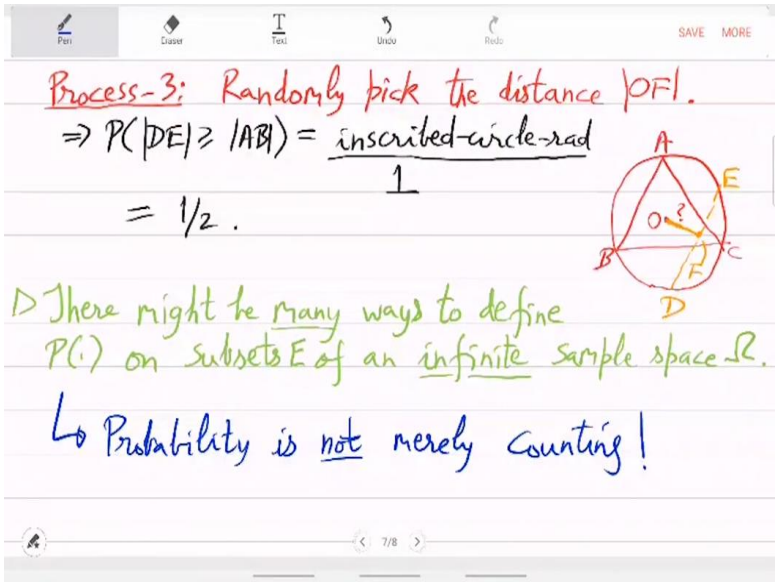
Probability for Computer Science
Prof. Nitin Saxena
Department of Computer Science and Engineering
Indian Institute of Technology - Kanpur

Module - 1
Lecture - 2
Examples and Course Outline

(Refer Slide Time: 00:13)



So, we are not picking F, but we are actually picking the length OF in a random way. And this length decides also the length of the chord DE, because it actually uniquely defines. Well, it will not uniquely define the chord, but it will at least uniquely define the length of DE. That is what the probability cares about. So, what are the good lengths? OF. Clearly from the previous circle, inscribed circular argument, the good length of OF is half or less.

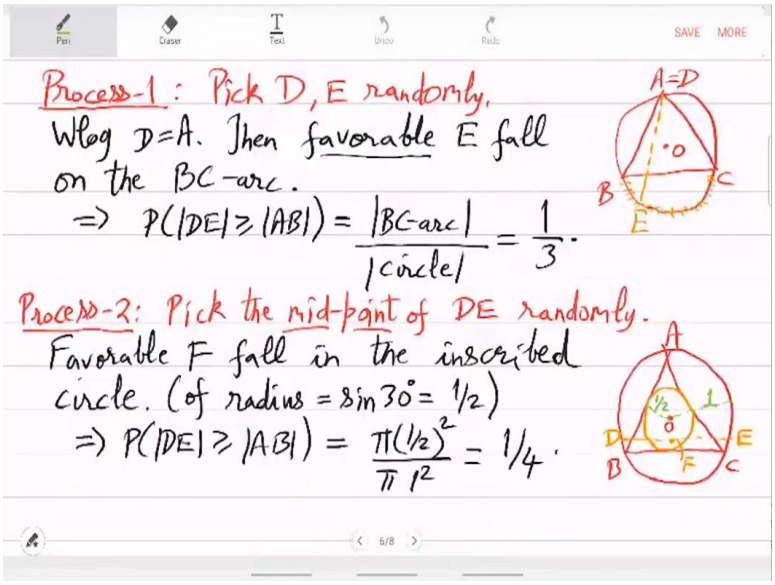
So, this means that probability that the length DE is at least the length AB. This is equal to the inscribed radius over the circumscribed circle radius, which is 1. So, that is half. Because, if you take an F which is, which we take OF to be bigger than half, then you go outside the inscribed circle and then you get smaller chords. So, you have 3 possibilities of the probability of same events. It is one fourth, one third and one half.

So, they differ so much? So, what is the paradox here? I mean, paradox must be clear, but how do you resolve the paradox? Where is the mistake in the math? So, the mistake is that, or the reason why you are getting the different probabilities is because you are assigning

probability in different ways to the sample space. Each of these 3 processes is defining a different probability function, that essentially is the resolution.

What you have to remember is, there might be many ways to define the probability function on subsets of an infinite sample space omega. So, you, for example, process 1 defined probability on D, E.

