CHAPTER 1

Conditional Probability

Let's say there is a virus going around, and there is a vaccine for it that requires two shots. You are working at a school, and you are wondering how effective the vaccines are. Some students are unvaccinated, some have had one shot, and some have had two shots. One day, you test all 644 students to see who has the virus. You end up with the following table:

	V_0	V_1	V_2
T_+	88 students	36 students	96 students
T_-	92 students	76 students	256 students

Here is what each symbol means:

- V_0: student has had zero vaccination shots

- V_1: student has had one vaccination shot

- V_2: student has had both vaccination shots

- T_+: student tested positive for the virus

- T_-: student tested negative for the virus

So, for example, your data indicates that 76 students who had only one of the two shots tested negative for the virus.

Your principal has a few questions. The first is, "If I put five randomly chosen students in a study group together, what is the probabiltiy that one of them has the virus?"

The first thing you might do is make a new table that shows the probability of a randomly chosen student being in any particular group. You just divide each entry by 644 (the total number of students).

	V_0	V_1	V_2
T_+	$p(V_0 \text{ AND } T_+) = 13.7\%$	$p(V_1 \text{ AND } T_+) = 5.6\%$	$p(V_2 \text{ AND } T_+) = 14.9\%$
T_-	$p(V_0 \text{ AND } T_-) = 14.3\%$	$p(V_1 \text{ AND } T_-) = 11.8\%$	$p(V_2 \text{ AND } T_-) = 39.8\%$

(In this table, we expressed the number as a percentage with a decimal point — you had to round off the numbers. If you wanted exact answers, you would have to keep each as a fraction: 36 students represents $\frac{9}{161}$ of the student body.)

1.1   Marginalization

Now we can sum across the columns and rows.

	V_0	V_1	V_2	sum
T_+	0.137	0.056	0.149	$p(T_+) = 0.342$
T_-	0.143	0.118	0.398	$p(T_+) = 0.547$
sum	$p(V_0) = 0.280$	$p(V_1) = 0.174$	$p(V_2) = 0.547$	

If a child is chosen randomly from the entire student body, there is a 34.2% that the student has tested positive for the virus, and there is 17.4% chance that the student has one shot of the vaccine.

This summing of the probabilities across one dimension is known as *marginalizing*. Marginalization is just summing across all the variables that you don't care about. If you don't care who has the virus, just the probability that a student has not received even one shot of the vaccine, you can simply marginalize all the vaccine statuses. To answer the principal's question, the easy thing to do is find the answer of the opposite: "if I put five randomly chosen students in a study group together, what is the probabiltiy that *none* of them has tested positive for the virus?"

The chance that a randomly chosen student doesn't have the virus ($p(T_-)$ is 54.7%. This means the chance that 5 randomly chosen students don't have the virus is $0.547 \times 0.547 \times 0.547 \times 0.547 \times 0.547 = 0.0489$ Thus, the probability of the opposite is $1.0 - 0.0489 = 0.951$

The answer, then, is "If you put 5 kids in a study group together, there is a 95.1 % probability that at least one of them has the virus."

1.2   Conditional Probability

Now the principal asks you, "What if I make a group of five kids who have had both shots of the vaccine? What are the odds that one of them has tested positive for the virus?"

This involves the idea of *Conditional probability*. You want to know the odds that a student doesn't have the virus, given that the student has had both shots of the vaccine.

There is a mathematical notation for this:

$$p(T_-|V_2)$$

That is the probability that a student who has had both vaccination shots will test negative for the virus.

How would you calculate this? You would count all the students who had a positive test *and* both vaccination shots, which you would divide by the total number of students who had both vaccination shots.

$$p(T_-|V_2) = \frac{256}{96 + 256} = \frac{8}{11} \approx 72.7\%$$

If we are working from the probabilities, you can get the same result this way. Divide the probability that a randomly chosen student had a positive test *and* both vaccination shots by the probabiltiy that a student had both vaccination shots:

$$p(T_-|V_2) = \frac{p(T_- \text{ AND } V_2)}{p(T_-)} = \frac{0.398}{0.547} \approx 72.7\%$$

Notice that this is different from $p(V_2|T_-)$, which is the probability that a student has had both vaccinations, given they tested negative for the virus.

Back to the principal's question: "If you have five students who have had both vaccinations, what is the probability that all of them tested negative for the virus?" The probability that one student is virus-free is $\frac{8}{11}$, so the probability that five students are virus-free is $\frac{8}{11}^5 \approx 0.203$. So, there is a 79.6% chance that at least one of the five has the virus.

1.3   Chain Rule for Probability

You just used this equality: For any events A and B

$$p(A|B) = \frac{p(A \text{ AND } B)}{p(B)}$$

This is more commonly written like this:

$$p(A \text{ AND } B) = \frac{p(A|B)}{p(B)}$$

This is an abstract way of writing the idea, but the idea itself is pretty intuitive. The probability that you are going to buy a ticket and win the lottery is equal to the probability that you buy a ticket times the probability that you win, given that you have bought a ticket. (Here A is "win the lottery" and B is "buy a ticket".)

This is known as *The Chain Rule of Probability*. We can chain together as many events as we want: The probability that you are going to die in the car that you bought with your winnings from the lottery ticket you bought is:

$$p(W \text{ AND } X \text{ AND } Y \text{ AND } Z) = p(W|X \text{ AND } Y \text{ AND } Z)p(X|Y \text{ AND } Z)p(Y|Z)p(Z)$$

where

- W = Dying in car accident
- X = Buying a car with lottery winnings
- Y = Winning the lottery
- Z = Buying a lottery ticket

In English, the equation says:

"The probability that you will die in a car accident, buy a car with lottery winnings, win the lottery, and buy a lottery ticket is equal to the probability that you buy a lottery ticket times the probability that you win the lottery (given that you have bought a ticket) times the probability that buy a car with those lottery winnings (given that bought a ticket and won) times the probability that you crash that car (given that you have bought the car, won the lottery, and bought a ticket)."

This is a draft chapter from the Kontinua Project. Please see our website (https://kontinua.org/) for more details.

Answers to Exercises

INDEX