**Baye’s Theorem**

And now let’s write down Baye’s Theorem. First consider the simple case of two events *a* and *b*.



So we have:



Now presume the set of variables aj are complete, mutually exclusive, and that we have P(a1, a2, …, an) or whatever. So then we can write P(b) = ΣajP(aj)P(b|aj). And then the probability of aj given some other event b is:



and so we have:



Important Point,

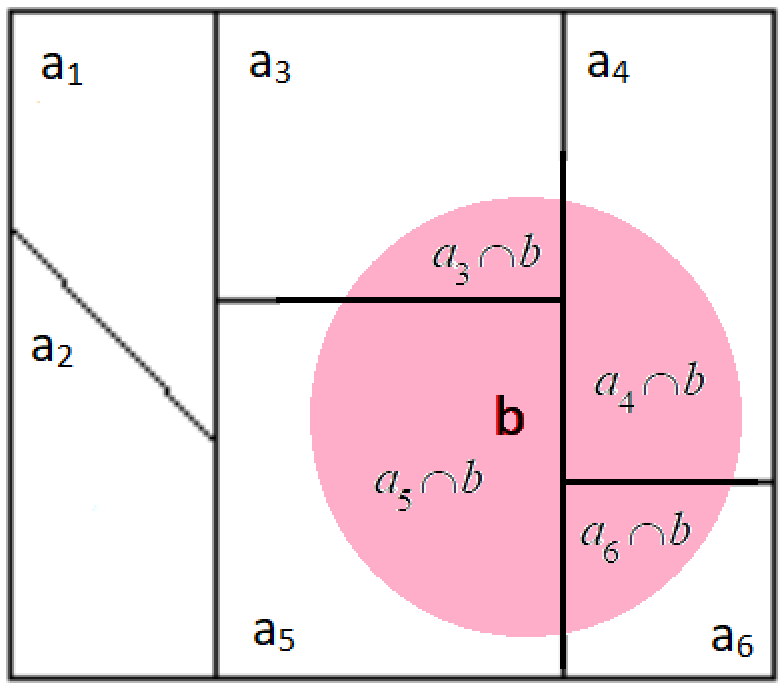
Baye’s theorem essentially quantifies scientific inductive reasoning. We have a set of mutually exclusive ‘hypotheses’ aj, with a priori probabilities P(aj). Each imply an event b with some probability P(b|aj), which could be 0, and could be 1, or anything in between. Then supposing we do an experiment and confirm b, we want know how to update the probability of these hypotheses → P(aj|b). All we need to know, to work this out, is the set of a priori probabilities P(aj), and the implications P(b|a­j).

In the special case of j = 2 we have:



Can observe that P(aj|b) is just a weighted average of all the P(ai|b)’s. And the weights are the a priori probabilities: P(ai)’s. If all the a priori probabilities are equal, then P(ai|b) is just P(b|ai)/(ΣP(b|aj).

We can visually represent Baye’s theorem with the diagram. The quadrilaterals are all the aj that together form the universal set (or at least together form a set of which b is a subset – so a1 and a2 don’t really matter), and b is the event that happened.



So we presume to know the probabilities P(aj) of all the aj’s. And we know all the conditional probabilities P(b|aj) of b [so basically we know all the intersection probabilities P(baj)]. And with that we can get any of the reverse conditional probabilities P(aj|b). Geometrically, this means we know all the areas of the events aj. And we know what fraction (b∩aj)/aj each aj event contributes to the area of b [or more simply, we know the intersection areas b∩aj] . And with that we can get the fraction (b∩aj)/b of area b contributes to each event aj. So could say,



Note this is just the given intersection area divided by the sum of all the intersection areas. Anyway, note we do *not* have:



‘cause if a happens regardless, then we’d have P(a) = 2. But since,



we *could* say,



and then, if a will happen regardless, we’d have:



as we should. We can have more than just one ‘b’. Consider following diagram,

Diagram

Description automatically generated with low confidence

Note 1 would take up the rest the space besides b1 and 2 would take up the rest of the space besides b2. Obviously there would be overlap between 1 and 2 just as there is between b1 and b2. The probability of any given ai given say, b1, is like before,



