Lecture 27: Probability (continued)

We saw a couple of examples where probability calculations may be counter-intuitive. These highlight the fact that you need to be *very careful* while discussing or thinking about probabilities!

Inclusion-exclusion

Let's play another dice game. You pick a number N between 1 and 6. Then, you roll three fair dice. If at least one of the numbers that come up match the number you picked, then you win; else, you lose. Is this a fair game?

It *does* seem like there is a 50/50 chance of you winning. Here is a "proof" of this claim. Let A_i be the event that dice i $(i = 1, 2, 3)$ comes up as N. Clearly, if the original dice are all fair, then the probability of each A_i is 1/6, i.e.,

$$P(A_i) = 1/6.$$

So the probability of the three of them combined is

$$P(A_1) + P(A_2) + P(A_3) = 1/6 + 1/6 + 1/6 = 1/2.$$

Seems like this is a fair game, right?

But there is a **bug** here. If instead of 3 dice, we have 7 dice, then the same logic applies — but then the sum of the probabilities is now 7/6, which does not make sense!

Really, what you are after is *not* the sum of the probabilities, but rather, the probability of the *union* of these events, i.e., we want to compute

$$P(A_1 \cup A_2 \cup A_3)$$

which, by the Principle of Inclusion and Exclusion, is given by:

$$P(A_1)+P(A_2)+P(A_3)-P(A_1\cap A_2)-P(A_2\cap A_3)-P(A_1\cap A_3)+P(A_1\cap A_2\cap A_3)$$

Here, $P(A_1 \cap A_2)$ is the probability of *both* dice 1 and dice 2 coming up N, which is 1/36. Similarly, $P(A_2 \cap A_3) = P(A_1 \cap A_3) = 1/36$. Finally $P(A_1 \cap A_2 \cap A_3) = 1/6 * 1/6 * /16 = 1/216$. Plugging in all the numbers, we get the probability of you winning:

$$P(A_1 \cup A_2 \cup A_3) = 1/2 - 1/12 + 1/216 = 0.421..$$

which means that it is *definitely* not a fair game!

Conditional probability

We now define what we call *conditional* probability. Given two events (say A and B), we define the *probability of A given B*, written as

$$P(A|B)$$

as the ratio of the probability of A and B happening, divided by the probability of B. So given that B happened, the probability that A also happened is given by:

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

Conditional probability is useful for modeling all kinds of scenarios. Here is an example.

Suppose that 2 teams play a best 2-out-of-3 series. Suppose that the probability of winning the first game for each time is 1/2, but the probability of winning *following a win* is 2/3. (Momentum is on your side, hence you are more likely to win.) Given that you have won the first game, what is the likelihood of you winning the series?

The way to solve this as usual is to identify the sample space. Since this is a best of 3, there are 6 outcomes:

- WW
- WLW
- WLL
- LWW
- LWL
- LL

(Draw the tree out to visualize this better.)

You can also compute the probabilities of each outcome (leaf):

- $P(WW) = 1/2 * 2/3 = 1/3$
- $P(WLW) = 1/2 * 1/3 * 1/3 = 1/18$
- $P(WLL) = 1/2 * 1/3 * 2/3 = 1/9$
- $P(LWW) = 1/2 * 1/3 * 2/3 = 1/9$
- $P(LWL) = 1/2 * 1/3 * 1/3 = 1/18$
- $P(LL) = 1/2 * 2/3 = 1/3$

Now, let A be the event that you win the series, and B is the event that you won the first game. It is easy to see that the probability of A is $1/3+1/9+1/18 = 1/2$, the probability of B is also $1/2$, and the probability of both happening ($A \cap B$) is $1/3 + 1/18 = 7/18$. Therefore, the probability of you winning the series, given that you won the first game, is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{7/18}{1/2} = \frac{7}{9}.$$

Therefore, you are very likely to win the series!

Testing and conditional probabilities

Again, when discussing conditional probabilities, intuition can be very confusing. Here is a simple example involving a medical test. (Similar applications arise in other diagnostic contexts as well, e.g. testing whether a chip is defective or not.)

Assume that 10% of the population has a certain disease. To diagnose this disease, there is a certain test, which is imperfect, that succeeds with the following parameters:

- If you have the disease, the test returns TRUE with 90% probability and FALSE with 10% (so the "false negative rate" is 10%)
- If you don't have the disease, the test returns TRUE with 30% probability and FALSE with 70% (so that the "false positive rate" is 30%.)

This looks like a reasonable test. Now, given that you have a random person who tested positive, what is the likelihood that the person actually has the disease? Seems like it is pretty likely, right?

We solve this using conditional probability. Again, the 4 outcomes here are:

- (Has disease, tested positive)
- (Has disease, tested negative)
- (Doesn't have disease, tested positive)
- (Doesn't have disease, tested negative)

and the probability for each of the above is:

- 1/10*9/10 = 0.09
- 1/10*1/10 = 0.01
- 9/10*3/10 = 0.27
- 9/10*7/10 = 0.63

Let A be the event that the person has the disease. The probability of this is $0.09 + 0.01 = 0.1$. Let B be the event that the person tested positive. The probability of this is $0.09 + 0.27 = 0.36$. The probability of $A \cap B = 0.09$. Therefore, the conditional probability $P(A|B)$ is the probability that the person has the disease, given that they tested positive, which is given by:

$$\frac{P(A \cap B)}{P(B)} = \frac{0.09}{0.36} = \frac{1}{4}.$$

So there is only a 1-in-4 chance that they actually have the disease!

Where is the catch here? Is the test bad? Actually, it is pretty good – the probability of the test being correct is the sum of the probabilities of the two outcomes – (Has disease, tested positive) and (Doesn't have disease, tested negative) – which is equal to $0.09 + 0.63 = 0.72$. So there is close to a 3-in-4 chance of the test being correct!

In summary, the test is very likely to be correct on average, but if the person had the misfortune of actually having the disease, the test is very likely to be wrong. Such weirdness is common while discussing conditional probabilities (and unfortunately, also leads to inaccurate interpretation of test results in many cases.)