Introducing Bayesian Analysis

Charles DiMaggio, PhD[1,2]

August 9, 2012

[1] Department of Anesthesiology, Columbia University, College of Physicians and Surgeons, New York

[2] Department of Epidemiology, Columbia University, Mailman School of Public Health, New York

1 What is Bayesian Analysis?

> A Bayesian is one who, vaguely expecting a horse, and catching a glimpse of a donkey, strongly believes he has seen a mule.

1.1 What are the chances?

Probability is at the heart of much rational thinking. [1] Consider the following quick series of calculations that might start your morning. What should I wear today? It's fall and usually cold, so a jacket might make sense. My weather app shows bright sun and temperatures in the 70's. But it was a free app, and it's wrong a lot of the time. OK. A light jacket it is.

At each stage, you are evaluating probabilities against evidence, then revising the probabilities. This is, essentially, Bayesian reasoning. Bayesian statistics allow us a method of calculating precise, numeric estimates of the kinds of probabilities that we often take for granted in our day-to-day reasoning. These kinds of estimates can be applied to the work of epidemiology.

[1] If you are not at all familiar with probability theory, I have a paper on the topic.

1.2 Conditional Probability

2 Deriving Bayes Theorem

Bayes Theorem, named after the Rev. Thomas Bayes[2], is strikingly simple in its derivation, but profound in it's implications. If you know the probability that event A occurs given that event B occurs, like the probability of the New York Mets winning the world series given that they recruited some hot new ace pitcher, Bayes theorem states that we can use that information to calculate the probability that event B occured, given event A occurred. In our example, that the Mets recruited some ace pitcher given that they won the World Series. In probability nomenclature, Bayes Theorem says that if we know $Pr[A|B]$ we can get at $Pr[B|A]$.

Deriving Baye's Theorem from first principles is fairly straightforward:

$$Pr[A \cap B] = Pr[B \cap A] \tag{1}$$
$$Pr[A \cap B] = Pr[A|B]Pr[B] \tag{2}$$
$$Pr[B \cap A] = Pr[B|A]Pr[A] \tag{3}$$
$$Pr[A|B]Pr[B] = Pr[B|A]Pr[A] \tag{4}$$
$$Pr[A|B] = \frac{Pr[B|A]Pr[A]}{Pr[B]} \tag{5}$$
$$Pr[B|A] = \frac{Pr[A|B]Pr[B]}{Pr[A]} \tag{6}$$

We are going to pay close attention to the denominator in those last two equations. By the law of total probability, $Pr[A] = Pr[A|B]Pr[B] + Pr[A|\overline{B}]Pr[\overline{B}]$, so

$$Pr[B|A] = \frac{Pr[A|B]Pr[B]}{Pr[A|B]Pr[B] + Pr[A|\overline{B}]Pr[\overline{B}]} \tag{7}$$
$$Pr[B|A] = \frac{Pr[A|B]Pr[B]}{\sum Pr[A|B]Pr[B]} \tag{8}$$

This equation is for categorical or discrete events, but when we work in probabilities we are frequently interested in probability *distributions* which sum to one, so the equation may more properly be written:

[2]Turnbridge Wells, England, c1763

$$Pr[B|A] = \frac{Pr[A|B]Pr[B]}{\int dB Pr[A|B]Pr[B]} \tag{9}$$

2.1 Example: Positive Predictive Value

The classic example of Bayes Theorem from clinical epidemiology is that of the sensitivity and specificity of a medical test vs. its positive and negative predictive values. in the 1990's, ELISA was used as a screening test for HIV. It was documented as being nearly 100% sensitive and 100% specific for detecting HIV. If the ELISA came back negative, the person was informed they were HIV negative. But, if the ELISA screening test came back positive, a confirmatory Western Blot test was performed. If both the ELISA and the Western Blot were positive, the person was informed they were HIV positive. If the Western Blot was negative, the sample was discarded and person was not informed about their positive screening test. Why? It's a matter of conditional probabilities.

