**MITOCW | conditional_probability**

You've tested positive for a rare and deadly cancer that afflicts 1 out of 1000 people, based on a test that is 99% accurate. What are the chances that you actually have the cancer? By the end of this video, you'll be able to answer this question!

This video is part of the Probability and Statistics video series. Many natural and social phenomena are probabilistic in nature. Engineers, scientists, and policymakers often use probability to model and predict system behavior.

Hi, my name is Sam Watson, and I'm a graduate student in mathematics at MIT.

Before watching this video, you should be familiar with basic probability vocabulary and the definition of conditional probability.

After watching this video, you'll be able to: Calculate the conditional probability of a given event using tables and trees; and Understand how conditional probability can be used to interpret medical diagnoses.

Suppose that in front of you are two bowls, labeled A and B. Each bowl contains five marbles.

Bowl A has 1 blue and 4 yellow marbles. Bowl B has 3 blue and 2 yellow marbles.

Now choose a bowl at random and draw a marble uniformly at random from it. Based on your existing knowledge of probability, how likely is it that you pick a blue marble? How about a yellow marble?

Out of the 10 marbles you could choose from, 4 are blue. So the probability of choosing a blue marble is 4 out of 10.

There are 6 yellow marbles out of 10 total, so the probability of choosing yellow is 6 out of 10.

When the number of possible outcomes is finite, and all events are equally likely, the probability of one event happening is the number of favorable outcomes divided by the total number of possible outcomes.

What if you must draw from Bowl A? What's the probability of drawing a blue marble, given that you draw from Bowl A?

Let's go back to the table and consider only Bowl A. Bowl A contains 5 marbles of which 1 is blue, so the probability of picking a blue one is 1 in 5.

Notice the probability has changed. In the first scenario, the sample space consists of all 10 marbles, because we are free to draw from both bowls.

In the second scenario, we are restricted to Bowl A. Our new sample space consists of only the five marbles in

Bowl A. We ignore these marbles in Bowl.

Restricting our attention to a specific set of outcomes changes the sample space, and can also change the probability of an event. This new probability is what we call a conditional In the previous example, we calculated the conditional probability of drawing a blue marble, given that we draw from Bowl A.

This is standard notation for conditional probability. The vertical bar ( | ) is read as "given." The probability we are looking for precedes the bar, and the condition follows the bar.

Now let's flip things around. Suppose someone picks a marble at random from either bowl A or bowl B and reveals to you that the marble drawn was blue. What is the probability that the blue marble came from Bowl A?

In other words, what's the conditional probability that the marble was drawn from Bowl A, given that it is blue? Pause the video and try to work this out.

Going back to the table, because we are dealing with the condition that the marble is blue, the sample space is restricted to the four blue marbles.

Of these four blue marbles, one is in Bowl A, and each is equally likely to be drawn.

Thus, the conditional probability is 1 out of 4.

Notice that the probability of picking a blue marble given that the marble came from Bowl A is NOT equal to the probability that the marble came from Bowl A given that the marble was blue. Each has a different condition, so be careful not to mix them up!

We've seen how tables can help us organize our data and visualize changes in the sample space.

Let's look at another tool that is useful for understanding conditional probabilities - a tree diagram.

Suppose we have a jar containing 5 marbles; 2 are blue and 3 are yellow. If we draw any one marble at random, the probability of drawing a blue marble is 2/5.

Now, without replacing the first marble, draw a second marble from the jar. Given that the first marble is blue, is the probability of drawing a second blue marble still 2/5?

NO, it isn't. Our sample space has changed. If a blue marble is drawn first, you are left with 4 marbles; 1 blue and 3 yellow.

In other words, if a blue marble is selected first, the probability that you draw blue second is 1/4. And the

probability you draw yellow second is 3/4.

Now pause the video and determine the probabilities if the yellow marble is selected first instead.

If a yellow marble is selected first, you are left with 2 yellow and 2 blue marbles.

There is now a 2/4 chance of drawing a blue marble and a 2/4 chance of drawing a yellow marble.

