Biomathematics.
Dr. Ranjith Padinhateeri
Department of Biotechnology
Indian Institute of Technology, Bombay

Lecture No. # 22
Probability Distribution

Hello, welcome to this lecture on biomathematics. In the last lecture, we have seen in the last lecture, we started discussing statistics; how statistics can be applied in biology and we took a simple example of a few examples like traveling to a college and number of students, number of marks for an exam and all that and we discussed the idea of average and standard deviation.

Now how do we, so the basic question we had is that we get a lot of data in from experiments, large amount of data and given this huge data, set of numbers data is essentially is a set of numbers. How do we make sense of these numbers? What we can learn from this numbers? We learn that average and standard deviation is some two simple things that we can learn and that has a meaning and that are two numbers from huge data set; we can extract two numbers, meaningful numbers: average and standard deviation.

What more can we do? That is, what we are going to discuss in this lecture? What more we can extract from the data? Or, how do we present the data so that we get, we have much more information, we can convey much more meaningful information? So, today again we have statistics as our main title. In statistics we will have, so this is the slide, in statistics we will have, we will discuss in specifically probability distribution.

So, we will go ahead and see what is probability. What do you mean by probability distribution? As we said, we have various experiments. One of the experiment is traveling to the college but, we will discuss a new experiment, a simple experiment that you all can do yourself; we do not need a big set up or anything; just you can do it in your class. Let us do a simple experiment of measuring height so the experiment is, measure height.

(Refer Slide Time: 03:03)

So, measure height of boys in your class in the simple experiment that anybody can do. You can measure height of all boys, let us say and/or girls. Here, we take example of boys because some numbers that we used more suitable here for this example. So, if we do this experiment let us say, you just go on and measure now all the heights of each and every student in your class and what do we get when we measure? We will be getting a big data set. Let us look at some represent something; what you will typically get? You will typically get a set of numbers.
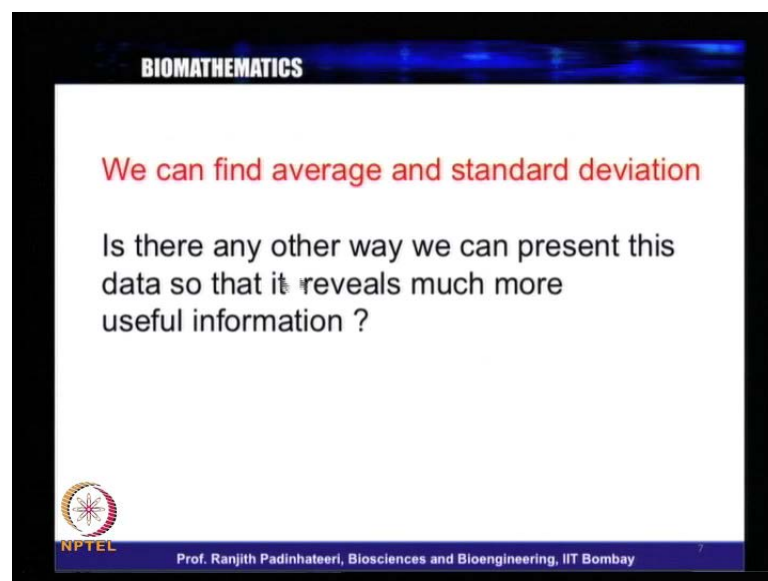
(Refer Slide Time: 03:55)

Let us see here is what the let us say in your class if you just look at the numbers. So you can have 150 centimeter, 171 centimeter, 165 centimeter, 140 centimeter; some numbers these are some kind of reasonable numbers you know that it will be like so the many of them are 150, 140,very few 180, is just 180. There is nothing bigger than 180 you see there this one person is 180 centimeters and like this is 1 person with like 131 centimeter very small short person and very tall person. And, others somewhere in between so you have like large number of such data set such numbers so this is not ending here so this just a continuous list.

So, let us say you do this measurement and how do we make sense of this data? We will have this huge data here and how do we make sense of this data? So, let us say we have 100 we measure heights of 100 such students as a we have 100 such numbers 150, 160, 130, 141, 120, 2 1 3 5 100 numbers; how do we make sense of this data?