(Refer Slide Time: 03:09)

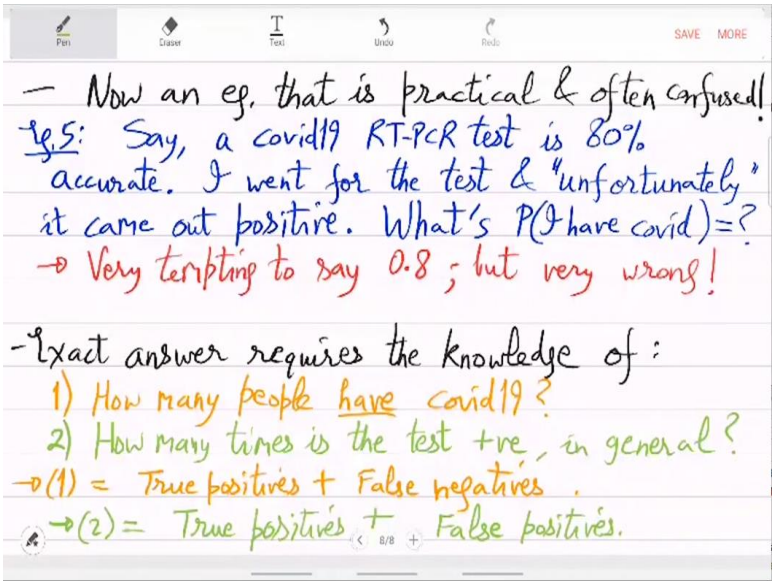


Process 2 defined it on midpoint of D, E and process 3 defined it on the length OF. And as you look at these different maps, you ultimately get different probabilities for this event DE greater than equal to AB. So, that is why we will need some formalisms, otherwise we get into this. If you only care about physical interpretation, then you might get confused in these examples.

So, in the end, we will formally define everything, but this is a good example to keep in mind. So, this also tells you that probability is not merely counting, because in infinite space, you cannot really count, counting does not make sense, but probability still makes sense. And to get a unique answer, you have to formalise it properly. So, here you see the difference between counting and probability.

In discrete space, you can say that they are equivalent, but not in infinite space. Probability is doing something else. So, now let us shift gears and move to a less abstract, more practical example, but still confusing. Let us go to example 5.

(Refer Slide Time: 05:09)



Now, an example that is practical and often confused. It can cause serious confusion, not only in an individual, but even in a society. And right now, it is very relevant. So, let us see this carefully. So, say a COVID-19 RT-PCR test is 80% accurate. So, you went, or I went to a lab which does this test and they say that it's accuracy is 80%. Now, so, I went for this test. And unfortunately, it came out to be positive, which you do not want to happen, but say it happened.

So, I took this test which came out to be positive. Now, my question is, what can you deduce from this information? So, what is the probability that I have COVID-19? Given this information that the test is positive, can I deduce? With what probability, with what confidence, with what chance can I say that I have the COVID virus or I do not have the COVID virus? Both the things are possible because the test is not perfect.

So, can I assign probabilities to this? Now, it is very tempting to say 0.8. The probability that I have COVID is 80%. But this is very wrong. This is what many people, in fact, most people would not realise that this 80% accuracy does not mean that this particular specific test that happened makes me COVID positive with 80% chance. So, calculating this is more complicated.

So, this exact answer; the formula for the exact answer, you will see later in the course; but without going into the formula, I will just say that exact answer depends on, it requires the knowledge of: 1, how many people in the, let us say in the country, have COVID-19 virus?

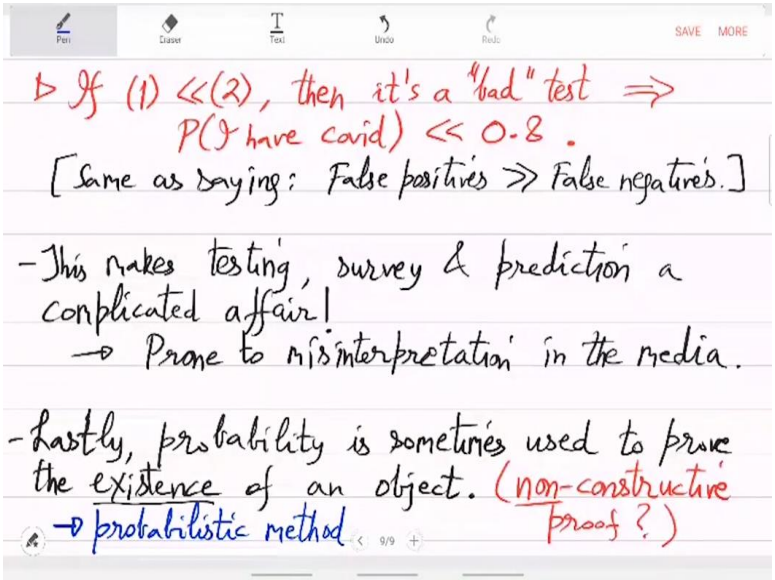
And second is, if the test was done on everybody, how many times is it, is the answer positive? So, how many times is the test result, in general?

