# The Bayesian approach for calculating the probability of a hypothesis given: the results of some test, statistical properties of the test, and the prior probability of the hypothesis (before any information from the test result)

In what follows I'm going to present Bayes' formula (Bayes' theorem) using the classic example of calculating the probability that a person has a particular disease given a positive test for it, along with information on the "accuracy" of the test and the *prevalence* of the disease (terms will be defined in the next page).

I'll first show how one can calculate the probability of disease given a positive test in a "hands on" way, and then show how one can do this systematically using Bayes' theorem.

Note other than the particular exposition (for better or worse), there is nothing "new" here (and this material can be found in most statistics books and by Google search). I've tried to "collect" and explain the commonly used formulas and terms used for the basic analysis of a diagnostic test. In particular, this was written to address queries in the Coursera Johns Hopkins University Data Science Specialization **Statistical Inference** Course, whose textbook is: [1] **Statistical Inference for Data Science**, Brian Caffo, Leanpub,
last updated on 2016-05-23

Alan E. Berger 28 May 2020

# Definitions

The **prevalence** of a disease is the fraction of the population being tested that actually has the disease. The prevalence can be viewed as *prior* information modifying the likelihood of disease derived from the information that the test was positive, and from the *sensitivity* and *specificity* of the test, defined below.

A **test** for a disease, for this discussion, is a measurement taken, for example the amount $\mathcal{A}$ of some biomolecule in the blood, and some **decision procedure** on whether the test indicates presence of the disease (e.g., if $\mathcal{A}$ is greater than some cutoff level $\mathcal{L}$, then report a **positive result** (disease is present)).
Procedures, such as *ROC (*receiver operating characteristic*) analysis,* for how to make a "good choice" for $\mathcal{L}$, is a related topic but outside our focus here.

The **sensitivity** of a test is the fraction of time the test will return a positive result (disease present) when the test subject has the disease.

The **specificity** of a test is the fraction of time the test will return a negative result (disease not present) when the test subject does not have the disease.

The **false alarm rate** (**FAR**) (or **false positive rate (FPR)**) of a test is the fraction of time the test will return a positive result (disease present) when the test subject does **not** have the disease (FAR = 1 – specificity).

The **false negative rate** of a test is the fraction of time the test will return a negative result (disease absent) when the test subject does have the disease
(false negative rate = 1 – sensitivity).

So let's see an example:
Suppose the test has a sensitivity of 0.95 (often given as percent, so 95%); and a specificity of 98% (these are the numbers quoted on a web site for a rapid strep test [1]). A value of 25% for the prevalence (in the population of children with symptoms sufficient to pursue testing) is in the range given in [2]. So these are "ball-park" numbers for an actual test (even if not current) so reasonable enough for our first example.

[1] https://www.physician360.co/news/what-is-a-rapid-strep-test-and-are-they-reliable

[2] http://southwestchildrenscenter.com/pdf/information-and-forms/article-library/articles-by-physicians/marshall-j-benbow-md/Streptococcal%20Pharyngitis.pdf

# A “direct” calculation without the Bayes' formula

Given these values on the test behavior and prevalence, we can do a calculation of **the probability of the hypothesis H of having the disease** given the **data D of a positive test result**, **P(H|D)**, also called the **positive predictive value (PPV)**, as follows. Suppose we had N = 4000 subjects being given the test (the value of N “scales out” of the calculation, but here is a convenient choice to make intermediate numbers come out as integers, which while not necessary to get P(H|D), looks more natural).

Then (prevalence = 25%) we would expect (**on average**) 0.25 * 4000 = 1000 of the 4000 to actually have the disease. From the given test sensitivity value of 0.95, we would expect 950 true positive tests from the 1000 subjects with the disease. With a specificity of 0.98, the false alarm rate is 0.02, so (**on average**) we would expect 0.02 * 3000 = 60 false positive test results. So (**on average**) there would be a total of 1010 positive test results of which 950 would have the disease, so

p(H|D) probability of having the disease given the data of a positive test result = 950 / 1010 = 0.940594

and the **false discovery rate** (the fraction of positive results that are false positives) is 60 / 1010 = 0.0594059 (I’ve kept this absurd number of digits to verify against a later calculation).

