Conditional Probability

Thus far, we have studied independant probability. However, in many situations, past observations can be effectively used to predict the future or unknown situations.

In order to illustrate this idea, let's consider a simple example which has shown up in many books of brain teasers (and also done in a fairly famous column by Marilyn Vos Savant who is alleged to have the highest IQ in the world):

"Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the other doors, opens another door, say No. 3, which has a goat. He then says to you, 'Do you want to pick door No. 2?' Is it to your advantage to take the switch?"

The answer, surprisingly is yes. Consider that when you first select a door, the odds are 1/3 of picking the car. Let's assume you always pick door # 1. So, the scenarios are:

1. The car is behind door #1. You pick door #1. The host shows you door#2 has a goat. You switch to door #3. You lose.
2. The car is behind door #2. You pick door #1. The host shows you door #3 has a goat. You switch to door #2. You win.
3. The car is behind door #3. You pick door #1. The host shows you door #2 has a goat. You switch to door #3. You win.

In other words, by taking the "switching" strategy, you win 2/3 times as opposed to 1/3 by not switching.

Medical Results

A more true-to-life example can be seen by interpreting the results of medical exames. Let us imagine that you go into the doctor's office and are given a routine test for a particular disease. Now, it's important for this discussion that it be routine. If you have any *a priori* (ahead of time) reason to suppose you have the disease then naturally, that changes the nature of the ensuing discussion.

Further, let's imagine that only 1% of the population (10/1000) have this disease. The medical company which makes the disease has issued the following table:

Cases (per 1000)

| | Has Disease | Doesn't have disease |
|---|---|---|
| Positive | 9 | 50 |
| Negative | 1 | 940 |

How would the medical company describe the accuracy of their test?

Well, they could say its 90% accurate. After all, 9 times out of 10, if you have the disease, you will come up positive on the test. The other 1 time, when you have the disease and come up negative, is known as a *false negative*.

They could also say that the test is 5% accurate. After all, of the roughly 990 people who don't have the disease, only 5% come up positive on the test. These cases are known as *false negatives*.

But in reality, the odds are somewhat different. What you'd really like to know is not:

> What are the odds, given that you have the disease, of having a positive result on the test?

But rather:

> What are the odds, given that you have a positive result on the test, of actually having the disease?

I want you to think about the difference between the two statements for a bit, because they really are all the difference in the world. Consider that out of every 1000 people who take the test, 59 of them get a positive. However, of those, only 9 really have the disease. 9/59=15%. In other words, even though you tested positive, the odds are actually quite low that you have the disease.

Bayes' Theorem

Statisticians have a fancy rule that they use to describe posterior probabilities. It is known as Bayes' Theorem. As an equation, it reads:

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{1}$$

The term, $P(A|B)$ means, the probability that some statement, $A$, is true (in our example – "You have the disease"), given that you already *know* that statement $B$ is true (in our example – "You tested positve). The term $P(A,B)$ is the probability that both are true. And $P(B)$ is the *a priori* probability that $B$ would come up true.

What's the probability of having the disease and coming up positive? 0.9% What's the *a priori* probability of coming up positive? 5.9%. And, thus, the probability of having the disease, given that you came up positive is 0.9/5.9=15%. Exactly as we found before.

DNA Evidence

The reason I bring all of this up is that conditional probability is one of the most misunderstood and misused forms of statistics. One particularly pervasive example is in the use of DNA evidence. Consider the following. A defendant, John Doe, has been arrested for murder. A bit of blood was found on the murder weapon, and, after doing a DNA test, the results came back a perfect match. Moreover, the lab report said that only 1/1 Million people in the country would have come up as a match on this test.

So, the prosecuter makes his case:

Using only the DNA evidence, the probability is 999,999/1,000,000 that Mr. Doe is guilty. Since the standard is "beyond a reasonable doubt" the jury has no choice but to convict.

The defense makes the argument:

No. There are 300,000,000 people in the U.S. That means that 300 would have the same DNA and thus, the probability, given that only 1 of them actually did commit the crime, is only 1/300, about 0.3%, and thus, the defendant should be aquitted.

Who's right?

In reality, we realize that there is an unspoken assumption, one that puts mathematics and legality at odds. Officially, a defendant is presumed innocent until proven guilty. In reality, though, we assume that because someone has been arrested that there is some reasonable chance that they are guilty. So the question is: What is the prior probability that a person on the stand is innocent? If we call it P(I), then, and P(+) is the probability of testing positive on the DNA test:

$$P(I|+) = \frac{0.000001 \times P(I)}{0.000001 \times P(I) + (1 - P(I))}$$

where 1-P(I) is simply the a priori probaility that the defendant (based on other evidence) is guilty. (He must be one or the other). Consider the cases where based on other evidence, the chance that he's innocent is 0.999. In that case, the DNA evidence overwhelmingly changes sour view, and the posterior probability of his innocence is merely 1/1000! Easily enough to convict.