What we have drawn here is called a tree diagram. The probability assigned to the second branch denotes the conditional probability given that the first happened.

Tree diagrams help us to visualize our sample space and reason out probabilities.

We can answer questions like "What is the probability of drawing 2 blue marbles in a row?" In other words, what is the probability of drawing a blue marble first AND a blue marble second?

This event is represented by these two branches in the tree diagram.

We have a 2/5 chance followed by a 1/4 chance. We multiply these to get 2/20, or 1/10. The probability of drawing two blue marbles in a row is 1/10.

Now you do it. Use the tree diagram to calculate the probabilities of the other possibilities: blue, yellow; yellow, blue; and yellow, yellow.

The probabilities each work out to 3/10. The four probabilities add up to a total of 1, as they should.

What if we don't care about the first marble? We just want to determine the probability that the second marble is yellow.

Because it does not matter whether the first marble is blue or yellow, we consider both the blue, yellow, and the yellow, yellow paths. Adding the probabilities gives us 3/10 + 3/10, which works out to 3/5.

Here's another interesting question. What is the probability that the first marble drawn is blue, given that the second marble drawn is yellow?

Intuitively, this seems tricky. Pause the video and reason through the probability tree with a friend.

Because we are conditioning on the event that the second marble drawn is yellow, our sample space is restricted to these two paths: P(blue, yellow) and P(yellow, yellow).

Of these two paths, only the top one meets our criteria - that the blue marble is drawn first.

We represent the probability as a fraction of favorable to possible outcomes. Hence, the probability that the first marble drawn is blue, given that the second marble drawn is yellow is 3/10 divided by (3/10 +3/10), which works out to 1/2.

I hope you appreciate that tree diagrams and tables make these types of probability problems doable without having to memorize any formulas!

Let's return to our opening question. Recall that you've tested positive for a cancer that afflicts 1 out of 1000 people, based on a test that is 99% accurate.

More precisely, out of 100 test results, we expect about 99 correct results and only 1 incorrect result.

Since the test is highly accurate, you might conclude that the test is unlikely to be wrong, and that you most likely have cancer.

But wait! Let's first use conditional probability to make sense of our seemingly gloomy diagnosis.

Now pause the video and determine the probability that you have the cancer, given that you test positive.

Let's use a tree diagram to help with our calculations.

The first branch of the tree represents the likelihood of cancer in the general population.

The probability of having the rare cancer is 1 in 1000, or 0.001. The probability of having no cancer is 0.999.

Let's extend the tree diagram to illustrate the possible results of the medical test that is 99% accurate.

In the cancer population, 99% will test positive (correctly), but 1% will test negative (incorrectly).

These incorrect results are called false negatives.

In the cancer-free population, 99% will test negative (correctly), but 1% will test positive (incorrectly). These incorrect results are called false positives.

Given that you test positive, our sample space is now restricted to only the population that test positive. This is represented by these two paths.

The top path shows the probability you have the cancer AND test positive. The lower path shows the probability that you don't have cancer AND still test positive.

The probability that you actually do have the cancer, given that you test positive, is

(0.001*0.99)/((0.001*0.99)+(0.999*0.01)), which works out to about 0.09 - less than 10%!

The error rate of the test is only 1 percent, but the chance of a misdiagnosis is more than 90%! Chances are pretty good that you do not actually have cancer, despite the rather accurate test. Why is this so?

The accuracy of the test actually reflects the conditional probability that one tests positive, given that one has cancer.

But in practice, what you want to know is the conditional probability that you have cancer, given that you test positive! These probabilities are NOT the same!

Whenever we take medical tests, or perform experiments, it is important to understand what events our results are conditioned on, and how that might affect the accuracy of our conclusions.

In this video, you've seen that conditional probability must be used to understand and predict the outcomes of many events. You've also learned to evaluate and manage conditional probabilities using tables and trees.

We hope that you will now think more carefully about the probabilities you encounter, and consider how conditioning affects their interpretation.