(Refer Slide Time: 05:25)

Two things we learnt or that we can find the average and standard deviation. We can say that 0 average is 155 centimeters or 150 centimeters. That will give you some idea and we can say as standard deviation is plus or minus 20. So, standard deviation is 20; so we can say that average is 150 and the standard deviation is 20 that means 150 plus or minus 20. This is something which we can say if you wish like using the idea that we learned in the last class the last lecture. But, the question is there any other way we can present this data so that it reveals much more useful information more than average in standard

deviation can give? Can we present this data instead of presenting all these many numbers here? Can we present this data in a different way such that this reveals much more information by just looking the way we present it?

So, we just write down these numbers is very bad idea; it is a like a very stupid idea in some sense or it is not like a very, it is not a great idea. It does not make much of a sense using lot of numbers that is what it means like many large, just many numbers it does not make much sense but, if we present it in a particular way. Is there a smart way of presenting this such that it makes much more sense by just seeing that you should be able to make out lot more information than either average or standard deviation? Or, just by one glance of these numbers does not make much sense so just more standard deviation and average.

We need lot much more information and how do we present this as that how do we present this numbers in one slide. Such that it reveals a lot of information so the answer to so this is the question is there a way is there any other way we can present this data.

(Refer Slide Time: 07:51)

So that it reveals much more useful information and the answer is distribution you can, we will discuss what is distribution means so but, the just the meaning of the word would tell you. We are taking about heights so, we want to present the distribution of heights. So how do we present the distribution of heights so let us look at this is one way?

(Refer Slide Time: 08:26)

We can say that take present in different range. So, let us say we can break down the data into ranges. How many students are there having a height and the range of 130 to 140 centimeters so this height ranges in centimeters? If the height ranges between 130 and 140 centimeters, how many students are there? Just one student. How many students are there having height between 140 centimeter and 140 centimeter and 50 centimeters?

Let us say there are a 17 students; this is why this call a number of students having height in this range so 150 and 160. How many students are there such that the heights they fall in between 150 centimeter and 160 centimeter there are 35 students?

There are 30 students having height in between 160 and 170 and there are 15 students having height in between 170 and 180 and there are 2 students having height in this range between 180 and 190. So, large number of students in the middle and a few students with very short few students and very long few students; so very tall students only two very short students only one student like very short, 2 tall students and all are somewhere in the middle. So, this is the typical data that you expect and this is what you have so this is actually a smart way of presenting the date because this gives you some idea but, if this is just a table and if you plot this table in a graph. This we always like you always like to present the data as a graph. That can make much more sense; so one way of presenting such a things are called histograms. So, when you have data this kind

of this kind of data they and if you plot it they appear as if like a histogram; so what is the histograms?

(Refer Slide Time: 10:54)
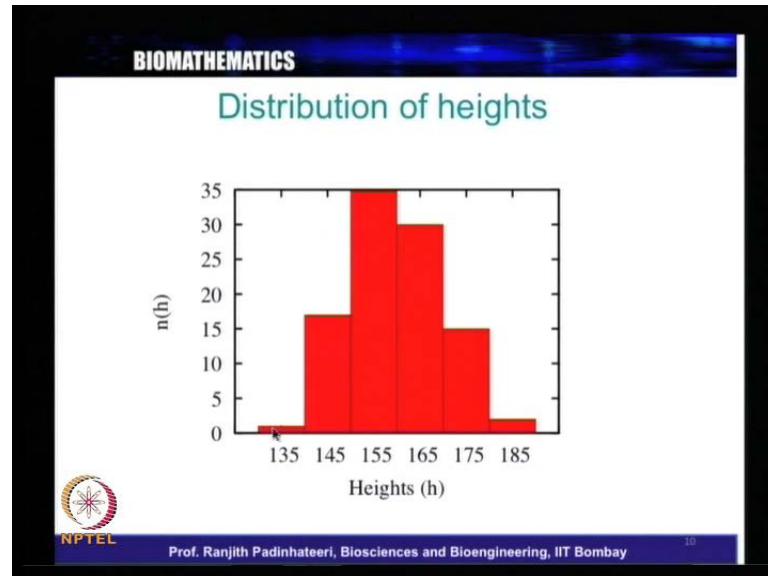
So see this, let us say between 130 and 140, this is what I mean 135; I just put some number in between here there is one student with 135 height between 130 and 140. So h is the heights and n of h is the number of students having height h. So what is the graph distribution of heights so n of h? You can call this distribution. So this essentially the same thing that we saw like if we can draw in this in a different way if you wish.