Consider what the sensitivity and specificity of a test tell us. The sensitivity of a test is the probability that a person will test positive given that they have the disease ($Pr[+|D)$. The specificity of a test is the probability that a person will test negative given that they do not have the disease ($Pr[-|\overline{D})$. It is an attribute of the test and invariant. But it is not what a patient and his or her health care provider are interested in. Rather than wondering if they will test positive if they have the disease, they are concerned whether they have the disease given that they tested positive (or don't have the disease, given that they tested negative). And this is a very different question. These measures are called the positive and negative predictive values of a test.

Consider the following table based on actual data:

elisa	western blot		
	pos	neg	
pos	498	4	502
neg	10	488	498
	508	492	1000

Let's use our derived Bayes Theorem formula to calculate the sensitivity of the ELISA screening test.[3] The sensitivity of a test is the $Pr[+|D]$ which we can calculate from the table as as:

[3]There are more intuitive ways of calculating these values from a simple table like this. But we want to illustrate the formula, so bear with me.

$$\frac{Pr[D|+]P[+]}{Pr[D|+]P[+] + Pr[D|-]P[-]} \tag{10}$$

$$\frac{(498/502)(502/1000)}{(498/502)(502/1000) + (10/498)(498/1000)} \tag{11}$$

In R, we write:

```
((498/502)*(502/1000))/(((498/502)*(502/1000)) + ((10/498)*(498/1000)))
```

And our result is, as advertised, a sensitivity of 98%.

The calculation for positive predictive value is:

$$Pr[D|+] = \frac{Pr[+|D]P[D]}{Pr[+|D]P[D] + Pr[+|\overline{D}]P[\overline{D}]} \tag{12}$$

$$= \frac{(498/508)(508/1000)}{(498/508)(508/1000) + (4/488)(492/1000)} \tag{13}$$

Which results in a positive predictive value of 99%. So, then, why the confirmatory Western Blot? Well, did you notice the prevalence of disease in this population? More than half of this group is HIV positive. In fact, they were a group of IV Drug Users in the New York City area. What if we were to make these calculations with a group with a prevalence more in line with the general population? Since the sensitivity and specificity of a test is an attribute of the test itself, we can apply those proportions to a hypothetical population of a million people, with a disease prevalence of 0.02%:

elisa	western blot		
	pos	neg	
pos	1960	784	9944
neg	40	990016	990056
	2000	998000	1000000

A positive predictive value using these numbers is:

$$\frac{(1960/2000)(2000/1000000)}{(1960/2000)(2000/1000000) + (784/998000)(998000/1000000)} \tag{14}$$

For a positive predictive value of about 20%.

We can see a couple of things here. First, Bayes Theorem works, and that's certainly nice enough. Second, we can use Bayes theorem to derive probabilities for separate

but related events, and that's actually quite provocative. Third, some terms we use in epidemiology, like prevalence, may have different names, like $Pr[D]$ in other statistical traditions. That is only to be expected, but can pose an initial obstacle to understanding Bayesian statistics.

2.2 A More Intuitive Approach to Bayes Theorem

Bayes Theorem can be a bit of a mind scrambler when you first encounter it. There is, though, a more intuitive approach to it than that of deriving probability formulas. A hint at this approach can be got at by considering the cells of the ELISA tables above. This is a familiar approach to epidemiologists. So for example, when we calculated the positive predictive value of ELISA in a population, we could have used the table cells and said that the probability of someone being diseased given that they have a positive test is simply the number of diseased persons among all those with a positive test, or (in the classic epi 2-by-2 table terminology $a/a + b$. Similarly the sensitivity of a test is $a/a + c$. All we are doing is restricting our attention (and basing our denominator) on either the row total or the column total.

