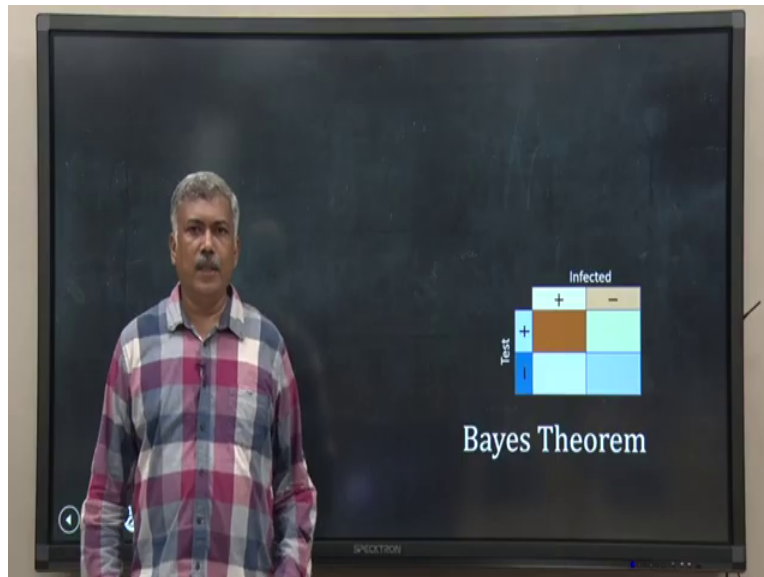


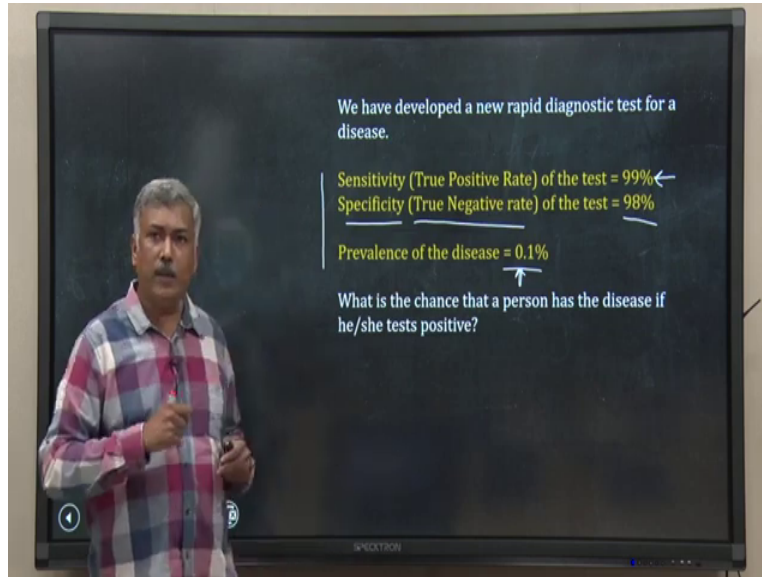
Data Analysis for Biologists
Professor Biplab Bose
Department of Biosciences & Bioengineering
Mehta Family School of Data Science & Artificial Intelligence
Indian Institute of Technology, Guwahati
Lecture: 6
Bayes Theorem and Likelihood

(Refer Slide Time: 00:32)



Hello, hope you are doing well. In this lecture, we will learn about Bayes Theorem. We will start with a numerical problem.

(Refer Slide Time: 00:40)



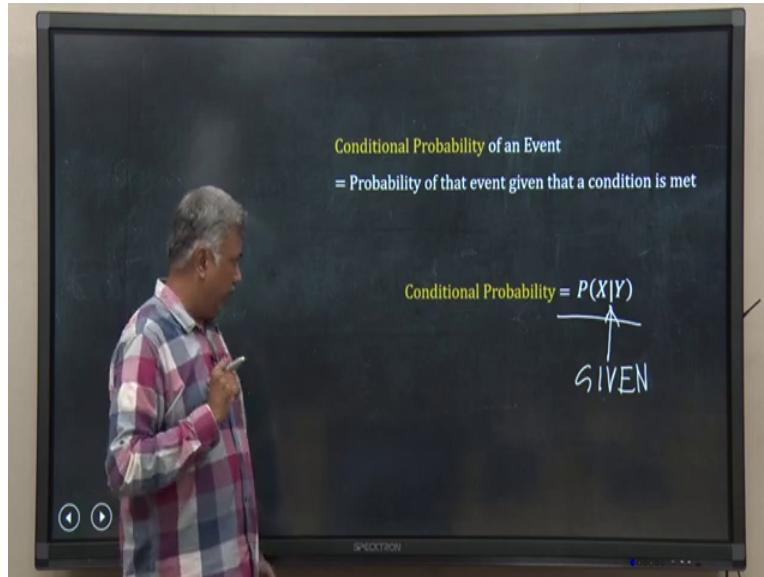
Suppose there is a disease and for that disease we have developed a rapid test, diagnostic test and its accuracy has been measured by a sampling, by testing on a sample of population. So, the information on accuracy is given here, the specificity is which is the true negative rate, specificity is that true negative rate.

That means, people who does not have the disease and they have tested negative in this test, this called specificity. is 98 percent whereas, the sensitivity of these tests, sensitivity mean true positive rate that means, if I am positive, what is the probability that it also tested positive when this diagnostic kit was tested? So, that is 99 percent.

And this disease is not so common, so, its prevalence is actually low, it is 0.1%, so, 1 out of 1000 people have this disease. So, suppose now, I walk into a clinic and I do not have any symptom or anything else, I walk into the clinic and ask them to do this rapid diagnostic test on me and the test say positive, my test say it is positive.

So, what is the chance or probability that I am also disease positive, means I am also, I also have the disease? To answer this type of question, we use Bayes theorem, but before we go into Bayes theorem and understand what it stands for, and how I can use this to answer this particular type of question, we have to understand conditional distribution.

(Refer Slide Time: 02:14)



What is conditional probability? Conditional probability will mean, conditional probability of an event mean what is the probability of that event given, and something has some condition has been met. For example, another event has happened. And we regularly use the idea of conditional probability in our day to day life, for example, right now, here it is often raining.

So, if I go out I have to take a umbrella. So, I decide whether to take the umbrella based upon my belief that whether it will rain or not. So, what do I do? I look out through the window and check whether there is a dark cloud or not, if there is dark cloud, I assume that probably there will be rain. So, that means I will take, I should take carry the umbrella.

So, here the probability that whether it will rain or not, that belief depends upon the condition that there is a dark cloud or light cloud or there is no cloud, something like that. So, this is what is conditional probability? And in mathematical language, in mathematical symbol, we represent conditional probability like this, $P(X|Y)$.

So, we call it $P(X|Y)$, this vertical line is given, it stands for given. So, this $P(X|Y)$ is the conditional probability of X given that the condition Y has met. Now, to understand mathematically what is actually conditional probability versus relation with other types of probability, I will take a simple biological example.

(Refer Slide Time: 03:54)

Gene A

Y: Mutated
N: Not Mutated

	Y	N
Gene B Y	50	30
Gene B N	20	100

Gene A

Y: Mutated
N: Not Mutated

	Y	N
Gene B Y	50	30
Gene B N	20	100

$n = 200$

Mutated
 \downarrow
 $P(A = Y)$

Gene A

Y: Mutated
N: Not Mutated

	Y	N
Gene B		
Y	50	30
N	20	100

$$P(A = Y) = \frac{50 + 20}{200} = 0.35$$

$$P(A = N) = \frac{30 + 100}{200} = 0.65$$

Gene A

Y: Mutated
N: Not Mutated

	Y	N
Gene B		
Y	50	30
N	20	100

$$P(A = Y) = \frac{50 + 20}{200} = 0.35$$

$$P(A = N) = \frac{30 + 100}{200} = 0.65$$

Marginal Probability

Suppose, there are two genes A and B, they often get mutated in particular type of cancer. So, I have collected 200 tumor sample from 200 people and then I have done some assay that you may have done, the whole genome sequencing or partial sequencing to check whether these two genes are mutated or not.