These are different questions and we do not have information for them. We do not know how many people have the virus in the country. And we also do not know how many times will the test be positive, because the test cannot be done throughout the nation, for everybody. The test is only done on a so small sample space. So, both these things, we do not have. So, in practice, they are somehow approximated.

And once we have the approximation, then I can say what is the probability that I have COVID, given that the test was positive. But this information is currently missing from example 5. So, to say more about this, this 1 is called; so, 1 is actually equal to what we say true positives plus false negatives. Because true positives are those people who have the virus and the result was positive. So, these are the true positives.

False negatives are those people who have the virus, but the result was negative. So, that is what number 1 is. What is number 2? So, this is only about test being positive. Whether the person has virus or not, we are not worried about that. So, this is true positives plus false positives. So, that is the information, extra information needed. So, you need true positives, you need false negatives and you need false positives. These 3 things you need, and then you can do a calculation.

(Refer Slide Time: 11:34)



Now, interestingly, if this number 1 is much smaller than number 2, which means that very few people have actually the virus COVID-19. On the other hand, the test is many times positive, in the general population. So, then it is a bad test. That means that the test is bad. So, then, it is a bad test which implies that probability that I have COVID is very small. It is much smaller than 0.8.

So, in case the test is a bad test, then the probability of I having COVID is not at all 0.8. It is much smaller. It may even be close to 0. So, basically, this test does not give you any information, except it confuses you, because the accuracy was touted as 80%. So, which is the same as saying; so, 1 much less than 2 is the same as saying that false positives are much more than false negatives.

1 much less than 2 means that this true positive plus false negative is much less than true positive plus false positive, which is cancelling out. It says that false positives are much more than false negatives. So, it is a bad test. So, taking the test and getting positive does not mean that you have the virus. So, you can see that these things are quite complicated. So, when somebody advertises 80% accuracy, that just means that the test was done on a small sample space.

And there, people who had the virus, test became positive, with 80% chance. And people who did not have the virus, test was negative with 80% chance. But from that, when you take the test, you cannot deduce anything. You need much more information. You need information about the whole sample space, the general population. So, this is a very relevant and contemporary example, which of course you must have thought about.

So, this is the reason why. So, this makes testing, survey and prediction a very complicated affair. So, it is prone to a misinterpretation, specially in social media or even media, because part of the things, I mean, the initial part is correct, 80% is correct, but then the conclusion, you cannot conclude 80%, because you are, those are 2 different things. So, this is highly prone to misinterpretation in the media, because people can be easily made to convince themselves that this 80% is the same as what you want.

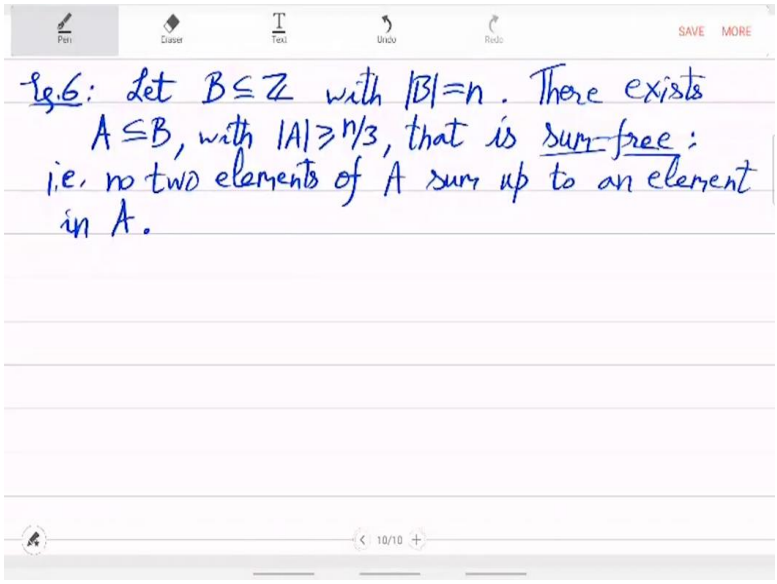
So, those were the examples. I hope these examples give you, not just remind you of basic probability, but also give you the challenges that lie in applying probability and why we will

need formalism, why we will need formulas. So, lastly, something that you might not have seen. So, lastly, another aspect of probability is to prove existence, usually of some discreet object. So, lastly, probability is sometimes used to prove the existence of an object.