(Refer Slide Time: 11:28)
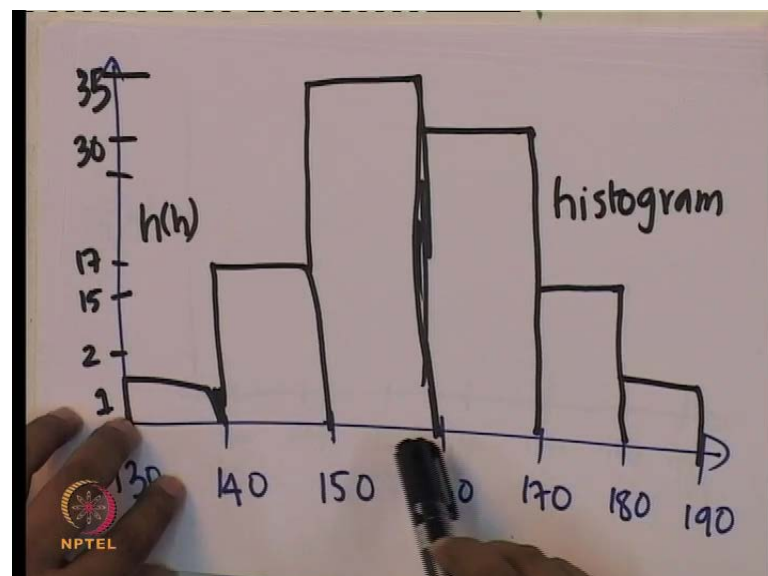
We can have drawn this x and n of h; so how many students are there between 130 and 140? So is 150 is 160 there is 170, 180 is 190; so these are the numbers we have 130 so let me write this is 130; let me write it little more clearly. So, this little more clear for you. Let me draw this little more clearly so you have let us say this is 130, 140, 150, 160, 170, 180, 190. So, how many students in this table? If you they have one student between 130 and 140.

So between 130 and 140 we have just one student so we mark it one so between 130 just one student between 140 and 150 there was 17 students so between 140 and 150 plus 17 students let us mark this as 170. So this is 17 between 150 and 160 how many students so this tables tell you between 150 and 160 there are 35 students. So between 150 and 1 sixty there are 30 5 students so let me mark this as 35 this is 35; so there are 35 students in this range. Between 160 and 170 there are 30 students. this is what this table says - 30 students in this range.

So 160 and 170, 30, so this is let us say this is 30 somewhere here. This is 30; so me where somewhere here 30 so roughly we can mark this way 30 students. And, 170 and 180 in this range there are 15 students; so this is 17 so this is 15; so 15 students in this range between 170 and 180; 15 students and 180 and 190 only two students.

So this is one and two is here; so this is like two students; so this kind of a block diagram if you wish so if you in a in a block again a manner something like a block. Like, so it is a block of 130, 140 another block of 140, 150 another block of 150, 160. So, another block here another block here another block here so such this kind of a plot is called is histogram if you wish. So this is called histogram so this is distribution so this is height in a particular range and this is number of students having that height h. So, this is what essentially plotted here in the in this exactly what we just drew is plotted here so there are between 130 and 140 I just mark by 135.

There are 1 and there is this 30, 35, this is 17 it is just above 15 this is 17 35 and this is 30. So, again this is 15 and this 2 so this is a histogram and we can call distribution. But, is this an accurate representation? In some senses it is not very accurate, why? Because, there are students with 130, 132, 133, you all put in one range so we do not actually distinguish between students having 139 height and 131 height one height 131 centimeter and 139 centimeter.

They are all in one range like let us look at here if a student having 150 one and 150 nine we put them in same range so we can actually reduce the range if your wish so instead of writing instead of writing 130 140 I if I wish I could write how many students between 130 and 132.