Now, I have found some result and I have represented that data for these 200-tumor sample, 200 patients in this tabular format. What does it shows? For example, I have 30 tumors where A, gene A is not mutated, but gene B is mutated. So, that is this one 30 whereas, there are 100 cases

out of these 200 where I have no mutation in A I have no mutation either in gene B also. So, in this way I represent the data.

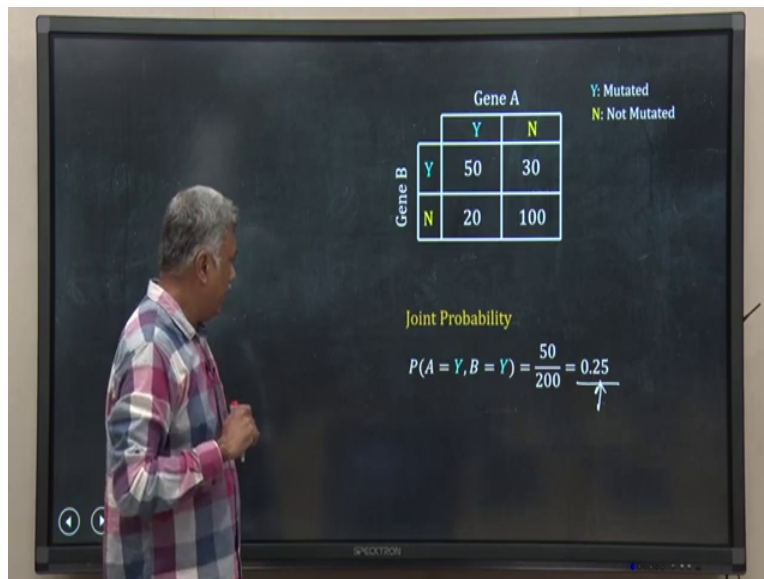
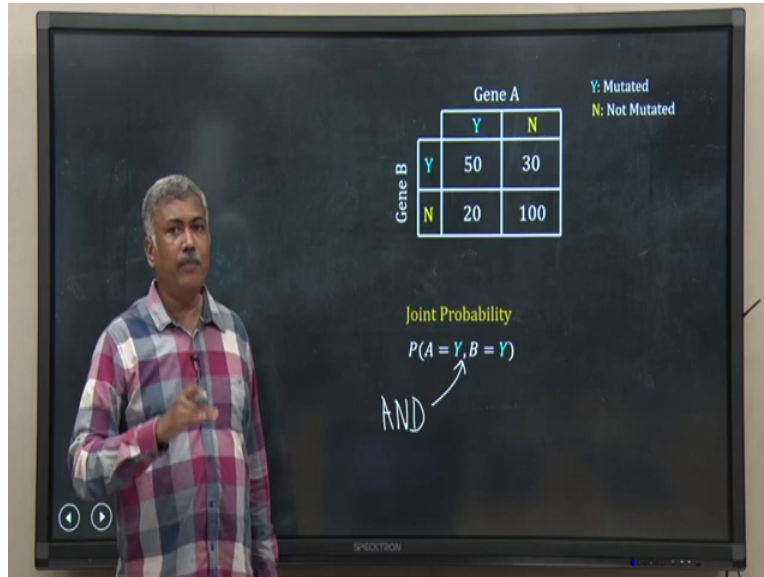
So, from these data, based upon these data, I want to calculate some probability. For example, I can ask, what is the probability that A, is mutated. Remember, we are representing, by Y we are representing it is mutated. So, I am asking you to calculate, what is $P(A=Y)$?

Now, I will take the assumptions that if I have good enough data, then the frequency of observation is equivalent to probability. So, to do this calculation to get the probability that A is mutated, what I will do, I will check the table. So, take the first row. In this first row, in both the cases for 50 and 20, gene A is mutated, out of how many, out of 200 sample total number of sample n small n is 200.

So, out of 200, 50 plus 20 is mutated sample. So, that means, my $P(A=Y) = 0.35$. Now, similarly I can ask what is the probability, that A is not mutated, then I will take this column. So, that is 30 plus 100. 130 cases, I do not have mutation out of 200. So, it is 30 plus 100 divided by 200.

And I get 0.65, that is $P(A=N)$. So, in this way, I can actually calculate the probability that gene B is mutated, or the probability that gene B is not mutated. These two types of probability, this probabilities, probability that A is mutated, probability is of A not mutated, these are called marginal probability. Marginal probability is a probability of some outcome, where it can happen in all possible ways, all possible conditions. So, that is the marginal probability. Now, let us learn another probability.

(Refer Slide Time: 07:03)



That is joint probability. Here, what I am asking you to calculate is the probability of A is mutated, and B is also mutated, this comma here represent AND, that means, what I am asking, $P(A = Y, B=Y)$. I want to calculate this probability $P(A = Y, B=Y)$.

Let us look into the data table, gene A and gene B both mutated in these 50 cases out of 200. So, that means my probability will be 50 divided by 200. And that is 0.25. So, $P(A = Y, B=Y)$, that means both are mutated is equal to 0.25. Now, let us look into conditional probability based on the same data set.

(Refer Slide Time: 08:03)

Gene A

	Y	N
Gene B	50	30
	20	100

Y: Mutated
N: Not Mutated

Conditional Probability

$$P(A = Y | B = Y)$$

Gene A

	Y	N
Gene B	50	30
	20	100

Y: Mutated
N: Not Mutated

Conditional Probability

$$P(A = Y | B = Y) = \frac{50}{50 + 30} = 0.625$$

		Gene A	
		Y	N
Gene B	Y	50	30
	N	20	100

Y: Mutated
N: Not Mutated

Conditional Probability

$$P(A = Y|B = Y) = \frac{50}{50 + 30} \leftarrow P(A=Y, B=Y)$$

$$= \frac{50/200}{(50 + 30)/200}$$

		Gene A	
		Y	N
Gene B	Y	50	30
	N	20	100

Y: Mutated
N: Not Mutated

Conditional Probability

$$P(A = Y|B = Y) = \frac{50}{50 + 30} \leftarrow P(A=Y, B=Y)$$

$$= \frac{50/200}{(50 + 30)/200}$$

\uparrow
 $P(B=Y)$

So, I want to calculate a particular type of conditional probability, I want to calculate what is the probability $P(A = Y|B = Y)$. So, how should we calculate that, first thing I should check the number of cases where B is mutated, where it is this row 50 plus 30, there are 80 cases out of 200, where B is mutated. Out of these 80 cases.

How many cases we have where A was also mutated. This first one, see both are Y, both are mutated. So, that means this, 50 out of 80 cases, we have both mutation in A and B. So, that means $P(A = Y|B = Y) = \frac{50}{50+30}$, equal 0.625. I can ask another conditional probability. For

example, I can ask what is the probability of A is mutated, given B is not mutated, $P(A = Y|B = N)$?

But before I go into it, just rearrange these data and check what do I get. So,

$$P(A = Y|B = Y) = \frac{50}{50+30}$$

Now, I divide numerator and denominator by the sample number 200. Right I can do that, and then what do I get?

$$P(A = Y|B = Y) = \frac{50/200}{(50+30)/200}$$

What is this, what is $(50/200)$, $(50/200)$ is nothing but this one divided by total numbers 200 that means, that is the joint probability of both A and B are mutated.

So, that is probability A is mutated given B is also mutated. Sorry, it will be joint, let me, not given it is joint probability. Let me erase this. So, this is a joint probability, probability of A is mutated given also B is mutated, both are mutated simultaneously and $P(A = Y, B = Y)$. So, that is the numerator. What is this one? This one, I can check 50 plus 30 that is the 80 cases, this one so, this one is nothing but probability that B is mutated. So, that is the marginal probability of mutation in B, B is mutated. So, I can write this one.

(Refer Slide Time: 11:04)

The image shows a man in a plaid shirt pointing at a blackboard. On the blackboard, there is a contingency table for Gene A and Gene B, and a derivation of a conditional probability formula.

