another example, imagine that the person flipping the coin looks to you like a bit of a swindler, and you also happen to know that there is a magic shop around the corner that sells coins with a bias of around 0.75. In this case you might employ a different prior.

```
prior.probability <- numeric(101)
prior.probability[0:101] <- 1
prior.probability[48:54] <- 3
prior.probability[73:78] <- 3
# Normalize; since it is a PDF, sum must be 1.0
prior.probability <- prior.probability / (sum(prior.probability))
barplot(prior.probability, names.arg = coin.bias)
```

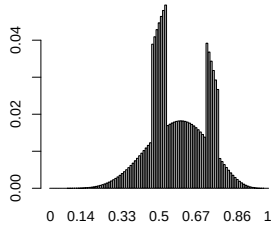



If we observe 6 heads out of 10 tosses in this case, we'd compute the posterior probability distribution as:

```
(p.d6 <- sum(dbinom(6, 10, coin.bias) * prior.probability))

## [1] 0.1083962

posterior.probability <- dbinom(6, 10, coin.bias) * prior.probability / p.d6
barplot(posterior.probability, names.arg = coin.bias)
```



As you can see, using this kind of Bayesian formulation, you can incorporate into your model just about any prior knowledge or beliefs that you feel is appropriate, and the result will be a rich and detailed description of the results of the interplay between that and your data. This can be a two-edged sword. If your prior is too definitive, you'll need to collect an overwhelming amount of data to contradict it. When critically evaluating studies that use Bayesian methods, it is a good idea to appreciate what prior was used. A weak or non-committal prior may be OK, but a strong prior should be backed up by a wealth of evidence, and hopefully not an arbitrary belief or biased expectation.

7 Going Further

The application of Bayesian methods gets interesting when there are multiple parameters in the model. Imagine that we extended our trivial example just a bit to involve two coins. The number of models to choose from now gets significantly larger: assuming that we continue to consider 101 possible biases per coin, there are now over 10,000 models. Evaluating that 'pesky' denominator just got a lot more challenging. The reward of dealing with this problem is that the result will be a joint distribution of the two coins that contains information about any interactions between the coins.

It turns out that you can get very good estimates of the joint distribution of these two parameters not by fully evaluating the denominator and all of the possible numerators, but rather by sampling combinations of parameter values according to specific rules that optimize the search. These methods are not practical to ever carry out by hand, but now that we all have laptops with several fast processors, they are accessible to just about everyone.

If you want to pursue this topic further, the text "Doing Bayesian Data Analysis" by John Kruschke is a great place for a beginner to start (make sure you consult the 2nd edition of this book).