This approach can be applied to tables of probabilities to arrive at a more intuitive appreciation of Bayes Theorem. Say we have two separate but related events that can occur. For example we toss a coin three times and are interested in (1) the number of heads, and (2) the number of times we 'switch' from head to tails. Here is the total sample space:

result	heads	switches	probability
TTT	0	0	1/8
TTH	1	1	1/8
THT	1	2	1/8
THH	2	1	1/8
HTT	1	1	1/8
HTH	2	2	1/8
HHT	2	1	1/8
HHH	3	0	1/8

We can summarize the probability of the *joint probability* of these two outcomes (total number of switches, total number of heads) in a single table:

	heads			
switches	0	1	2	3
0	1/8	0	0	1/8
1	0	2/8	2/8	0
2	0	1/8	1/8	0

taking this one step further, we can sum the joint probabilities to arrive at the

marginal probabilities

	heads				
switches	0	1	2	3	
0	1/8	0	0	1/8	2/8
1	0	2/8	2/8	0	4/8
2	0	1/8	1/8	0	2/8
	1/8	3/8	3/8	3/8	

The final row, is the marginal probability for the number of heads. The final column is the marginal probability for the number of switches. Notice that the marginal probabilities sum to one when we add them up across the final row or column. This makes sense, because we are accounting for every possible outcome and one of them *must* occur.

Consider how we can use this table to arrive at a conditional probability. The probability of one switch, given one head is $(2/8)/(3/8) = 2/3$ All we need to do is look at the cell for one switch that is contained in the column for one head. Now, say we wanted to know the probability of one head given one switch. Now, we restrict our attention to the row that corresponds to one switch, and select the cell for one head to arrive at $(2/8)/(4/8) = 1/2$. This is really what we are doing when applying Bayes Theorem. The joint probability of a single cell can be seen relative to the column total or the row total. This simple idea of joint and marginal probabilities will become exceedingly important when we begin to discuss sampling approaches to solving Bayesian problems. [4]

3 Applying Bayes Theorem to Models and Data

All this talk of probabilities is well and good, but we know that as epidemiologists and scientists we live in a world of regression equations, models and data. It is a relatively simple step from the world of probability to the world of models and data.

[4] Another, intuitive example comes from Michael Starbird and involves (as it so often does in probability) playing cards. It begins with the question: What is the probability of choosing a red card that is a face card. It is, clearly, the same as the probability of choosing a face card that is also a red card. This is saying that $Pr[face \cap red] = [red \cap face]$ Think of the marginal and conditional probabilities upon which the joint probabilities are based. The joint probability of choosing a face card that is red is the product of the marginal probability of choosing a face card (12/52) and the conditional probability of choosing a red card given that we have chosen a face card (6/12). Alternatively, the joint probability of choosing a red card that is a face card is the product of the marginal probability of choosing a red card (26/52) time the conditional probability of choosing a face card given that we have chosen a red card (4/12) Feel free to satisfy yourself that those two calculations equal the same probability. Then, picture the two marginals meeting at the same 'square' on a table of joint probabilities.

We need only adapt the table of joint probabilities so that the rows correspond to data (our observations) and the columns correspond to models (our parameters).

Our model values are the probabilities of observing the data values given our choice of model and parameters. For example, say we are flipping a coin [5] and have three candidates for a binomial model parameter $\theta = .25, \theta = .5, \theta = .7$ We generate data by flipping our coin 12 times. Say we observe 3 heads. We can calculate the *likelihood* of observing that data given our model for each of the three possible parameters by applying the binomial formula: $Pr[data|\theta] = \theta^3(1-\theta^9)$ Let's set up a quick function in R:

```
prob<-function(x){
  (x^3)*((1-x)^9)
}

v<-c(0.25,0.5, 0.75)

prob(v)
```

The three likelihoods that correspond to our three possible $\theta's$ are :

```
[1] 1.173198e-03 2.441406e-04 1.609325e-06
```

The *maximum* likelihood corresponds to a value for θ of 0.25. [6]