		Gene A	
		Y	N
Gene B	Y	50	30
	N	20	100

Y: Mutated
 N: Not Mutated

Conditional Probability

$$\begin{aligned}
 \rightarrow P(A = Y|B = Y) &= \frac{50}{50 + 30} \\
 &= \frac{50/200}{(50 + 30)/200} \\
 &= \frac{P(A = Y, B = Y)}{P(B = Y)}
 \end{aligned}$$

Gene A
Y: Mutated
N: Not Mutated

	Gene A	Y	N
Gene B	Y	50	30
	N	20	100

Conditional Probability

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

$$P(A = Y | B = N) = \frac{P(A = Y, B = N)}{P(B = N)}$$

Gene A
Y: Mutated
N: Not Mutated

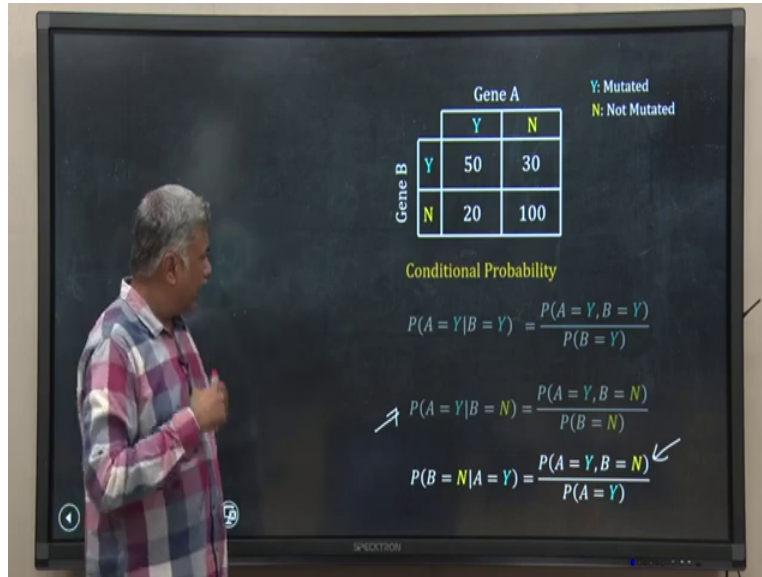
	Gene A	Y	N
Gene B	Y	50	30
	N	20	100

Conditional Probability

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

$$P(A = Y | B = N) = \frac{P(A = Y, B = N)}{P(B = N)}$$

$$P(B = N | A = Y)$$



So, what do I get, I get

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

So, as I said, I can use the same formulation to calculate other conditional probability also. For example, consider this one, $P(A = Y|B = N)$.

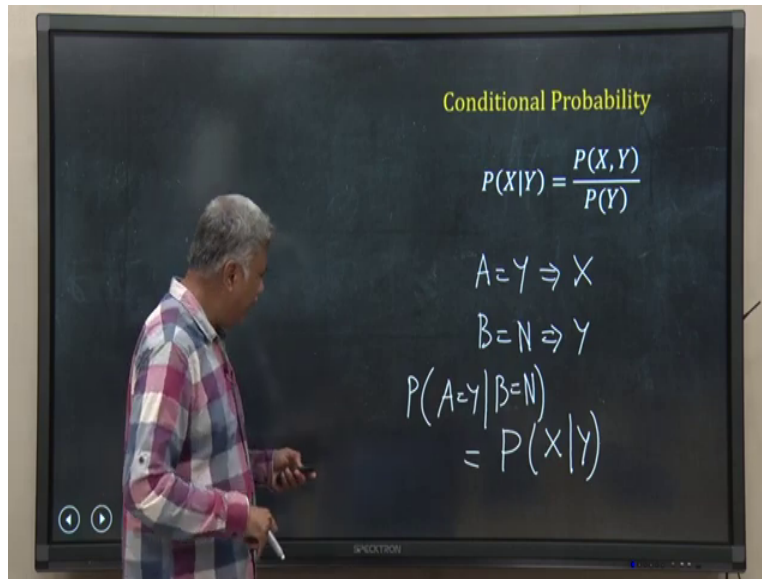
So, what will happen in that case, using the previous example, if I follow the previous example, I will write this is equal to probability that A is mutated and B is not mutated. This is a joint probability between A has mutation and B does not have mutation divided by probability that B does not have mutation, there is a marginal probability of B equal to N, B does not have mutation.

Similarly, I can also calculate the probability that B is not mutated given A is mutated, A is equal to Y, what should that be, that will be equal to probability that A is mutated and B is not mutated, divided by the marginal probability of A is mutated. Just pay attention what we are doing here, in the earlier example, I was writing probability of A is mutated given B is not mutated.

In this case, I am just calculating the reverse of that one. I am calculating the probability that B is not mutated given A is mutated. So, I can actually swap this questions, we have to remember here when we are talking about conditional probability, we are not talking about causality. Mutation of A is not causing mutation of B, we are not saying something like that.

Conditional dependence means conditional probability means there is some probabilistic observational association, I am observing both the event to happen together with certain probability. That is all, we are not saying this one causes that one, we are not saying mutation in A causes mutation or B, something like that. So, this is the way we can actually calculate conditional probability for different cases.

(Refer Slide Time: 13:26)



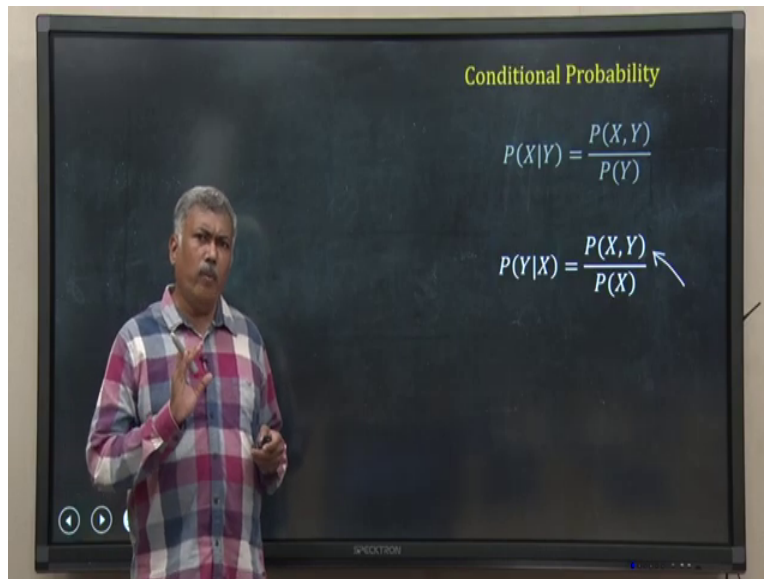
So, I can generalize this one and I can write and that is what you will see in most of the textbook, the conditional probability,

$$P(X|Y) = \frac{P(X,Y)}{P(Y)}$$

There is nothing confusing here, I can explain what is X and Y taking the earlier example, for example, in the earlier case, suppose A equal to Y, A is mutated, that may have, may be represented as X.

And suppose B is not mutated. That can be represented by Y, then probability that A is equal to Y, A is mutated given that the B is not mutated, is equal to probability, I can replace A and B these things by X and Y. Probability of X given Y. This is the way we can in a short form. That is why we will write probability of X given Y is equal to probability of X and Y, joint probability, divided by the marginal probability, probability of Y, that is all. Now, I can again swap the position and I can say.

