Conditional probability and Bayes' theorem (https://eli.thegreenplace.net/2018/conditional-probability-and-bayes-theorem/)

March 13, 2018 at 05:32 **Tags** Math (https://eli.thegreenplace.net/tag/math)

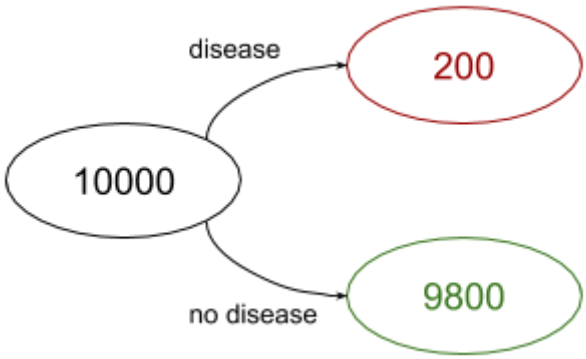
One morning, while seeing a mention of a disease on Hacker News, Bob decides on a whim to get tested for it; there are no other symptoms, he's just curious. He convinces his doctor to order a blood test, which is known to be 90% accurate. For 9 out of 10 sick people it will detect the disease (but for 1 out of 10 it won't); similarly, for 9 out of 10 healthy people it will report no disease (but for 1 out of 10 it will).

Unfortunatly for Bob, his test is positive; what's the probability that Bob actually has the disease?

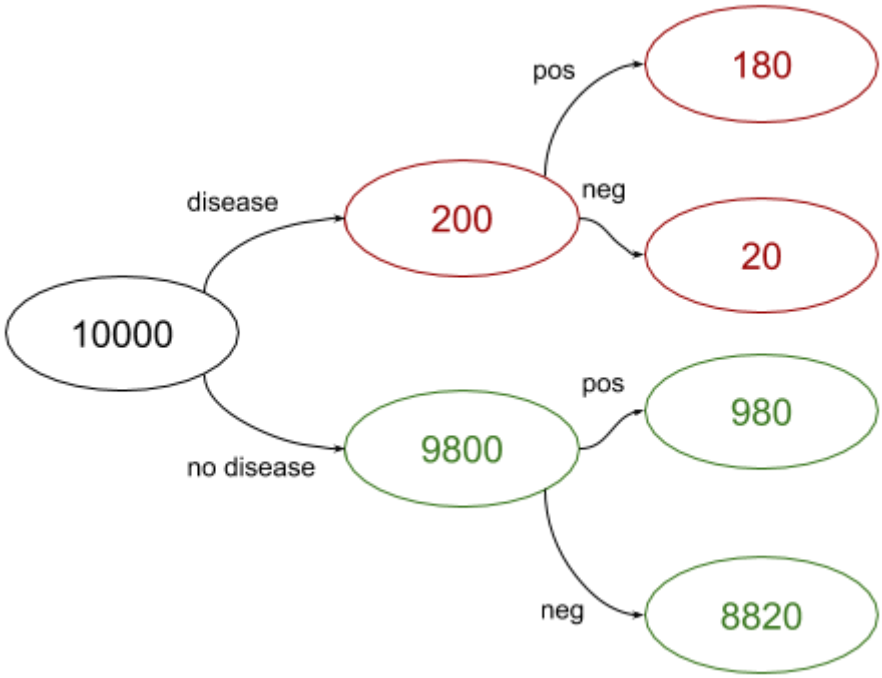
You might be tempted to say 90%, but this is wrong. One of the most common fallacies made in probability and statistics is mixing up conditional probabilities. Given event D - "Bob has disease" and event T - "test was positive", we want to know what is $P(D|T)$ - the conditional probability of D given T. But the test result is actually giving us $P(T|D)$ - which is distinct from $P(D|T)$.

In fact, the problem doesn't provide enough details to answer the question. An important detail that's missing is the *prevalence* of the disease in population; that is, the value of $P(D)$ without being conditioned on anything. Let's say that it's a moderately common disease with 2% prevalence.

To solve this without any clever probability formulae, we can resort to the basic technique of counting by cases. Let's assume there is a sample of 10,000 people [1]; test aside, how many of them have the disease? 2%, so 200.