And we can do more complicated things, like the probability of some ai given b1 and b2 (I’ll symbolize as b1b2). This would just be:



And all of these probabilities are presumably known. Geometrically, they require knowledge of areas of ak, and of fractional areas bi/ak, and of fractional areas (bi∩bj)/ak. I guess we could generalize to say, the probability of some event E(**b**) involving the b’s, is:



Now I’d like to look at a more general form of Baye’s theorem. When we use the formula subjectively, we’re supposed to know the overall a priori probabilities for the mutually exclusive events aj. We often don’t know what P(ai) is itself, at least subjectively. We only know these probabilities in light of our other presently held beliefs. So we can say that we know P(aj|c). And then we’d like to work out how our estimate of the probability of aj changes upon additional information that the even b occurred. First let’s establish this result:



Note it has the same form as P(ab) = P(b)P(a|b), just with the addition of the extra c behind the |. And now consider:



where in the last line we use that prior result we established. So we have:



And we’ll observe that this looks the same as Baye’s rule, just with an extra *c* behind all the *a*’s. Diagram would look something like this:

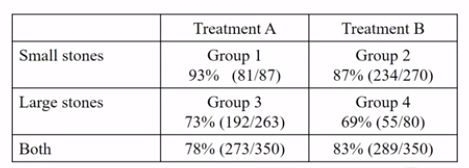
Chart, pie chart

Description automatically generated

P(ai|bc) is the fraction of (b∩c)’s area taken up by ai. P(aj|c) is the fraction of c’s area taken up by aj. P(b|ajc) is the fraction of (aj∩c)’s area taken up by b.

**Simpson’s Paradox**

So say a malady like kidney stones comes in two varieties – small and large. And you test two treatments A and B. Looks like you’re testing on different groups, however. Paradox is that treatment A, say, can be more successful in treating both small and large stones, individually, and yet have a lower overall success rate than treatment B.



**Example**

Consider two mutually exclusive hypotheses, *a* and . And some consequence, b. Baye’s rules says.



Would just like to note a few things.



So however much the probability of *a* goes up, compensates so that P(a|b) + P(|b) = 1. Suppose P(b|a) = P(b|). Then,



And likewise we’d find P(|b) = P(). In other words, the probability doesn’t change. That makes sense.

**Example**

Say your girlfriend said something odd to you. It could’ve been because she was upset with you or not. You’re trying to decide which is more probable. In general, the likelihood that she’s grumpy is about 15%. The probability she would’ve said that to you if she were grumpy is about 100%, and the probability she would’ve said it if she weren’t is about 25%. Well,



So most likely, she’s not mad. But it’s close. Better take her out to dinner anyway.

**Example**

Say the overall probability of breast cancer is 10%. And there is a test which will accurately confirm you have it 97% of the time, but also give false positive 5% of the time. It says you have it. What are the chances? Well diagram looks like this:

Shape, square

Description automatically generated

Our event space is spanned by C = having cancer, and = not having cancer. Our event P = testing positive. And we have:



Filling stuff in,



Wow that’s a lot less than I’d have thought. But this is the amount of P’s area taken up by C.

**Example**

Here’s one where we have a sort of composite Bayes rule. A sudden outbreak of a novel respiratory disease caused the medical system of a country to rush to test as many people in the population as possible in order to evaluate the transmissibility of the disease. Because the disease is novel, several new testing regimes were proposed. The government agency overseeing the selection of tests chose two tests as permissible. The LR2 test only requires the collection of saliva, whereas the NR8 test requires the insertion of a large cotton swab deep into the sinuses to collect a sample. Because of the extreme discomfort of the NR8 test for their patients, doctors are twice as likely to conduct the LR2 test.

