

Statistics – Tool or Weapon?

The application and misapplication of Bayes' theorem

There is an old saying: “Statistics don’t lie... but statisticians sometimes do.” This maxim is a corollary of the well known fact that ‘Garbage In produces Garbage Out’ (*GIGO*). It reaffirms the observation that, no matter how sophisticated any given computer program may be, if you feed it input data that is flawed, the output you will get will also be flawed. Despite this however, most people are usually impressed by complication: the more complicated the computerized algorithm, the more persuasive the results are for anyone who is gullible. The effectiveness of complicated statistics is a case in point.

Most of us have heard the terms *a priori* and *a posteriori* and know they refer respectively to things that *precede* or *follow* an event. But most people do not know how to calculate nor to interpret an ‘*a posteriori* probability’ even though they may have occasionally heard that term. Such quantities are often used in science and we ought to know how to interpret them and how to judge the appropriateness of using them. For this we can turn to Bayes’ theorem.

Bayes’ theorem (*BT*) is a statistical principle that is quite enlightening if properly applied. But as with many powerful tools, it can be dangerous if it is misused. *BT* was specifically developed to answer questions about the validity of tests. For example, consider the following test situations and then ask how much credence ought to be put in the results of these procedures:

1. Suppose you undergo a medical test for cancer. The test diagnoses you as having cancer. In this event, what is the probability that you indeed do have cancer?
2. A test is initiated to determine who is the best teacher in a school employing 30 teachers. This test has been previously shown to identify the best teacher from a similar sized group of teachers with 95% accuracy. After using this test to arrive at a winner, what is the probability we have correctly identified the best teacher on the faculty?

Bayes’ Theorem provides a quantitative handle on the accuracy (believability) of answers to questions such as 1, and 2, (above). It lets us know how much faith we should put in the results of any given test when it is applied in some specific situation. But as with all mathematical procedures (computerized or not), we must be ever alert to the *GIGO* principle.

We can describe any simple test (diagnostic procedure) by the diagram shown in figure 1 below. This diagram is generally applicable to any testing procedure but is labeled here specifically in terms of question 1, above. The two quantities shown at the left are the ‘inputs’ and are called the *a priori* probabilities – the probabilities of occurrence and non-occurrence in the general population of the quality we are testing for (before we perform the test):

$p(C)$ = The probability that any person picked at random from the general population has cancer. Obtaining an accurate estimate of this is crucial but is often difficult.

$p(\text{not } C)$ = The probability that a typical person does **not** have cancer = $1 - p(C)$ because a person either does or does not have cancer. (These two events are mutually exclusive.) And a probability of 1 implies certainty.

The phrase *a priori* is used to describe them because these quantities are known *prior* to applying the test.

The probabilities that describe the accuracy of the test itself are the following four ‘conditional probabilities’:

$p(D|C)$ = The probability that a person will be ‘correctly diagnosed’ as having cancer given that he or she really does already have cancer.

$p(\text{not } D|C)$ = The probability that a person will be diagnosed as being cancer free given that he or she does have cancer. This called a ‘false negative’ diagnosis.

$p(D|\text{not } C)$ = The probability of a cancer free person being diagnosed as having cancer. This result is called a ‘false positive’ diagnosis.

$p(\text{not } D|\text{not } C)$ = The probability of a cancer free person being ‘correctly diagnosed’ as being cancer free.