(Refer Slide Time: 14:45)

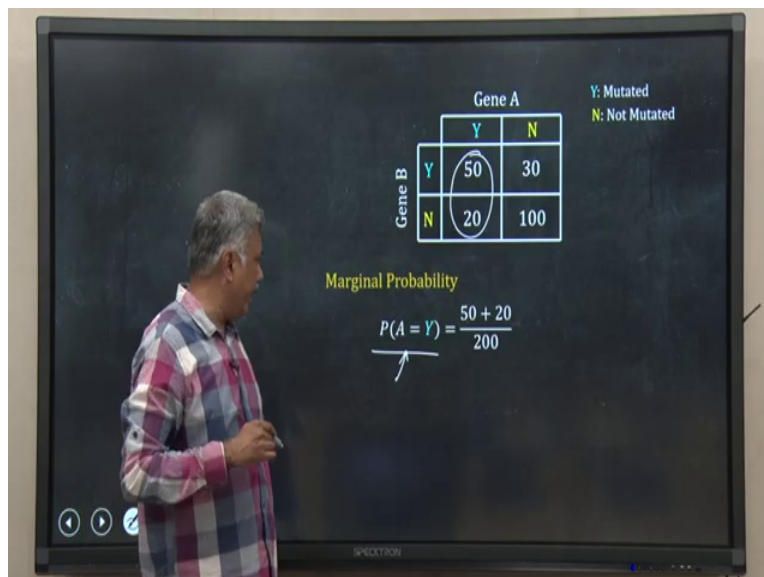


What is the probability, the conditional probability of Y given X,

$$P(Y|X) = \frac{P(X,Y)}{P(X)}$$

Now let me go back to the original mutation table and learn something new about a marginal probability.

(Refer Slide Time: 15:08)



		Gene A	
		Y	N
Gene B	Y	50	30
	N	20	100

Y: Mutated
 N: Not Mutated

Marginal Probability

$$P(A = Y) = \frac{50 + 20}{200} = \frac{50}{200} + \frac{20}{200} \leftarrow$$

		Gene A	
		Y	N
Gene B	Y	50	30
	N	20	100

Y: Mutated
 N: Not Mutated

Marginal Probability

$$P(A = Y) = \frac{50 + 20}{200} = \frac{50}{200} + \frac{20}{200}$$

$$= P(A = Y, B = Y) + P(A = Y, B = N)$$

$P(X, Y)$
 $= \frac{P(X, Y)}{P(Y)}$

		Gene A	
		Y	N
Gene B	Y	50	30
	N	20	100

Y: Mutated
N: Not Mutated

Marginal Probability

$$P(A = Y) = \frac{50 + 20}{200} = \frac{50}{200} + \frac{20}{200}$$

$$= P(A = Y, B = Y) + P(A = Y, B = N)$$

$$= P(A = Y | B = Y)P(B = Y) + P(A = Y | B = N)P(B = N)$$

		Gene A	
		Y	N
$P(X) = \frac{P(X Y)P(Y)}{200} + \frac{P(X N)P(N)}{200}$			

Y: Mutated
N: Not Mutated

Marginal Probability

$$P(A = Y) = \frac{50 + 20}{200} = \frac{50}{200} + \frac{20}{200}$$

$$= P(A = Y, B = Y) + P(A = Y, B = N)$$

$$= P(A = Y | B = Y)P(B = Y) + P(A = Y | B = N)P(B = N)$$

So, I have the same table, and I want to calculate the marginal probability of A is mutated, $P(A = Y)$. So, I can do that simply, I can check, sum this column and divide by 200. Now rearrange this term $(50 + 20)/200$. Let me rearrange the term. So, by rearranging term what I get $(50/200) + (20/200)$.

What is $(50/200)$. Let me check, this is 50. And this is divided by 200. That means, that is the joint probability that both A and B mutated, A and B equal to Y. And what is 20? This one is a joint probability, that gene A is mutated, and gene B is not mutated. So, let me write that. So, the first term $(50/200)$ is probability that A is equal to Y and B is equal to Y.

And the second term, (20/200) is probability that A equal to Y, and B equal to N, that is B is not mutated. So, this is essentially the, what I am showing this marginal probability of A equal to Y is summation of two joint probabilities. Now, just now what we have learned, we have learned in the conditional probability.

$P(X|Y)$ is equal to joint probability of X and Y divided by probability of Y. So, that means I can write the joint probability in terms of this conditional probability and this marginal probability. And that is what I will do here. So, what I will write, $P(A = Y, B = Y)$ can be written as the product of a conditional probability and a marginal probability, what is that? $P(A = Y|B = Y)$ into $P(B = Y)$.

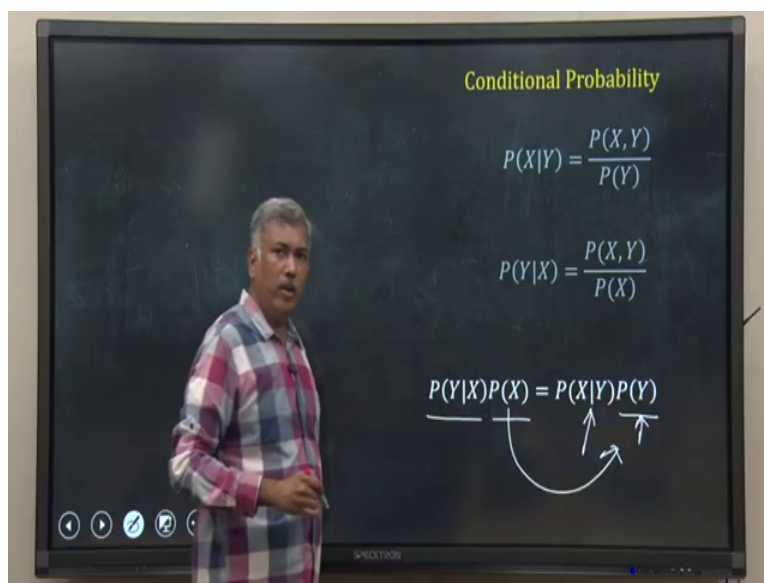
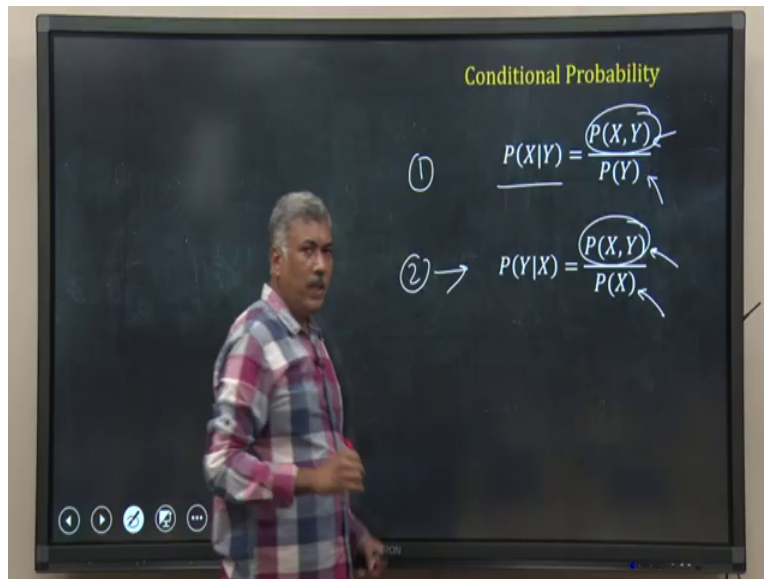
So, it is a multiplication, the product of a marginal probability and a conditional probability. Similarly, take the second term, probability, the joint probability of A equal to Y and B equal to N not mutated, what I will get, I will get probability that A is Y given B is equal to N into probability of B equal to N.

So, what I have got, I have got some generalized relation. If you tell me probability of X, that is the marginal probability of X, then that should be,

$$P(X) = P(X|Y) P(Y) + P(X| \neg Y) P(\neg Y)$$

This symbol, this means NOT Y. So, probability of X given not Y into probability of not Y. So, I am summing these two things and getting the probability of X. This is a generalized statement. Now, I will enter into the Bayes theorem.