When the LR2 test is conducted, there are two possible outcomes: If the patient has the disease, the LR2 test has a positive result 75% of the time. If the patient does not have the disease, the LR2 test has a false positive result 25% of the time.

When the NR8 test is conducted, there are two possible outcomes: If the patient has the disease, the NR8 test has a positive result 100% of the time. If the patient does not have the disease, the NR8 test has a positive result 50% of the time.

In a town where exactly 20% of the population has contracted the novel respiratory disease, a townsperson selected uniformly at random from the population was sent to the doctor for their mandatory screening for the illness. The result of the test came back as positive. What is the probability the person has the disease?

Define the probabilities of being positive, and of testing positive whether you’re positive or negative.



and also the probability of taking one test vs. the other.



Then from Baye’s Rule, breaking down the possibilities, we have (PT means ‘testing positive’):



**Example**

Say we have two events. x can be boy or girl, and y can be tall or short. So this effectively P(x,y).

|  |  |  |
| --- | --- | --- |
|  | y = short | y = tall |
| x = boy | 0.2 | 0.4 |
| x = girl | 0.3 | 0.1 |

With this, we can get any other probability of interest. So let’s calculate stuff. The individual probability distributions are…note trying to do multiple calculations at once, so x and y are column matrices x0 = boy, x1 = girl, and y0 = short, y1 = tall :



Note, importantly, that ΣxP(x) = 1, and ΣyP(y) = 1. Conditional probability distributions are as follows. Here I’m calculating it strictly from the formula.



But can also calculate them, shorthand, like this:



Evidently, these events are not independent variables since P(x|y) isn’t independent of y, and neither does it equal P(x), and neither does P(y|x) equal P(y).

**Example**

Let’s consider a symmetric distribution….

|  |  |  |
| --- | --- | --- |
|  | y = short | y = tall |
| x = boy | 0.1 | 0.2 |
| x = girl | 0.2 | 0.5 |

With this, we can get any other probability of interest. So let’s calculate stuff. The individual probability distributions are…note trying to do multiple calculations at once, so x and y are column matrices x0 = boy, x1 = girl, and y0 = short, y1 = tall :



Note, importantly, that ΣxP(x) = 1, and ΣyP(y) = 1. Conditional probability distributions are as follows.



But can also calculate them, shorthand, like this:



So a symmetric table gives identical marginal distributions, P(x), P(y), which makes sense. Are they independent? It would seem not necessarily on general grounds as symmetric distributions don’t imply independence. And they are not in fact, since P(x|y) is not independent of y.

**Example**

What about an independent distribution? Say P(A) = (0.2, 0.8) and P(B) = (0.4, 0.6). Then our table would be as below….

|  |  |  |
| --- | --- | --- |
|  | y = short | y = tall |
| x = boy | 0.08 | 0.12 |
| x = girl | 0.32 | 0.48 |

It isn’t obvious from just looking at the table that the variables are independent. Now let’s look at the conditional probabilities. Again I’ll do it the formal way for the first set,



and the not-so-formal way for the second:



And we see from the conditional probabilities confirmation of their independence since e.g. P(x|y) is independent of y, and moreover, is equal to P(x) itself.

**Example: Induction**

Inductively, we know that if a hypothesis’ prediction is confirmed, then its probability of being true increases. This follows from Baye’s rule. Let H and be our two hypotheses, analogous to the a’s. And let B be the prediction. And say that P(B|H) = 1, while P(B|) = 0. Then,



And this indicates that the probability of H, given its confirmed prediction B, is greater than it was before.

Can also just say, without splitting hypotheses into H and , that:



**Example: Contrapositive Law**

Have to do this later sometime, but can generalize the contrapositive deduction law (A→B) → (-B→-A) to probabilities.

Suppose P(|a) = 1, i.e., P(b|a) = 0. In other words a → -b. But then if we have b, then we must have -a, by the contrapositive law. And this checks out, as:



which is the same as P(|b) = 1. Or how about this:



Does it follow that P(A) = 0? Let’s see,