(Refer Slide Time: 16:48)
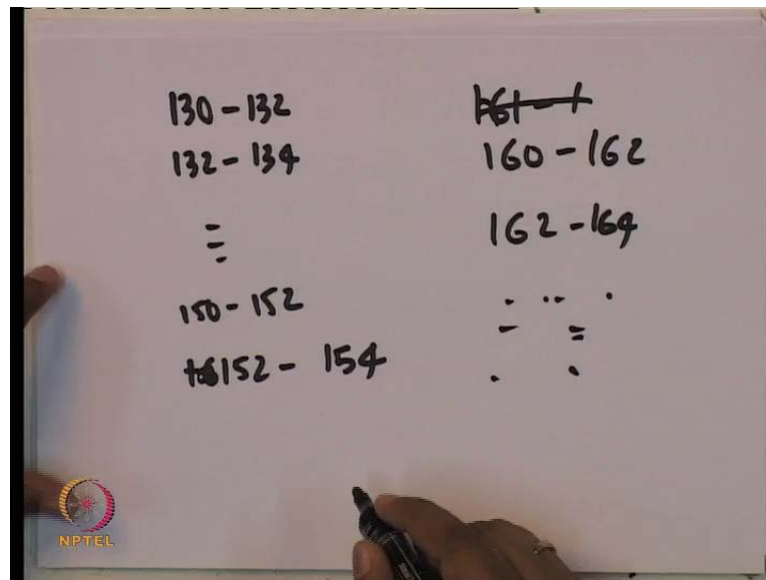
So, we can range make this in 2 centimeter and let us say 1560 and 152 160 similarly, 152 154 so like 161 sorry 150, 160, 162, 164 so like this you can write you can in take a interval of two so then you can you can make the intervals smaller and smaller. Then that will be a better description of the data.

(Refer Slide Time: 17:40)
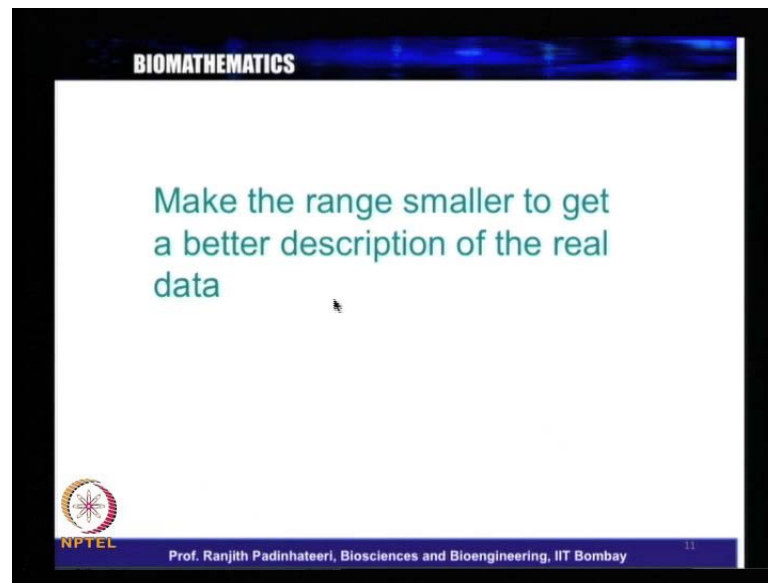
Essentially, you can make the range smaller to get a better description of the real data. Now, how small it should be how small the range should be let us say that when you measure this height. You do not have lecture in this; here tape which you measure can only measure this in a same centimeters you cannot measure less than a centimeter.

Let us say you cannot measure millimeter, you can only measure in centimeters then you can present the data in the range of centimeter if your wish. So, then this is nothing better that there is a best description of the data. So, then you can ask what is a question how many students having height 131 centimeter. How many students having height 132 centimeter? How many students having height 34 centimeter? How many students having height 151 centimeter? Each centimeter by centimeter you can have this data present it.