(Refer Slide Time: 19:01)



So, let me look into the basic conditional probability.

$$P(X|Y) = P(X,Y) / P(Y)$$

The second equation that I can write, this one, is,

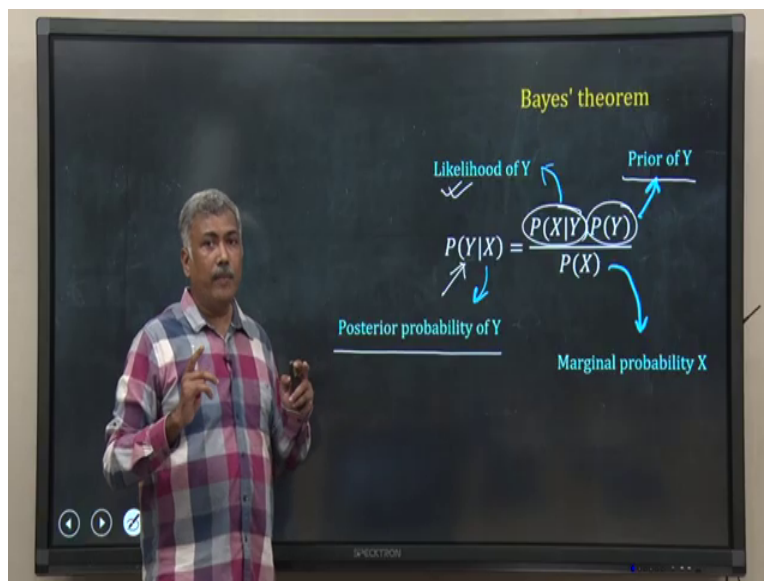
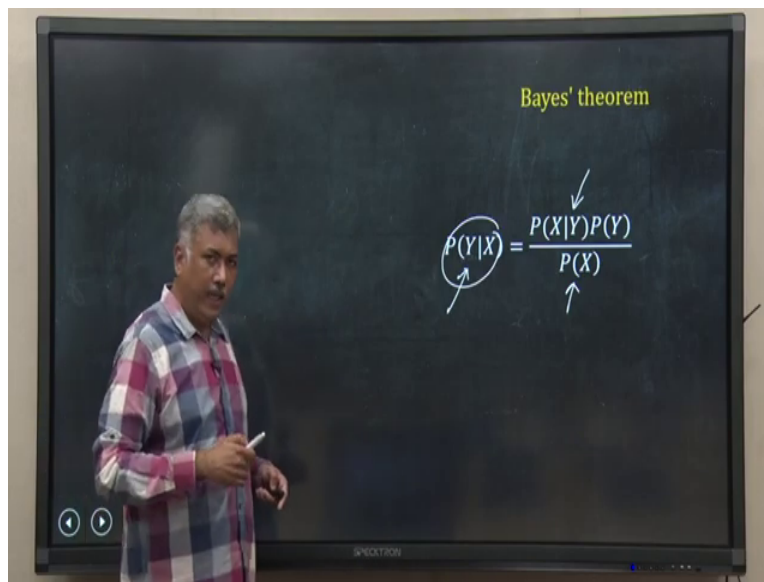
$$P(Y|X) = P(X,Y) / P(X)$$

So, both these equation 1 and 2 has one thing common, P(X, Y), P of X and Y. So, now I can actually rearrange term and equate these two equation and then what do I get? I get something like this.

$$P(Y|X) = P(X|Y) * P(Y) / P(X)$$

Its simple rearrangement, you can try it. Now, what I do, I take this probability of X on this side, it will go in the denominator.

(Refer Slide Time: 20:08)



And I get the Bayes theorem. What does it says? It says the conditional probability,

$$P(Y|X) = P(X|Y) * P(Y) / P(X)$$

This relationship we get, this is the Bayes theorem, and we will use this one to answer the initial question, the initial numerical problem with which we started today's lecture.

Now, before I go into it, just to remind you that this term each of these term, like this one, this one, all has certain terminologies that we have some, have some name, for the technical name for them. And they are like this, for example, this one probability, conditional probability, $P(Y|X)$ is called posterior probability of Y whereas, $P(Y)$ here in this Bayes equation will be called Prior of Y. Whereas, this, the first term, $P(X|Y)$ will be called likelihood of Y. Now, you may wonder why they have such names, I will explain them in time. So, let us go back now, and try to solve the problem with which we started our lecture today, the diagnostic kit problem.

(Refer Slide Time: 21:27)

We have developed a new rapid diagnostic test for a disease.

Sensitivity (True Positive Rate) of the test = 99%
Specificity (True Negative rate) of the test = 98%

Prevalence of the disease = 0.1%

What is the chance that a person has the disease if he/she tests positive?

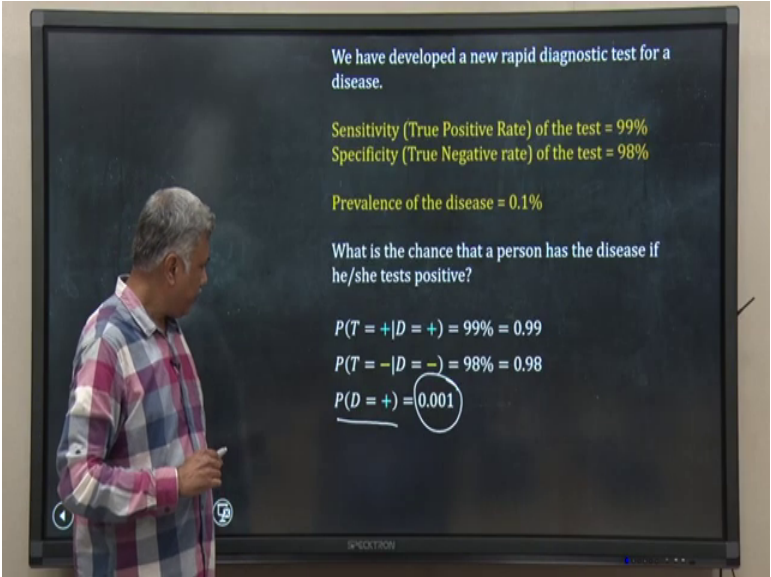

$P(T = + | D = +) = 99\% = 0.99$

We have developed a new rapid diagnostic test for a disease.

Sensitivity (True Positive Rate) of the test = 99%
Specificity (True Negative rate) of the test = 98%

Prevalence of the disease = 0.1%

What is the chance that a person has the disease if he/she tests positive?

$$P(T = + | D = +) = 99\% = 0.99$$
$$P(T = - | D = -) = 98\% = 0.98$$


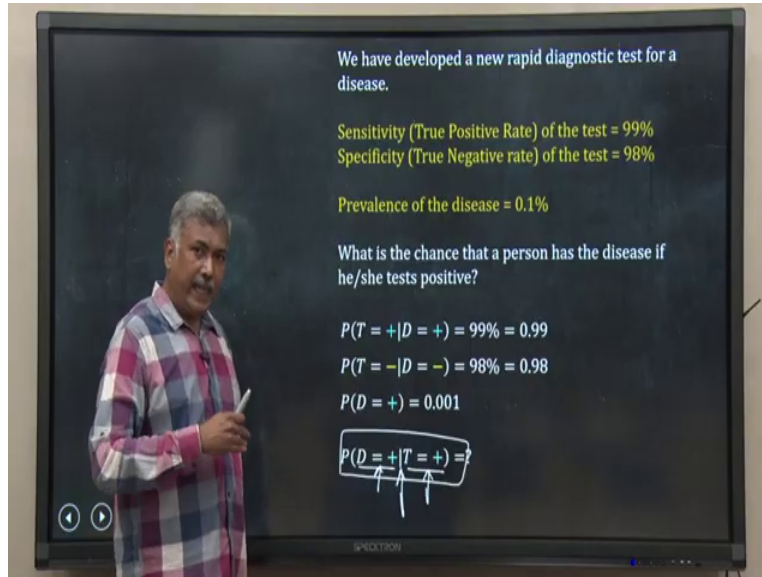
We have developed a new rapid diagnostic test for a disease.