Notice something: we have arrived very neatly at the $Pr[data|\theta]$, but much like in the case of our screening tests, is this really what we want to know? We are, in fact, interested in $Pr[\theta|data]$. But we now know we can get at that with Bayes Theorem:

$$Pr[\theta|data] = \frac{Pr[data|\theta]Pr[\theta]}{Pr[data]} \tag{15}$$

Which leads us back to our table of joint and marginal probabilities:

data	model (θ value)	(marginal)
(data value)	$Pr[\theta\|data] * Pr[data] = Pr[data\|\theta] * Pr[\theta]$	$Pr[data]$
(marginal)	$Pr[\theta]$	

[5] There is a lot of coin flipping, card playing and plucking colored objects from urns in probability theory...

[6] This is, in fact, the maximum likelihood estimate or MLE

And we can again intuit that applying Bayes Theorem is like restricting our attention to one or the other set of marginal probabilitis.

4 Why do Bayesian Analysis

You may by this point have a better appreciation of what Bayesian analysis is. But why should you care? From a purely statistical perspective, we analyze data for three main reasons: to estimate parameters, to predict, and to compare models. Bayesian analysis offers advantages in all three areas. First, we can *directly estimate* the probability of parameters. This is simply not possible with more traditional frequentist statistics. Second, we can *directly predict* data values because we can explicitly calculate the probability of a value given all possible values (the evidence) Third, model comparisons are enhanced because Bayesian analyses inherently *account for model complexity.* In general, we penalize more complex models because they are more likely than simpler models to account for *any* data, including random data. A simpler model, in Baysian terms, will parcel out the total probability to few values. If the data are closer to those values, the model will be more likely. A more complex model. with more possible parameter values will be less likely to have a strong probability for any particular value.

But perhaps the most compelling reason to do Bayesian analysis is that it offers an attractively intuitive yet methodologically rigorous way of updating our knowledge. If our knowledge is scant or non-existent, the results will be, essentially, the same as the maximum likelihood estimate. But, if we have some prior or pre-existing knowledge, shouldn't we use it? Rational thought is at its core a matter of updating our beliefs, and the very role of data is to influence those beliefs.

5 Objections to Bayesian Analysis

Bayesian analysis has been, and in some quarters continues to be, the subject of debate and controversy. There are a number of objections, but three common ones are that the calculations, especially for realistic and complicated models, can be statistically intractable, priors are capricious, idiosyncratic or unknowable, and our 'beliefs' should not influence data-based analyses.

The intractability issue was, in fact, true for many years. Bayesian calculations involve summing the probability of the data (the 'evidence' in the denominator of Bayes Theorem). This can be a difficult integral. For many years we were limited to 3 possible approaches. We could choose a prior that was conjugate to the likelihood which makes the posterior come out the same as the prior. We could approximate the true but intractable function with other functions that were easier to work with.

Or, we could approximate the integral by using a grid of values. Each of these approaches had their limitations, particularly for the kind of complex problem with which we are most often interested.

Over the last 10 or 15 years, though, simulation based approaches, particularly Markov Chain Monte Carlo methods, and the concomitant availability of sufficient computing power to usefully apply them, have largely obviated the issue of calculations, leading to an explosion of interest in Bayesian methods.

As for capricious priors, well I suppose that could be a problem in theory, but it is rarely one in practice. In fact, priors almost invariably come from the published literature, generally agreed upon facts or accepted theories. At the very least, Bayesian analysis forces us to consider our assumptions.

Finally, as I alluded to above, our beliefs or prior assumptions are invariably the starting point for the advancement of knowledge. In scientific endeavors, those beliefs are usually informed by previous evidence, sometimes in the form of data and studies, sometimes in the form of individual experience, sometimes in the form of received wisdom. We instinctively evaluate additional evidence, experience or information in light of those beliefs or prior assumptions. Bayesian analysis simply formalizes and quantifies the process.