Of the people who have the disease, 90% will test positive and 10% will test negative. Similarly, of the people with no disease, 90% will test negative and 10% will test positive. Graphically:



Now we just have to count. There are 980 + 180 = 1160 people who tested positive in the sample population. Of these people, 180 have the disease. In other words, given that Bob is in the "tested positive" population, his chance of having the disease is 180/1160 = 15.5%. This is *far* lower than the 90% test accuracy; conditional probability often produces surprising results. To motivate this, consider that the number of *true positives* (people with the disease that tested positive) is 180, while the number of *false positives* (people w/o the disease that tested positive) is 980. So the chance of being in the second group is larger.

Conditional probability

As the examples shown above demonstrate, conditional probabilities involve questions like "what's the chance of A happening, given that B happened", and they are far from being intuitive. Luckily, the mathematical theory of probability gives us the precise and rigorous tools necessary to reason about such problems with relative elegance.

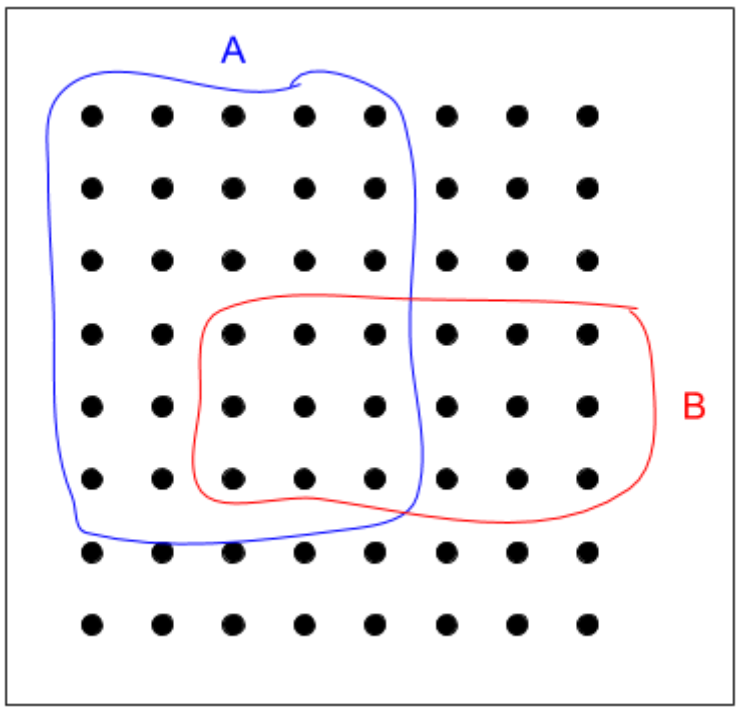
The conditional probability $P(A|B)$ means "what is the probability of event A given that we know event B occurred". Its mathematical definition is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Notes:

- Obviously, this is only defined when $P(B) > 0$.
- Here $P(A \cap B)$ is the probability that both A and B occurred.

The first time you look at it, the definition of conditional probability looks somewhat unintuitive. Why is the connection made this way? Here's a visualization that I found useful:



The dots in the black square represent the "universe", our whole sampling space (let's call it S, and then $P(S) = 1$). A and B are events. Here $P(A) = \frac{30}{64}$ and $P(B) = \frac{18}{64}$. But what is $P(A|B)$? Let's figure it out graphically. We know that the outcome is one of the dots encircled in red. What is the chance we got a dot also encircled in blue? It's the number of dots that are both red and blue, divided by the total number of dots in red. Probabilities are calculated as these counts normalized by the size of the whole sample space; all the numbers are divided by 64, so these denominators cancal out; we'll have:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{9}{18} = \frac{1}{2}$$

In words - the probability that A happened, given that B happened, is 1/2, which makes sense when you eyeball the diagram, and assuming events are uniformly distributed (that is, no dot is inherently more likely to be the outcome than any other dot).

Another explanation that always made sense to me was to multiply both sides of the definition of conditional probability by the denominator, to get:

$$P(A|B)P(B) = P(A \cap B)$$

In words: we know the chance that A happens given B; if we multiply this by the chance that B happens, we get the chance both A and B happened.

Finally, since $P(A \cap B) = P(B \cap A)$, we can freely exchange A and B in these definitions (they're arbitrary labels, after all), to get:

This is an important equation we'll use later on.