Sensitivity (True Positive Rate) of the test = 99%
Specificity (True Negative rate) of the test = 98%

Prevalence of the disease = 0.1%

What is the chance that a person has the disease if he/she tests positive?

$$P(T = + | D = +) = 99\% = 0.99$$
$$P(T = - | D = -) = 98\% = 0.98$$
$$P(D = +) = 0.001$$



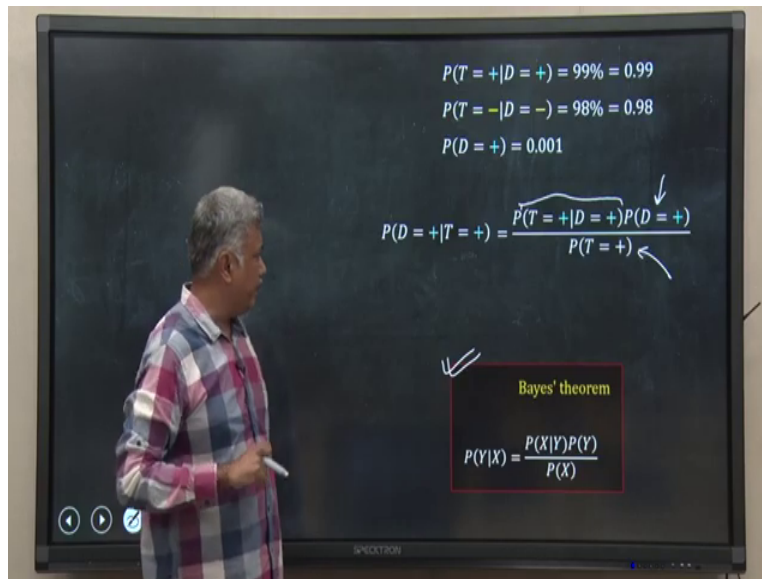
So, this is the problem again, the sensitivity is 99 percent, specificity is 98 percent, prevalence is 0.1 percent. I have to calculate if I enter in the clinic and do this diagnostic test on me. And if I come out positive, then what is the probability that I have the disease. This data is given in a word format. So, I will convert them in symbolic format.

So, what I can write, the sensitivity, sensitivity is what, true positive rate that means, if somebody has disease and the test has been used on that person and test result also come positive. So, the probability that T is positive given the person has disease, D positive, positive means we are getting positive result. So, probability that T is positive given D is also positive is 99 percent. So, that means this is 0.99.

Let us look into the specificity, specificity is true negative rate that means, a person who is negative given, that person is also testing negative in my test. So, this is a probability, the conditional probability that test is negative, test report is negative given that the person does not have the disease D is negative. So, that is 98 percent and I can write that 0.98.

How should I represent the prevalence, prevalence is the probability of having disease D equal to plus and that is 1 percent. So, that means, 0.001 quite a low prevalence. Now, I will use this data to calculate this question's answer. So, what I have to calculate, I have to calculate the conditional probability of having disease, disease positive given the test result has come positive. I have done the test on me, the test result has come positive, I want you to tell me the probability that I have the disease, $P(D = +|T = +)$ that I have to calculate.

(Refer Slide Time: 23:48)



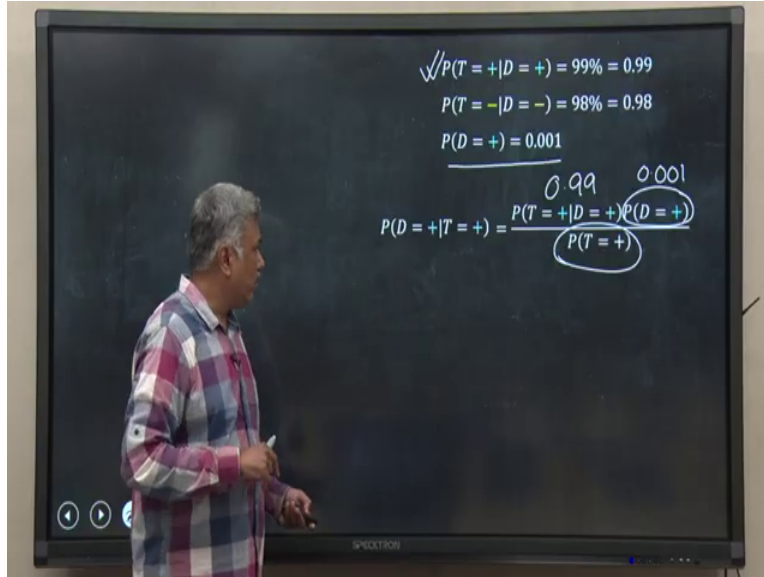
So, how should I do that, to do this calculation I use the Bayes theorem and just here I have shown the Bayes theorem that I will use. So, what I can consider, I can consider Y. I can write Y is nothing but D positive. Whereas, I can consider X, this X is test is equal to positive. So, now I can use this Bayes theorem to write the conditional probability $P(D = + | T = +)$.

So, what will I get?

$$P(D = + | T = +) = \frac{P(T = +, D = +) * P(D = +)}{P(T = +)}$$

Just look into the Bayes theorem. And see that we are writing the same thing actually, but not in terms of X and Y. But in terms of D positive T positive, that is what we are doing. Otherwise it is Bayes theorem. So, now I have to put a numerical value.


(Refer Slide Time: 25:08)




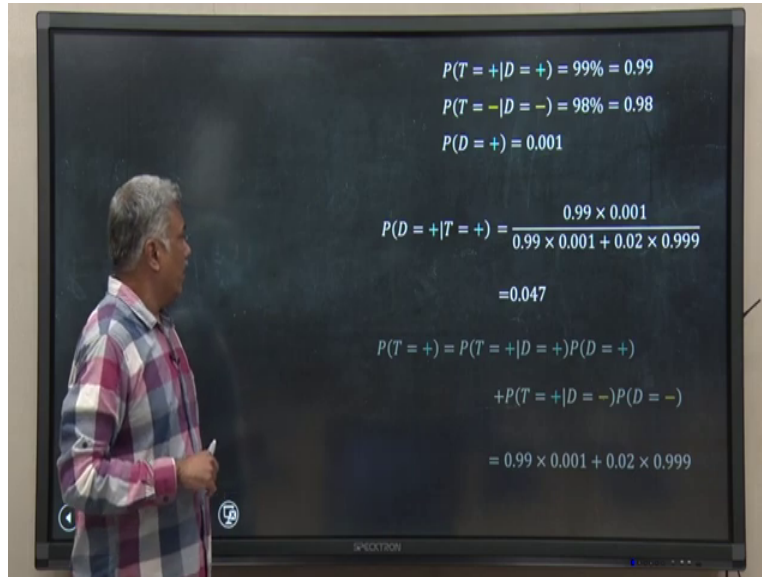
So, let me start with numerical values. So, D is positive, $P(D = +)$ is given here 0.001, let me write down good. $P(T = + | D = +)$ that is also given here. That is, given here, 0.99. So, that is 0.99. But how should I get this denominator, $P(T = +)$.

Now, remember, test can be positive for two types of people, when we were assaying and testing the diagnostic kit, we have taken lots of people, some people who are really positive in disease, they may have come positive. At the same time, some of the people who are negative in disease, may also have come positive. So, test positive can happen for disease positive people also, disease negative people also. So, I have to calculate the whole probability. So, I have to use the relation between marginal probability and conditional probability that we have discussed earlier. So, I will write it in this way.

(Refer Slide Time: 26:20)

$$\begin{aligned}P(T = +|D = +) &= 99\% = 0.99 \\P(T = -|D = -) &= 98\% = 0.98 \\P(D = +) &= 0.001\end{aligned}$$
$$P(D = +|T = +) = \frac{P(T = +|D = +)P(D = +)}{P(T = +)}$$
$$P(T = +) = P(T = +|D = +)P(D = +) + P(T = +|D = -)P(D = -)$$


$$\begin{aligned}P(T = +|D = +) &= 99\% = 0.99 \\P(T = -|D = -) &= 98\% = 0.98 \\P(D = +) &= 0.001\end{aligned}$$
$$P(D = +|T = +) = \frac{P(T = +|D = +)P(D = +)}{P(T = +)}$$
$$P(T = +) = P(T = +|D = +)P(D = +) + P(T = +|D = -)P(D = -)$$
$$= 0.99 \times 0.001 + 0.02 \times 0.999$$




$$P(T = +) = P(T = +|D = +)P(D = +) + P(T = +|D = -)P(D = -)$$

So, this whole summation gives me the total probability of having test positive.