These statistics don't look too bad (depending on the costs of false positives and false negatives).

To see how much the prevalence effects these results, now look at the situation where we have the same test characteristics (sensitivity = 0.95, specificity = 0.98), but now the disease is much rarer, so say the prevalence is only 1%

Here it is convenient to consider N = 40,000 as the number of subjects being tested. Then (on average) we would have: 400 subjects with the disease so 39,600 without the disease. Out of the 400, we would expect to get 380 (true) positive test results; but out of the 39,600 subjects without the disease the expected (average) number of false positive test results would be 0.02 * 39600 = 792 and so the expected total number of all the positive test results would be 380 + 792 = 1172

Then P(H|D) = 380 / 1172 = 0.324

and the false discovery rate is 792 / 1172 = 0.676, so the **prevalence**, which in Bayesian terminology is the ***prior probability* of the disease**, can have a big influence on the positive predictive value of a test.

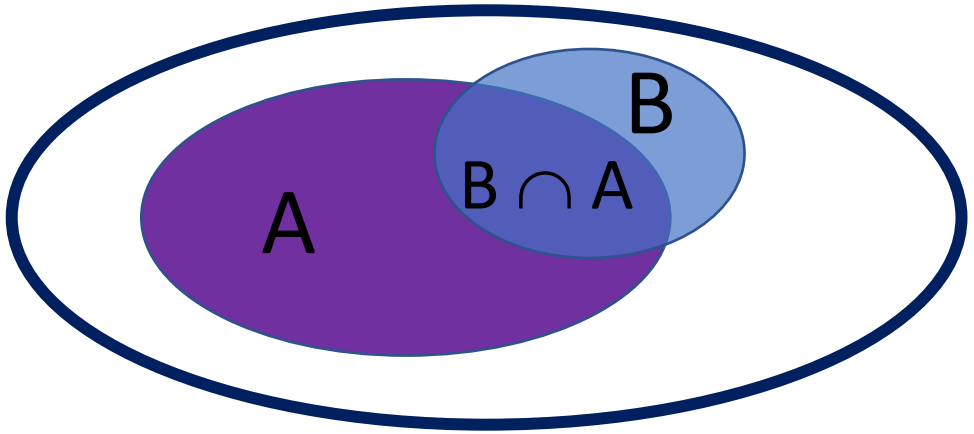
Now let's formulate this in terms of Bayes' formula.

For Bayes' formula, we need the formal definition of **conditional probability**, **P(B|A)**, the **probability of event B *given that event A is true*** (or has taken place). This is read as "the probability of B given A or "P of B given A" and is **defined by**

$$P(B|A) \equiv P(B \cap A) / P(A) \quad \text{(assuming } P(A) > 0)$$

Here $P(B \cap A)$ can be viewed as P(both B and A occur).

In terms of a Venn diagram:
P(B|A) is the probability of
$B \cap A$ relative to considering as
possible outcomes only those
within A, so in that case "we need to"
divide $P(B \cap A)$ by P(A) to have the total
(maximal) probability be 1: P(B|A) should be 1 when B contains A. As illustrated in the next page, this formal definition well coincides with its "verbal description".



As an example to see that this definition makes sense, consider the outcomes of rolling a die (a cube with 6 faces labeled 1,2,3,4,5,6). These six numbers are the possible outcomes of the experiment of tossing it (we are not considering the unlikely case that it comes to rest balanced on an edge (or vertex)), and we assume that it is "fair", i.e., each outcome is equally likely (with a probability P = 1/6).