(Refer Slide Time: 19:01)
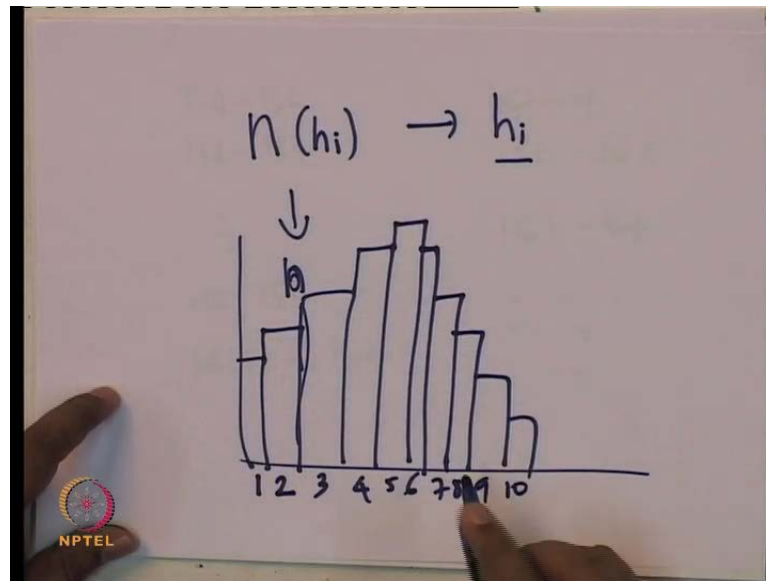
So let us say <mark>let us say</mark> you have the data presented such a way that n of h I is the number of students having height h I h I a height could be like 30 one so this height could be 131 150 one one sixty two one any number in short you can call this now let us say you plot this as a distribution now if you plot this as a distribution you will have many histograms let us say you have a histogram.

Let us say you have a histogram like this so I call this 1, 2, 3, 4, 5, 6, 7, 8, 9, 10; so what are these are ranges you took 10 ranges. So just like here in the previous example we had how many ranges? we had like one range, two second range, third range, fourth range, fifth range, sixth range, so we had 6 histograms – 1, 2, 3, 4, 5, 6. Similarly, let us say if your presenting the data height of students between 130 and 180 centimeter by centimeter; so you can have some so you can have 130 here 180 here.

(Refer Slide Time: 20:33)
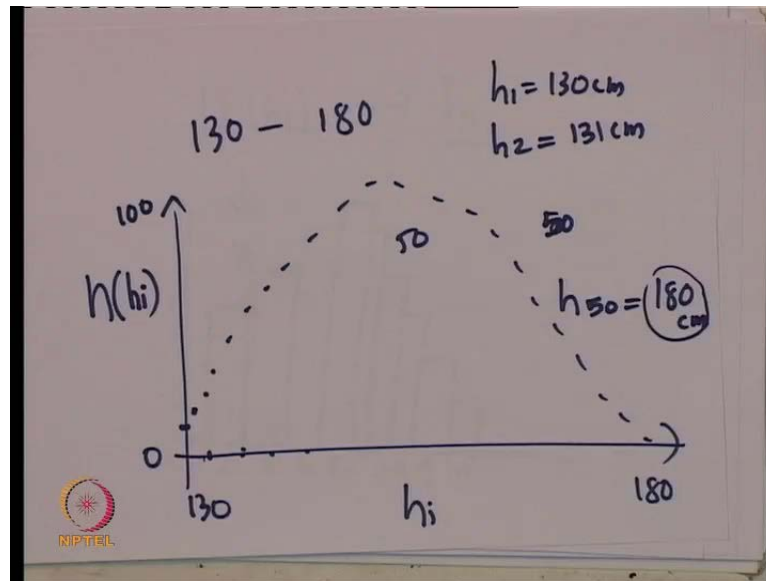
So, how many students having 130? May be one student; how many students 131? How many students 132? So you can have for each centimeter so you can have like 50 points so you can have 50 points like this for each centimeter how many students having so this is h i. So h 1 h 2 h 3 h 4 h 5; so h 1 is 130 centimeter, h 2 is 131 centimeter similarly, h 50 is 180 centimeter, so you can write this way; basically you can have h I versus n of h I graph which might look some which might have some particular shape.

So then let us say it has this particular shape n of h I versus h I if you take for centimeter by centimeter that means you ask the question, how many students having 130 centimeter that is this is number. So, this could go from 0 to some particular number let us say 100.