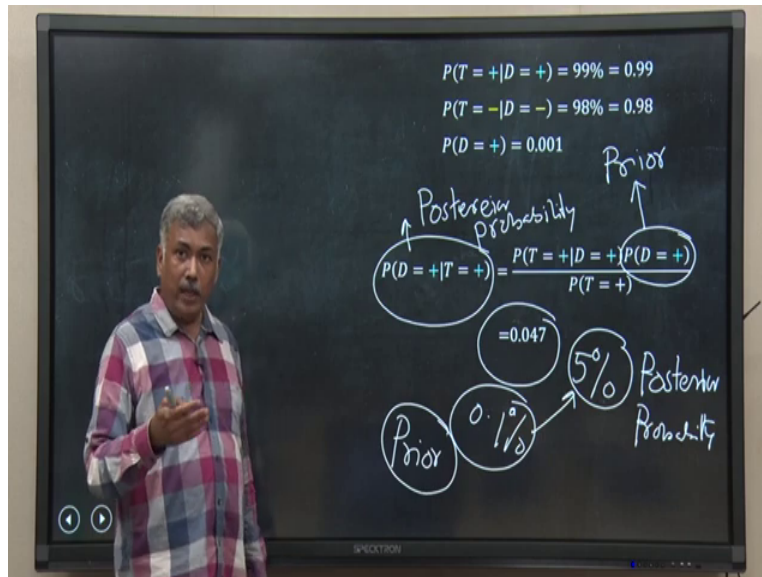
So, I will put that value here, the denominator, and I have done the calculation, you can do the calculation yourself also, D negative, you can get because the probability of D positive, and then you can calculate the D negative, you can also calculate this one. So, now, if I put these values, these values this whole thing in the denominator here, and do the calculation, the calculation looks like this.

And numerically, it gives me 0.047. That means 5% chance, and that looks quite paradoxical isn't it, that we have a very high accuracy, specificity, sensitivity are almost 99%, 98%. But if I test a random person without any symptom, enter into the clinic and use this test and it comes positive, the mathematics is telling me I have actually 5% chance of having the disease.

This looks paradoxical. And sometimes it is quite unbelievable for most of us, but that is, it is true, the calculation is correct, we have to remember that disease, as well as test all are probabilistic. And as the prevalence of the disease becomes very low and low, even though I have a very high accurate test, that test may not tell me, not confirm me that 100% I have the disease. I will always have a probabilistic result.

And I will show you, give you an example. Rather, I will give you a home test to calculate how this prevalence of the disease can affect our calculation. I will come to that. So, here in this case, I have a 5% chance that I will have the disease. And obviously I have to go for some other tests, other symptom things and all these things the physician will check before he or she decide that whether I really have the disease or not.

(Refer Slide Time: 28:43)



So, now look into the same result in a bit different way to understand the meaning of those terms. Remember, we said in a Bayesian theorem, we have those terms, Posterior probability, Prior probability, Likelihood and all these things, let us try to understand what are those? So, I have written down the same thing that just what we are calculated.

So, now try to think over it. I am talking to you and we know this disease happens and the prevalence is 0.0%. You have no symptom. You have not done any test, about these diagnostic tests for the disease. And if I ask you, what is the probability, what is the chance that you have this disease?

Let us think about it just for a few seconds, think about it. What is the chance? You have no symptoms, you have not done the test, you know the prevalence is 0.0%. That is what information you know from the newspaper. And if I ask you what is the probability or chance that you have

the disease, your common sense will obviously say, I know only from newspaper that it has prevalence of 0.1%.

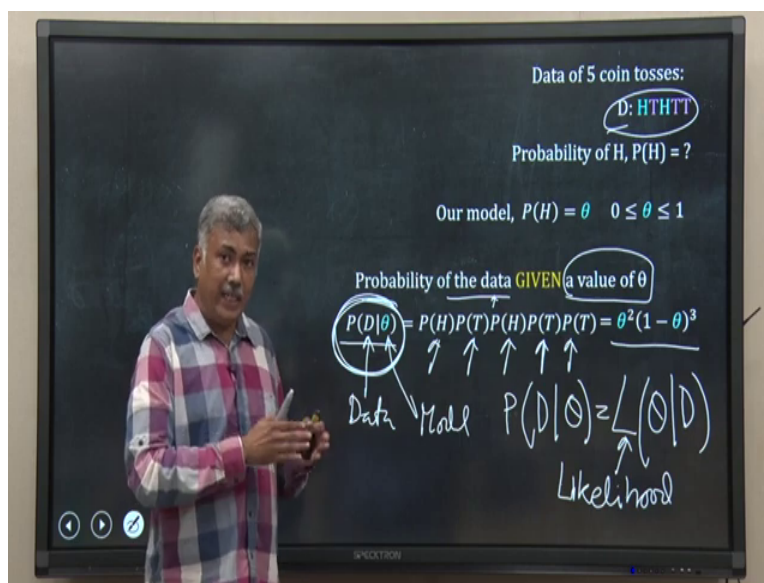
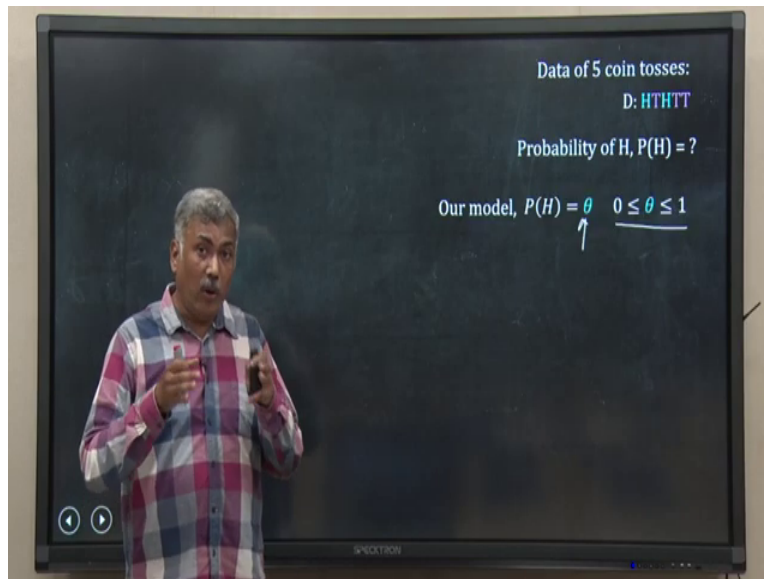
So, I have the chance of 0.001, 0.1% chance I have, that is right. So, this is your prior belief, before you have symptoms or test. This is your belief so, this is your prior belief. So, now you go to do the test and you came positive in the test. So, now you use your result and the Bayesian calculation and the Bayesian calculation now say that your chance of having the disease is 5% , earlier you thought it is 0.1%.

Now, you think it is 5%, just because you have done a test and positive result has come. So, this is your posterior, meaning now you will believe so, it is a posterior probability. And this is your prior, before the test this is what is your belief. So, that is why this term here, probability of disease positive, the prevalence is actually our prior.

And now, when you have done the test, and the probability of you having disease is my posterior probably. So, this is not just true for this diagnostic test, if you are building models to fit a data or something, you may have some prior belief how good is the model, you may have multiple model and you may have a judgment, you may have a belief that, the third model will fit the data better, it explains the data better.

Now you do some fitting, and then you calculate the probability, the correctness of your model. So, that is your posterior probability. And when I am talking on models, beyond this diagnostic test only, and they are where the likelihood concept becomes very clear. So, let me explain that with a simple model.

(Refer Slide Time: 31:52)



So, suppose I have a coin. I do not know whether it is fair or not, I do not know what is the probability of head and probability of tail. I tossed it, I tossed it 5 times. And the result I have got is HTHTT, 2 head 3 tails. So, that is my data. That is why I have written is as D. Now, I want to know what is the probability of having head for this coin?