Then, for example, the probability that the outcome is either 1 or 3 or 6 (we'll call any of these a "successful" outcome), **given that** the outcome was an odd number (i.e., either 1, 3, or 5), is P = 2/3. This is calculated via

the number of *possible* **successful** outcomes / the **total** number of *possible* outcomes,

where **possible** here means **possible given that the outcome was on odd number**; and so indeed this is 2 (since 6 is not a possibility given that the outcome was an odd number) divided by 3 (since the 3 odd numbers 1,3,5 have been given to be the possible outcomes).

Going through the formal definition with B the event the die came up 1 or 3 or 6, and A the event that the die came up with an odd number (1 or 3 or 5), we have
$B \cap A$ is the set of outcomes consisting of 1 or 3 and so by the formal definition,

$P(B|A) \equiv P(B \cap A) / P(A)$ which here is $(2/6) / (3/6) = 2/3$ as it "should be"

Now we are ready to get the first form of Bayes' formula. We can rewrite the conditional probability definition as

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

Interchanging A and B in the above:

$P(B \cap A)$ (which is the same as $P(A \cap B)$) $= P(A|B) * P(B)$ and so

**$P(B|A) * P(A) = P(A|B) * P(B)$**

From the previous slide, **P(B|A) * P(A) = P(A|B) * P(B)**

OK, so what does this have to do with probability of disease given a positive test?
well, we're getting there …

The above can to better advantage for our purposes be rearranged as

**P(B|A) = P(A|B) * P(B) / P(A)**

And if we now replace B by H (e.g., the **hypothesis** of having the disease) and A by D (e.g., the **data** of a positive test)
we have:

**P(H|D) = P (D|H) * P(H) / P(D)** **Bayes' formula version 1**

This is to be "read as": the probability of the hypothesis given the data (**the posterior probability,** also called **the positive predictive value PPV**) equals the probability of the data given the hypothesis <this is called **the likelihood**> times the **prior probability** of the hypothesis divided by **the probability of the observed data**.

We're now 1 step away from being able to "turn the crank" with this formula and routinely compute the probability of disease given a positive test and the properties of the test and the prevalence.
Copying in from the last page we have

**P(H|D) = P (D|H) * P(H) / P(D)** **Bayes' formula version 1**

We need to get a more computationally convenient form for P(D). Now for each subject, for this example, they either have the disease (H) or they don't (let's just denote that by they are **well (W)**). Hence P(H or W) = 1 and H and W are mutually exclusive (in math notation $P(H \cup W) = 1$ and $H \cap W = \varnothing$ ).
Hence **P(D)** = **P(D ∩H) + P(D ∩W)** since if d is an element of D then it is in one (but not both) of (D ∩H) and (D ∩W)

And then using the conditional probability formula
$P(D \cap A) = P(D|A) * P(A)$ twice, once with A being H and once with A being W, we get the commonly seen version of Bayes' formula (also called Bayes' theorem) for the case of 2 mutually exclusive possible "results", here H and W:

**P(H|D) = P (D|H) * P(H) / [P(D|H)*P(H) + P(D|W)*P(W)]** **Bayes' theorem**

Now let’s use Bayes' theorem (for 2 mutually exclusive possible results H and W):

**P(H|D) = P (D|H) * P(H) / [P(D|H)*P(H) + P(D|W)*P(W)]** **Bayes**

to recalculate in a systematic way P(H|D) for the first case of prevalence = 25%:
so D means the data of having gotten a positive test result; the prevalence (prior probability of disease) P(H) = 0.25; the sensitivity P(D|H) = 0.95; and the false alarm rate P(D|W) = 0.02 (from FAR = 1 – specificity). Since H and W are mutually exclusive and are the only possibilities, P(W) = 1 – P(H) = 0.75 Hence, putting these values into Bayes' formula gives

the probability of disease given the data D a positive test (the **PPV**)
P(H|D) = 0.95 * 0.25 / [0.95*0.25 + 0.02*0.75] = 0.2375 / [0.2375 + 0.015]
= 0.2375 / 0.2525 = 0.940594

Which, as advertised, agrees with the earlier calculation.

The next slide summaries this using the terms often encountered in the arena of medical tests.