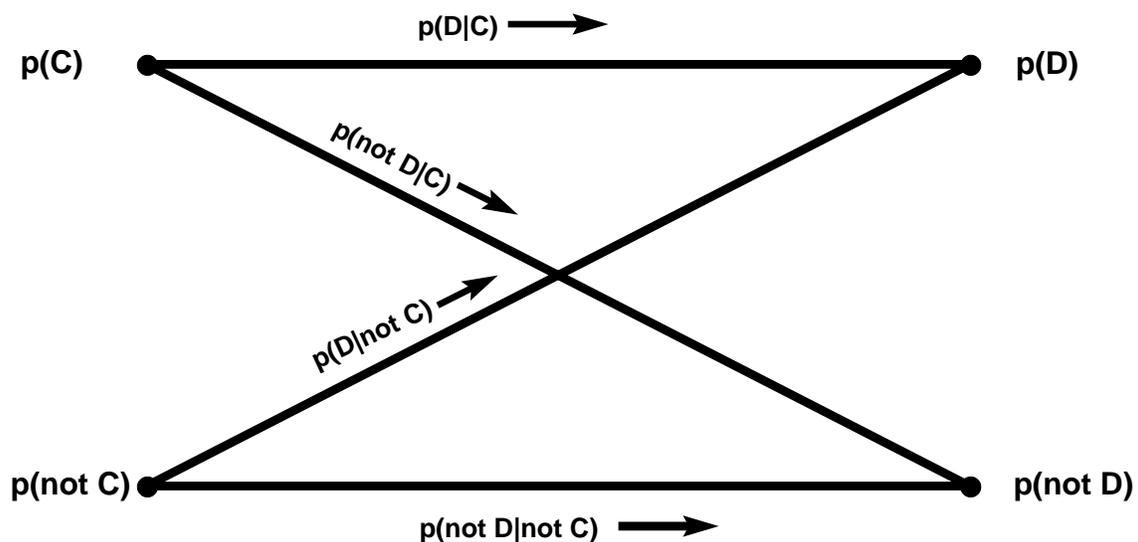


Figure 1. Graphical description of a diagnostic test. The quantities ‘C’ and ‘not C’ are mutually exclusive – there are no other possibilities.

Presumably, all four of the conditional probabilities listed above have been accurately determined by laboratory testing on known populations before this diagnostic test is actually given to anybody or applied in the field. They describe the outputs the test yields when applied to *known* test subjects.

Of course, what we (and especially the patients who will be subjected to this test) are interested in knowing is, in the case of the cancer test: What is the probability that I actually *do* have cancer given that the test says I do? Can I believe the diagnosis? This is the *a posteriori* (after the test) probability $p(C|D)$. It will not be equal to the conditional probability $p(D|C)$ unless the test is a 'perfect test'. A perfect test is one where $p(D|C) = p(\text{not } C|\text{not } D) = 1$. Such a test never gives any false positive or false negative diagnoses. But no real test is perfect, so how can we determine the value of $p(C|D)$? How believable is the diagnosis?

Bayes' Theorem

Bayes' genius lay in realizing he could use a well known logical principle to answer this question: The probability of two independent events, *C* **and** *D*, both being true is called the *joint probability* $p(CD)$ and is equal to:

$$p(CD) = p(C|D) p(D) \tag{1}$$

That is to say the probability of *C* and *D* **both** being true is equal to the product of the probability of *C* being true given that *D* **is** true, times the probability that *D* is true in the first place. Think about it. It is just common sense.

This joint probability can also be written as

$$p(CD) = p(D|C) p(C) \tag{2}$$

This expression is saying that the probability that **both** *C* and *D* are true is also equal to the product of the probability of *D* being true given that *C* **is** true, times the probability that *C* is itself actually true.

Example: Suppose I only go to the movies on Fridays. And, on any given Friday, the chances are 50%-50% that I will go. If I pick a day at random, what is the probability I will go to the movies that day? Answer: The probability any randomly chosen day will be a Friday is 1/7. Therefore $(1/7) \times (50\%)$ is approximately 7.1%.

Setting expressions (1) and (2) equal to each other, yields

$$p(C|D) p(D) = p(D|C) p(C) \tag{3}$$

and, solving for the quantity we are interested in yields:

$$p(C | D) = \frac{p(D | C)p(C)}{p(D)} \quad (4)$$

We know the values of the two numerator probabilities. But what is the value of the denominator? From the diagram (figure 1) we see that there are two paths by which we can get to the result, D (the upper right-hand node). Therefore:

$$p(D) = p(C)p(D | C) + p(\text{not } C)p(D | \text{not } C) \quad (5)$$

Making this substitution in the denominator yields:

$$p(C | D) = \frac{p(D | C)p(C)}{p(C)p(D | C) + p(\text{not } C)p(D | \text{not } C)} \quad (6)$$

This expression (6) is known as Bayes' Theorem and is easy to use *if* we know the conditional probabilities that describe the test and also the *a priori* probability of occurrence of the event we are looking for, $p(C)$. This expression gives us the answer to question 1 in the third paragraph above. All we have to do is fill in the numerical properties of our diagnostic test together with the *a priori* probability of occurrence of cancer in the general population (assuming we know it or can estimate it) and we will have our answer.

Let's use Bayes' Theorem (expression 6) to answer question #2.

We are trying to determine who is the best teacher in a school employing 30 teachers. And we are going to use a test which has been shown previously to identify the best teacher from similar groups of teachers with 95% accuracy (both false negatives and false positives have been only 5%). Expression (6) says: If this test is used to identify the best teacher (out of 30 in the faculty) the probability that the person identified really *is* the best is:

$$\frac{(0.95)(1/30)}{(1/30)(0.95) + (29/30)(0.05)} = 39.6\% \quad (7)$$

What? A less than 50% - 50% chance of being right? That's worse than just tossing a coin!