Independence of events

By definition, two events A and B are *independent* if:

$$P(A \cap B) = P(A)P(B)$$

Using conditional probability, we can provide a slightly different definition. Since:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

And $P(A \cap B) = P(A)P(B)$:

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

As long as $P(B) > 0$, for independent A and B we have $P(A|B) = P(A)$; in words - B doesn't affect the probability of A in any way. Similarly we can show that for $P(A) > 0$ we have $P(B|A) = P(B)$.

Independence also extends to the complements of events. Recall that $P(B^C)$ is the probability that B *did not* occur, or $1 - P(B)$; since conditional probabilities obey the usual probability axioms, we have: $P(B^C|A) = 1 - P(B|A)$. Then, if A and B are independent:

$$P(B^C|A) = 1 - P(B) = P(B^C)$$

Therefore, B^C is independent of A. Similarly the complement of A is independent of B, and the two complements are independent of each other.

Bayes' theorem

Starting with equation (1) from above:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$$

And taking the right-hand-side equality and dividing it by $P(B)$ (which is positive, per definition), we get Bayes's theorem:

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

This is an extremely useful result, because it links $P(B|A)$ with $P(A|B)$. Recall the disease test example, where we're looking for $P(D|T)$. We can use Bayes theorem:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)}$$

We know $P(T|D)$ and $P(D)$, but what is $P(T)$? You may be tempted to say it's 1 because "well, *we know the test is positive*" but that would be a mistake. To understand why, we have to dig a bit deeper into the meanings of conditional vs. unconditional probabilities.

Prior and posterior probabilities

Fundamentally, conditional probability helps us address the following question:

> How do we update our beliefs in light of new data?

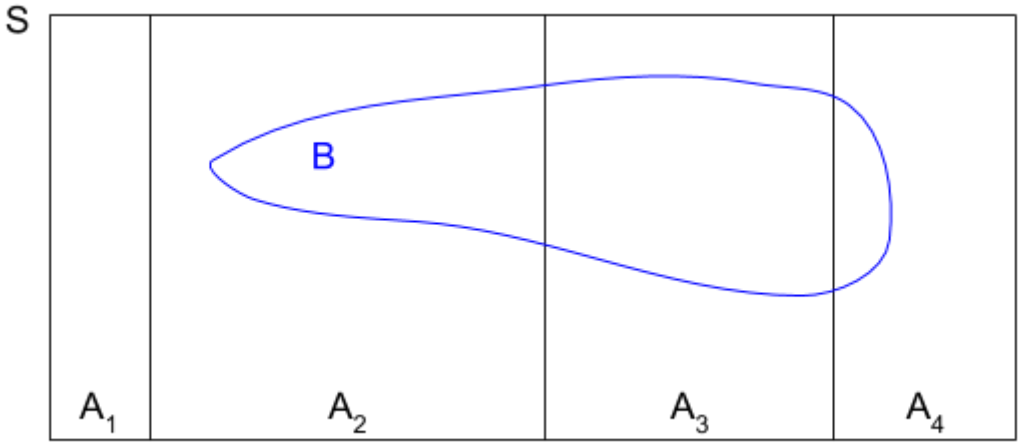
Prior probability is our beliefs (probabilities assigned to events) before we see the new data. *Posterior* probability is our beliefs after we see the new data. In the Bayes equation, prior probabilities are simply the un-conditioned ones, while posterior probabilities are conditional. This leads to a key distinction:

- $P(T|D)$: posterior probability of the test being positive when we have new data about the person - they have the disease.
- $P(T)$: prior probability of the test being positive before we know anything about the person.

This should make it clearer why we can't just assign $P(T) = 1$. Instead, recall the "counting by cases" exercise we did in the first example, where we produced a tree of all possibilities; let's formalize it.

Law of Total Probability

Suppose we have the sample space S and some event B. Sometimes it's easier to find the probability of B by first partitioning the space into disjoint pieces:



Then, because the probabilities of A_n are disjoint, we get:

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + P(B \cap A_3) + P(B \cap A_4)$$

Or, using equation (1):

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3) + P(B|A_4)P(A_4)$$

Bayesian solution to the disease test example

Now we have everything we need to provide a Bayesian solution to the disease test example. Recall that we already know:

- $P(T|D) = 0.9$: test accuracy
- $P(D) = 0.02$: disease prevalance in the population

Now we want to compute $P(T)$. We'll use the law of total probability, with the space partitioning of "has disease" / "does not have disease":

$$P(T) = P(T|D)P(D) + P(T|D^C)P(D^C) = 0.9 * 0.02 + 0.1 * 0.98 = 0.116$$

Finally, plugging everything into Bayes theorem:

$$P(D|T) = \frac{P(T|D)P(D)}{P(T)}$$

$$\begin{aligned} P(D|T) &= \frac{P(T|D)P(D)}{P(T)} \\ &= \frac{P(T|D)P(D)}{0.116} \\ &= \frac{0.9 * 0.02}{0.116} = 0.155 \end{aligned}$$

Which is the same result we got while working through possibilities in the example.