I cannot say 0.5, how do I know it is 0.5. Although blindly many times we believe the coins are fair, but in this case, I do not know. So, what I will do, I will create a model, I will assume a

model that means, I assume a value for $P(H)$, probability of getting head and suppose that value is θ . And θ can vary from 1 to 0.

To start with, you can assume $\theta = 0.8$ or something like that. So, that is your model, θ is your model. So, considering this model is true, I calculate what is the probability of the data? What is data? Data is this observation, probability of the data given the value of θ that I have assumed, a value of θ .

So, that will be probability of D given θ , D is the data, how I will calculate it, each of this coin toss are independent. So, that means what I will get, I will get,

$$P(D|\theta) = P(H)*P(T)*P(H)*P(T)*P(T)$$

You multiply all of them and as probability of head is θ , probability of tail is $(1 - \theta)$.

So, then it is $\theta^2 * (1 - \theta)^3$. Now, if you have assumed $\theta = 0.8$, you can plug the value and can calculate the probability. Now, you change your belief, you change your model, you say no, I believe now that θ is 0.4. So, you calculate again the same conditional probability, probability of data given the model.

Now, in this way, you can keep on changing your model and calculate what is $P(D|\theta)$, data given the model, this is data, this is model. And then you find a particular value of θ which gives you the highest probability then you say, I now have got the highest probability. So, I believe possibly, this is the right model, that is a right value of θ .

So, what are you testing here? What you are testing, although you are calculating probability of data given θ , you are actually testing the likelihood, the correctness of your model, which model is most likely. So, that is why probability of data given model is called the likelihood of the model given the data. We use a curly L to represent likelihood. So, that represents Likelihood. So, this probability, probability, conditional probability of D given θ is actually likelihood of θ , likelihood a model given the data. So, now let me move further.

(Refer Slide Time: 35:11)

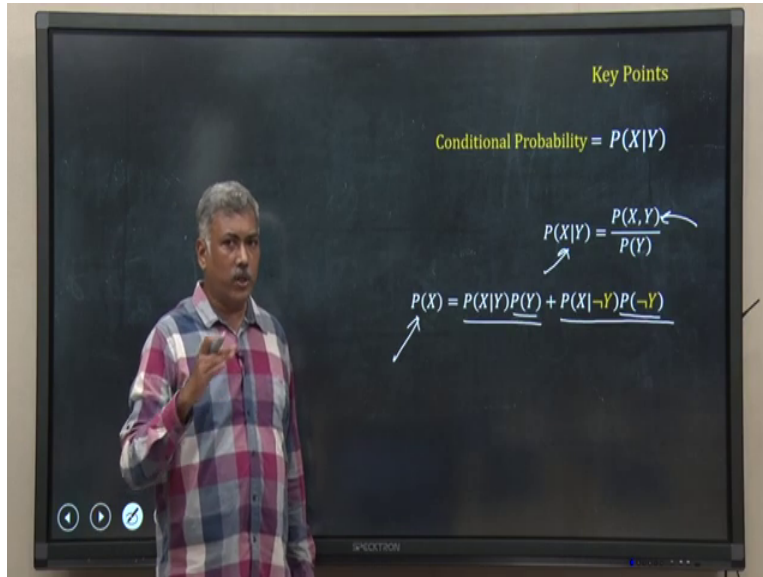
Data of 5 coin tosses:
D: HTHTT
Probability of H, $P(H) = ?$
Our model, $P(H) = \theta \quad 0 \leq \theta \leq 1$
Probability of the data GIVEN a value of θ
 $P(D|\theta) = P(H)P(T)P(H)P(T)P(T) = \theta^2(1-\theta)^3$
$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}$$

Actually, you can go beyond this likelihood only, you can use the Bayes theorem again, to actually also calculate the posterior probability of your model. If you have multiple models, you can assume their probability distribution also. And you can do the calculation, I will not go into that just to inform you. Actually, you can use Bayes theorem in totality to actually choose model or infer model from your data.

(Refer Slide Time: 35:34)

Key Points
Conditional Probability = $P(X|Y)$
$$P(X|Y) = \frac{P(X, Y)}{P(Y)}$$

$$P(X) = P(X|Y)P(Y) + P(X|\neg Y)P(\neg Y)$$



So, now, let me jot down what we have learned in this lecture. The first thing we have learned, we have learned about conditional probability, $P(X|Y)$. And now we have expanded this definition of conditional probability. And what we have learned is that the conditional probability

$$P(X|Y) = P(X, Y) / P(Y).$$

Similarly, you have the marginal probability;

$$P(X) = P(X|Y) P(Y) + P(X|-Y) P(-Y)$$

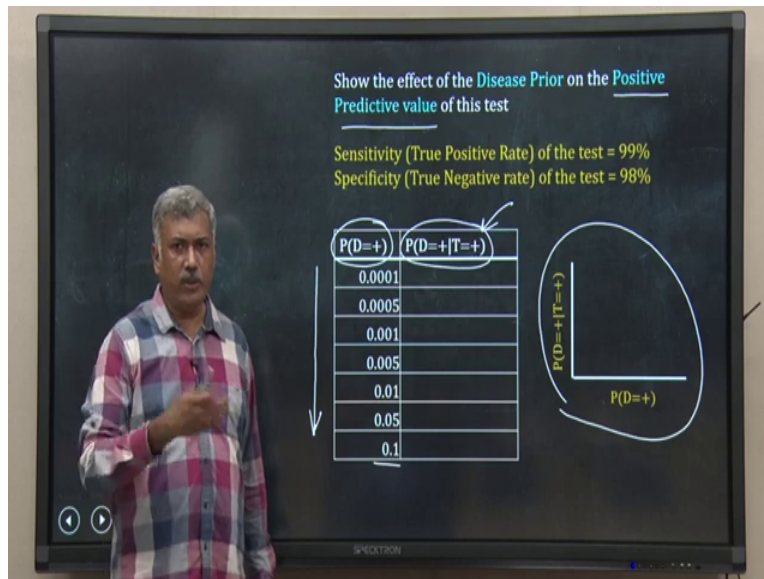
These two rules, these two relationships are very useful to do lots of conditional probability calculation.

Based on this idea of conditional probability, we reach the Bayes theorem. What Bayes theorem says, the conditional probability,

$$P(Y|X) = P(X|Y) * P(Y) / P(X)$$

So, this one, $P(Y|X)$ is called the posterior probability. Whereas the denominator $P(X)$ is the marginal probability. $P(Y)$ is the prior probability and $P(X|Y)$ is the likelihood of Y. That is all for this lecture. Before I leave you, I will leave you with a problem.

(Refer Slide Time: 37:21)



As I said, many times it is unbelievable to know that a assay which is so accurate having high specificity and sensitivity, when I use the Bayes theorem, and I calculate the probability of a person who has tested positive whether he has a disease or not becomes very low, that probability is very low. For example, in our example, it was only 5%.

So, actually this calculation, the probability that you are disease positive given you have tested positive, what we call positive predictive value of a test, we call it positive predictive value of a test and it is represented by this one, probability of D is positive given T is positive, actually depends upon the prevalence of the disease.

Now, remember, many a times prevalence of the disease may have been calculated, many a times we may have guessed, if you are new disease spreading across the globe, then you may not have adequate data, you have small sample size from here and there. And from that, you may have estimated or made a educated guess of the prevalence, what is the prevalence?

Prevalence is probability of disease equal to positive. So, what I want you to test at home is that take multiple value for this probability, this prevalence, probability of D equal to positive, this is positive starting from 0.0001 to 0.1, all these values. And then you use the same thing, same sensitivity and specificity that we use for our calculation and calculate this one.

Probability of having disease given test is positive, which equal positive predictive value, and then plot it here. You can do it manually and then make a plot in Excel or any other software if you have access to and draw the plot. That is your home task. Please try that. This will clear lots of idea about Bayes theorem, and it will make you more confident to use Bayes theorem in other problems. That is all for this lecture. Happy Learning.