Now let's restate Bayes' theorem

**$P(H|D) = P(D|H) * P(H) / [P(D|H)*P(H) + P(D|W)*P(W)]$** **Bayes** **(note the denominator is P(D))**

directly in terms of testing for a condition:

**H = hypothesis = the subject has the condition (or disease) being tested for**

**W = is well (does not have the condition),**

**D = the data (here meaning the test result is "positive" indicating has the condition)**

Then:

**P(H|D) is the probability of having the condition given a positive test result, also called the positive predictive value (PPV)**

**P(D|H) is the probability of a positive test result given that the subject has the condition, which is the sensitivity of the test**

**P(H) is the probability the subject has the condition, as known or estimated before any test was done, which is the prevalence** or **prior probability** of disease

**P(D|W) is the probability of a positive test result given that the subject is well, which is the false positive rate (FPR) (also called the false alarm rate) and equals 1 – the specificity (specificity = P(negative test result | W) )**

**P(W) is the probability the subject does not have the condition, as known or estimated before any test was done, which equals 1 – the prevalence**

**Also,**

**P(W|D) is the probability the subject is well when the test result was positive, also called the false discovery rate (FDR)**

**Note** $P(H|D) + P(W|D) = 1$ (since here either H or W is true, and also not both at the same time)

**Then Bayes' theorem for testing for a condition becomes: (note << means "much less than"; ≈ means "is well approximate by")**

**positive predictive value = sensitivity * prevalence / (sensitivity*prevalence + false positive rate*(1-prevalence))**

**When the prevalence is << FPR, and sensitivity ≈ 1, then PPV ≈ (prevalence / FPR) = prevalence / (1 – specificity)**

# Some observations on the **Bayes factor** (defined below)

Repeating, we have

$$P(H|D) = P(D|H) * P(H) / [P(D|H)*P(H) + P(D|W)*P(W)]$$ **Bayes**

and hence also, the probability for the subject being well (W) given the data D of a positive test (note the denominator is just the probability of the data D, P(D):

$$P(W|D) = P(D|W) * P(W) / [P(D|H)*P(H) + P(D|W)*P(W)]$$ **Bayes for W**

Now here the subject is either well or sick (no Schrödinger's cat scenarios allowed). If we sum these two Bayes equations we get $P(H|D) + P(W|D) = 1$ as expected. If we take the ratio of each side of these two equations we get (the denominator P(D) cancels out):

$$P(H|D) / P(W|D) = \{P(H) / P(W)\} * \{P(D|H) / P(D|W)\}$$

This says the ratio of the probability the subject is sick (given a positive test) to the probability the subject is well (given a positive test) equals
the ratio of the prior probability for sick to the prior probability for well (before we had the test result) times the term in red which is called the **Bayes factor**.

# The **Bayes factor P (D|H) / P (D|W)**

The **Bayes factor** is the ratio of the likelihood of the data D given the subject is sick (H) to the likelihood of the data D given the subject is well (W) <this is called **the diagnostic likelihood ratio of a positive test** in the Statistical Inference Course>

It is the amount by which the ratio of prior probabilities P(H) / P(W) (**the prior odds of sick to well**) is multiplied to get the ratio of posterior probabilities P(H|D) / P(W|D) of sick to well (**the posterior odds of sick to well**) (given the data D of a positive test). Hence the equation

**P(H|D) / P(W|D) = {P(H) / P(W)} * {P (D|H) / P (D|W)}**

can be phrased as

The posterior odds of sick to well = the prior odds of sick to well * the Bayes factor

(this equation also holds if D denotes a negative test result, in which case the Bayes factor is also called **the diagnostic likelihood ratio of a negative test**).