Conditioning on multiple events

We've just computed $P(D|T)$ - the conditional probability of event D (patient has disease) on event T (patient tested positive). An important extension of this technique is being able to reason about multiple tests, and how they affect the conditional probability. We'll want to compute $P(D|T_1 \cap T_2)$ where T_1 and T_2 are two events for different tests.

Let's assume T_1 is our original test. T_2 is a slightly different test that's only 80% accurate. Importantly, the tests are *independent* (they test completely different things) [2].

We'll start with a naive approach that seems reasonable. For T_1, we already know that $P(D|T_1) = 0.155$. For T_2, it's similarly simple to compute:

$$P(D|T_2) = \frac{P(T_2|D)P(D)}{P(T_2)}$$

The disease prevalence is still 2%, and using the law of total probability we get:

$$P(T_2) = P(T_2|D)P(D) + P(T_2|D^C)P(D^C) = 0.8 * 0.02 + 0.2 * 0.98 = 0.212$$

Therefore:

$$P(D|T_2) = \frac{P(T_2|D)P(D)}{P(T_2)} = \frac{0.8 * 0.02}{0.212} = 0.075$$

In other words, if a person tests positive with the second test, the chance of being sick is only 7.5%. But what if they tested positive for both tests?

Well, since the tests are independent we can do the usual probability trick of combining the complements. We'll compute the probability the person is *not* sick given positive tests, and then compute the complement of that. $P(D^C|T_1) = 1 - 0.155 = 0.845$, and $P(D^C|T_2) = 1 - 0.075 = 0.925$. Therefore:

$$P(D^C|T_1 \cap T_2) = P(D^C|T_1)P(D^C|T_2) = 0.845 * 0.925 = 0.782$$

And complementing again, we get $P(D|T_1 \cap T_2) = 1 - 0.782 = 0.218$. The chance of being sick, having tested positive both times is 21.8%.

Unfortunately, this computation is wrong, *very* wrong. Can you spot why before reading on?

We've committed a fairly common blunder in conditional probabilities. Given the independence of $P(T_1|D)$ and $P(T_2|D)$, we've assumed the independence of $P(D|T_1)$ and $P(D|T_2)$, but this is wrong! It's even easy to see why, given our concrete example. Both of them have $P(D)$ - the disease prevalence - in the numerator. Changing the prevalence will change both $P(D|T_1)$ and $P(D|T_2)$ in exactly the same proportion; say, increasing the prevalence 2x will increase both probabilities 2x. They're pretty strongly dependent!

The right way of finding $P(D|T_1 \cap T_2)$ is working from first principles. $T_1 \cap T_2$ is just another event, so treating it as such and using Bayes theorem we get:

$$P(D|T_1 \cap T_2) = \frac{P(T_1 \cap T_2|D)P(D)}{P(T_1 \cap T_2)}$$

Here $P(D)$ is still 0.02; $P(T_1 \cap T_2|D) = 0.9 * 0.8 = 0.72$. To compute the denominator we'll use the law of total probability again:

$$P(T_1 \cap T_2) = P(T_1 \cap T_2|D)P(D) + P(T_1 \cap T_2|D^C)P(D^C) = 0.72 * 0.02 + 0.1 * 0.2 * 0.98 = 0.034$$

Combining them all together we'll get $P(D|T_1 \cap T_2) = 0.42$; the chance of being sick, given two positive tests, is 42%, which is twice higher than our erroneous estimate [3].