How can a test that is 'known to be 95% correct' give so unreliable a result? If you look carefully at the numbers in this example (expression 7) you will see that the low value of the result is primarily caused by the second factor in the numerator (1/30). If that number were instead, say, 15/30, (In that case you would be asking, "Who is in the upper half of the faculty?"), then the result does come out to be 0.95 or 95%. That's better! This test can identify who are *among* our best (above average) teachers quite reliably. But we cannot reliably identify *the* 'best' out of thirty.

The moral of the story is this: ***If you are looking for the occurrence of something that is comparatively rare in the general population, your test has to be almost perfectly reliable or you will be fooling yourself if you believe the result you get from using it.***

So, if you are going to use a test (to try to calculate an *a posteriori* probability), always remember the following three rules:

1. The ‘correct’ conditional probabilities of the test, $p(D|C)$ and $p(\text{not } D|\text{not } C)$, must be extremely high.
2. If the thing (event, quality) you are seeking to identify is relatively rare (generally occurs with low probability), or if you have only a very rough idea about the value of that *a priori* probability, you ought to re-think the question you are asking.
3. Ask yourself whether the situation at hand actually calls for a probability computation at all or something different.

This third rule (above) is a subtle but extremely important one. For example, consider the following story: One night two hikers camped out in a tent in Indonesia. They had a few drinks before retiring and so slept very soundly until one of them awoke abruptly and screamed. His companion jumped up and asked what was going on. The first camper yelled, “A tiger just poked his head into our tent!” The second camper (being a good statistician) began to compute the *a posteriori* probability that a tiger had actually been there. The one who had seen the tiger yelled, “I SAW THE DAMNED THING WITH MY OWN EYES! Don’t start with the probability arguments, you dolt!”

This is, word-for-word, the answer that should be given in the following case as well:

An astronomer obtains an image of a highly redshifted object (*QSO*) that appears to be in front of a low redshifted galaxy. For example, see: <http://www.thunderbolts.info/tpod/2004/arch/041001quasar-galaxy.htm> Other astronomers are unconvinced and demand that he should evaluate the *a posteriori* probability that the *QSO* is indeed closer to us than the galaxy.

In this case, examining data is not a matter of ‘probabilities’ (neither *a priori* nor *a posteriori*). It is simply a question of do you believe the evidence or not. If not, then you must be prepared to say why not. Are you accusing the presenter of the evidence of counterfeiting it? Are you saying the *QSO* is an ‘artifact’ and not really there? To raise probabilistic arguments in cases where the evidence is ‘in your face’ is simply an evasion. It is dissembling. It is dishonest. When you have *prima facie* evidence of something, you do not need to initiate a ‘test’ to determine *a posteriori* probabilities. It is therefore incorrect to refer to ‘*a posteriori* probabilities’ when no test, as such, has been performed.

For example, astronomer Halton Arp has presented a long series of images of unusual concentrations of BL Lac objects relatively near Seyfert galaxies. In order to quantify his observation, Arp calculated the average density of these objects (the number of them per

square degree) over the entire sky. He then compared that BL Lac density measurement to their observed densities in small areas centered on Seyfert galaxies. He determined the ratio of those densities, to be $> \sim 10,000$. That is to say, the probability of finding a BL Lac near to a Seyfert galaxy is at least 10,000 times greater than the probability of finding one alone in an equal sized area chosen randomly on the open sky.

Arp said¹ that one of his detractors snorted, “Ridiculous!” He claimed no one would believe such a probability figure. And besides, “it is *a posteriori* – computed after you found the effect.” This critic wanted to reject Arp’s evidence – he had no inclination to acknowledge its existence – and he grabbed at the nearest straw he could find – denouncing Arp’s probability calculation as being an inaccurate *a posteriori* probability.

First of all, it wasn’t an *a posteriori* probability. As we have seen, an *a posteriori* probability quantifies how much you can trust the results presented by some kind of *test*. Arp was not performing any kind of test. He was simply comparing the observed density of a certain type object in one place on the sky contrasted with any other location. Arp ‘saw the tiger in the tent with his own eyes’ – he did not need to perform any kind of test in order to verify those observations.

Throwing around the descriptor ‘*a posteriori*’ in a pejorative attempt to belittle Arp’s work, clearly demonstrates that the critic either does not understand probability theory – or hopes that we don’t.

But we do.

Don Scott

¹ *Seeing Red*, pub Apeiron, Montreal, 1998, P.49.