# Calculation of P(H|D) using the Bayes factor

We can calculate the **PPV** (positive predictive value) **P(H|D)** without directly calculating P(D) in the case that there are only two mutually exclusive possibilities H and W. We have

$$P(H|D) / P(W|D) = \{P(H) / P(W)\} * \{P(D|H) / P(D|W)\} \qquad \textbf{equation (1)}$$

and since we are assuming the only possibilities are H and W, we also have

$$P(H|D) + P(W|D) = 1 \quad \text{and so} \quad P(W|D) = 1 - PPV$$

Denote the right-hand side of equation (1) by $\mathcal{R}$ which is then

$$\mathcal{R} = \{\text{prevalence} / (1 - \text{prevalence})\} * \{\text{sensitivity} / (1 - \text{specificity})\}$$

Then from equation (1), $PPV = (1 - PPV) * \mathcal{R}$ so $PPV = \mathcal{R} - PPV * \mathcal{R}$ and hence

$$PPV = \mathcal{R} / (1 + \mathcal{R})$$

and $\mathcal{R}$ involves only the standard quantities sensitivity, specificity and prevalence

# A bonus example – the famous "Monte Hall" problem

The setting is a game show, there are 3 doors, behind 1 of the doors (chosen at random before the show started) is a substantial prize, behind the other two doors is nothing (or a booby prize). The contestant picks a door (without loss of generality, say door 1). The host of the game show knows where the prize is (but neither the host nor anyone else can change which door the prize is behind). If the prize is behind door 1 (the door the contestant picked), the host opens one of the two other doors (door 2 or door 3) with a 50% probability of opening either one, showing the prize is not there. If the prize is behind door 2 or door 3, the host opens door 3 or door 2, respectively (the one of the two doors the contestant did not pick that does not have the prize). The host then asks the contestant whether he or she wants to switch their pick to the other unopened door.

The common quick instinctive response is **it doesn't matter**: the prize is behind one of the two remaining unopened doors so its "50-50"
That is **wrong**!! It does not take advantage of the additional information the host provided. Additional information -- sounds just like Bayesian …

The “**right way**” to think about this is: if, instead of opening any doors, the host always told the contestant (regardless of whether or not the contestant had picked the door with the prize) that the contestant could keep whatever was behind the door they picked, **or** could keep the prize if it was **behind either of the 2 other unopened doors**. Well since the prize was put at random behind one of the 3 doors, the probability it was behind the door the contestant picked is 1/3, and the probability it was behind one of the other two doors is 2/3; so the contestant doubles their probability of getting the prize by switching. If the host had only let the contestant simply change their mind and pick another door, the contestant would be no better off. But, back to the original format (where the host opens a door without the prize), what the host has in effect done is tell the contestant that they can pick only one of the other two doors – BUT “I’ll tell you one that doesn’t have the prize” – so in the 2/3 of the cases where the prize is behind either door 2 or door 3 the contestant gets to pick the correct door.

That’s a whole lot of thinking to get to the correct result. Using Bayes’ Theorem is much easier.

# The Monte Hall problem solved by Bayes

For concreteness, say the contestant picked door 1 and the host opened door 3 (no prize). Let H be the hypothesis that the prize is behind door 1 ("door 1") and W be the hypothesis that the prize is behind door 2 ("door 2"). The prior probabilities (before the host opened any doors) are P(H) = P(W) = 1/3

Now after we are given the data D that the host opened door 3, we have P(D|H) = ½ since the rules are that when the prize is behind door 1 (picked by the contestant) the host opens door 2 or door 3 with probability 0.5 for each.

But P(D|W) = 1 since if the prize is behind door 2 (hypothesis W) the host had no choice but to open door 3. Also, P(prize behind door 3|D) is 0. Then equation (1)

**P(H|D) / P(W|D) = {P(H) / P(W)} * {P(D|H) / P(D|W)}** **equation (1)**

"says" that P("door 1" given host opened door 3) / P("door 2" given host opened door 3) =
{(1/3) / (1/3)} * {(1/2) / 1} = ½

Since P(H|D) + P(W|D) = 1, we then get P("door 2" given host opened door 3) = 2/3 while P("door 1" given host opened door 3) = 1/3, so the contestant should switch, and Bayes' theorem lets us determine this systematically without needing "clever thinking".