So, sometimes it is also used to prove the existence of an object by calculating, you calculate probability of an event. If it comes out to be positive, then it means that object actually exists. So, this is a non-constructive proof. It may be constructive, but in general, these proofs are non-constructive proofs, because they only tell you the existence, they do not construct that object.

We will see some examples towards the end of this course. So, one example I can give you immediately. So, this is by the way called probabilistic method. Probabilistic method, specially in combinatorics, achieve this.

(Refer Slide Time: 18:09)



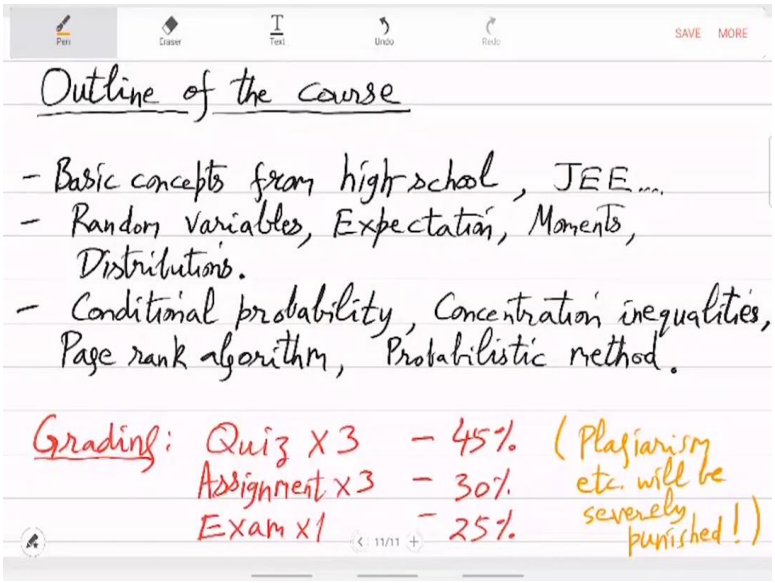
So, an example is: So, let B be a subset of integers and let it have n distinct integers. What you want to show is, there exists a subset A of B of size quite large, more than a third, more than n by 3. There exists a large subset that is sum-free. So, what is sum-free? Sum-free just means there are no two elements whose sum is a third element. So, no two elements of A sum up to an element in A.

So, this is what we want to show that. For any subset B of integers, there is always a sum-free subset that is large. This seems to be a difficult to prove statement, because there is no hope of a constructive proof, because you do not even know B. B can be anything. So, we will see

a proof or you can try to prove this by using probability. Those are the directions in which this course will move.

We will formalise concepts, we will define them and then we will apply them to many different examples problems, some of which may not have to do anything with probability, like this example 6. This is not about probability, per se.

(Refer Slide Time: 20:31)



So, finally, the outline of the course. So, we will do from next time, basic concepts from high school, the probability questions that you solved, those kinds of questions. Then we will look at random variables, expectation, moments, examples of distributions, probability distribution examples and some more advanced things which are very important in computer science applications, which is conditional probability; like this COVID testing example, that you can understand rigorously using conditional probability.

Concentration inequalities. Again concentration inequalities is something very important in computer science. Concentration basically means that, if I know that expectation or average of a random variable is something, let us say 10, then what is the probability that in practice you get a number like 20 or a number like 5. When the average is 10, what is the chance of being far away from it? far more or far less?

So, there are very useful inequalities about that, giving you probability upper bounds, also called tail inequalities. Then an idea about the PageRank algorithm; for example, which is used in or a variant of which is used in Google search. And finally the probabilistic method.

And finally, students who are registered inside IIT Kanpur, for them grading. So, tentatively, what we can do is, we can have quizzes, assignments and one exam.

It is a modular course. So, looking at the timeline, we can have 3 quizzes, we can have 3 assignments, we can have one End-Sem. So, quizzes, each can be 15%. So, that gives you 45% weightage. Assignment, we can have 10%. So, 30% weightage. And remaining 25% is for the End-Sem exam. So, based on your performance, you will be given a grade. Any version of plagiarism will be severely punished. So, plagiarism, etcetera will be severely punished, which obviously includes cheating from other students.