Bayes theorem with conditioning

Since conditional probabilities satistfy all probability axioms, many theorems remain true when adding a condition. Here's Bayes theorem with extra conditioning on event C:

$$P(A|B \cap C) = \frac{P(B|A \cap C)P(A|C)}{P(B|C)}$$

In other words, the connection between $P(A|B)$ and $P(B|A)$ is true even when everything is conditioned on some event C. To prove it, we can take both sides and expand the definitions of conditional probability until we reach something trivially true:

$$\begin{aligned}
P(A|B \cap C) &= \frac{P(B|A \cap C)P(A|C)}{P(B|C)} \\
\frac{P(A \cap B \cap C)}{P(B \cap C)} &= \frac{P(A \cap B \cap C)P(A|C)}{P(A \cap C)P(B|C)} \\
\frac{P(A \cap B \cap C)}{P(B \cap C)} &= \frac{P(A \cap B \cap C)P(A \cap C)}{P(A \cap C)P(B|C)P(C)}
\end{aligned}$$

Assuming that $P(A \cap C) > 0$, it cancels out (similarly for $P(C) > 0$ in a later step):

$$\begin{aligned}
\frac{P(A \cap B \cap C)}{P(B \cap C)} &= \frac{P(A \cap B \cap C)}{P(B|C)P(C)} \\
\frac{P(A \cap B \cap C)}{P(B \cap C)} &= \frac{P(A \cap B \cap C)P(C)}{P(B \cap C)P(C)} \\
\frac{P(A \cap B \cap C)}{P(B \cap C)} &= \frac{P(A \cap B \cap C)}{P(B \cap C)}
\end{aligned}$$

Q.E.D.

Using this new result, we can compute our two-test disease exercise in another way. Let's say that T_1 happens first, and we've already computed $P(D|T_1)$. We can now treat this as the new *prior* data, and find $P(D|T_1 \cap T_2)$ based on the new evidence that T_2 happened. We'll use the conditioned Bayes formulation with T_1 being C.

$$P(D|T_2 \cap T_1) = \frac{P(T_2|D \cap T_1)P(D|T_1)}{P(T_2|T_1)}$$

We already know that $P(D|T_1)$ is 0.155; What about $P(T_2|D \cap T_1)$? Since the tests are independent, this is actually equivalent to $P(T_2|D)$, which is 0.8. The denominator requires a bit more careful computation:

$$P(T_1|T_2) = \frac{P(T_1 \cap T_2)}{P(T_1)}$$

We've already found $P(T_1) = 0.116$ previously, using the law of total probability. Using the same law:

$$P(T_1 \cap T_2) = P(T_1 \cap T_2|D)P(D) + P(T_2 \cap T_2|D^C)P(D^C) = 0.9 * 0.9 * 0.02 + 0.1 * 0.2 * 0.98 = 0.034$$

Therefore, $P(T_1|T_2) = \frac{0.034}{0.116} = 0.293$ and we now have all the ingredients:

$$P(D|T_2 \cap T_1) = \frac{0.8 * 0.155}{0.293} = 0.42$$

We've reached the same result using two different approaches, which is reassuring. Computing with both tests taken together is a bit quicker, but taking one test at a time is also useful because it lets us *update our beliefs* over time, given new data.

Computing conditional probabilities w.r.t. multiple parameters is very useful in machine learning - this would be a good topic for a separate article.

---

[1] This actual number of people is arbitrary, and it could be anything else; in formulae it cancels out anyway. I picked 10,000 because it's a nice number ending with a bunch of zeros and won't produce fractional people for this particular example.

[2] You may be suspicious of this assumption - how can two tests for the same disease be independent? Being suspicious about probability independence assumptions is a good idea in general, but here the assumption is reasonable.

Note that we assume independence given D; in other words, that $P(T_1|D)$ and $P(T_2|D)$ are independent. We know the person is sick, and we know that T_1 turned positive - does this affect T_2? Depends on the test; some tests definitely test related things, but some may test unrelated things (say the first looks for a particular by-product of sick cells while the second looks for a gene that is known to be correlated with disease prevalence). It's possible to find plausible connections between almost anything though, so all independence assumptions are "best-effort".

[3] My intuition for understanding why it's higher is that there's a tug of war between the test accuracy and the prevalence (the lower the prevalence, the higher the test accuracy has to be to produce reasonable predivtive value). But when we recompute with two tests, we still use prevalence just once in the formula, so the two tests combine forces against it.

For comments, please send me an email (mailto:eliben@gmail.com), or reach out on on Twitter (https://twitter.com/elibendersky).

Back to top