How many students having 132 centimeter how many students having 130 four centimeter how many students having 180 centimeter so you ask this question so you have two things you have
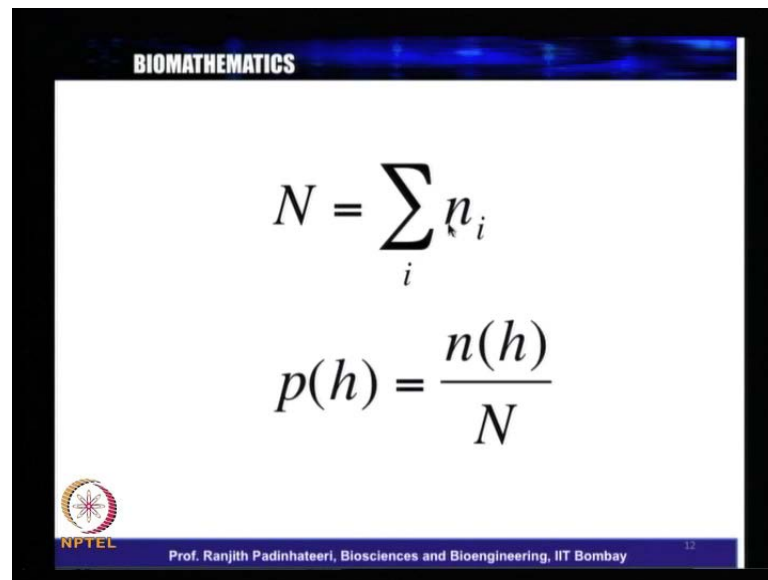
(Refer Slide Time: 22:36)

n of h I number of students having height h I now if you do this sum over I n of h i.

What is this mean if I write sum over I is equal to one two 50 there are 50 different heights why does this mean why does why does this imply.

So let us think about it so let us expand this when we expand this sum over so this is by n of h one plus n of h two plus n of h three plus dot dot dot n of h 50 so let us say h one is 130 centimeter so this is number of students I month 30 one centimeter number of students having height 132 centimeter number of students having height 130 three centimeter plus dot dot dot number of students having height 180 centimeter.

So this will give you total number of students right because if you this involves all the students number of having height 131 to 180 like all the we are measuring centimeter by centimeter and we are having this kind of a histogram and then we are doing this sum and at the end what will you get is the total number of students so that is what essentially shown in this slide here
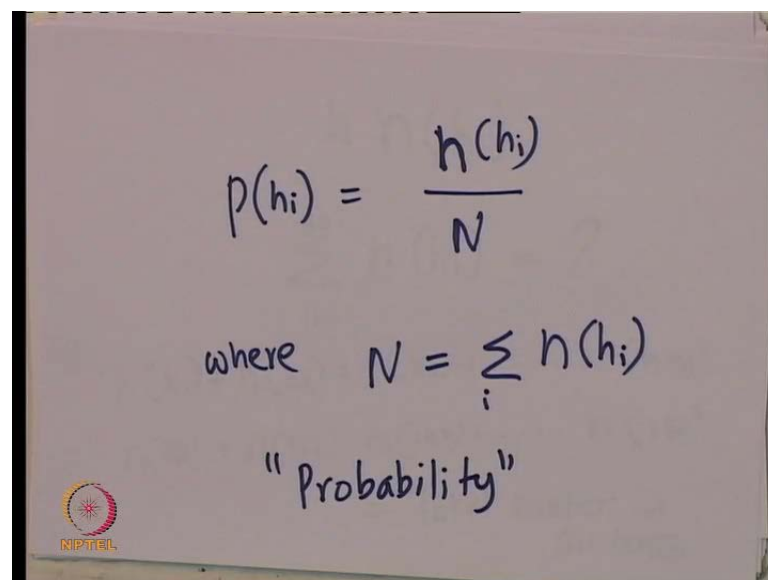
(Refer Slide Time: 24:32)

Sum over I n I by n I in n of h I i get something on the n which is the total number of students and if I divide this n of h by n I can define something on p of h.

In other wards if you wish you can you can write it as

(Refer Slide Time: 24:52)

p of h I is equal to n of h I divided by n where n is equal to sum over I n of h i.
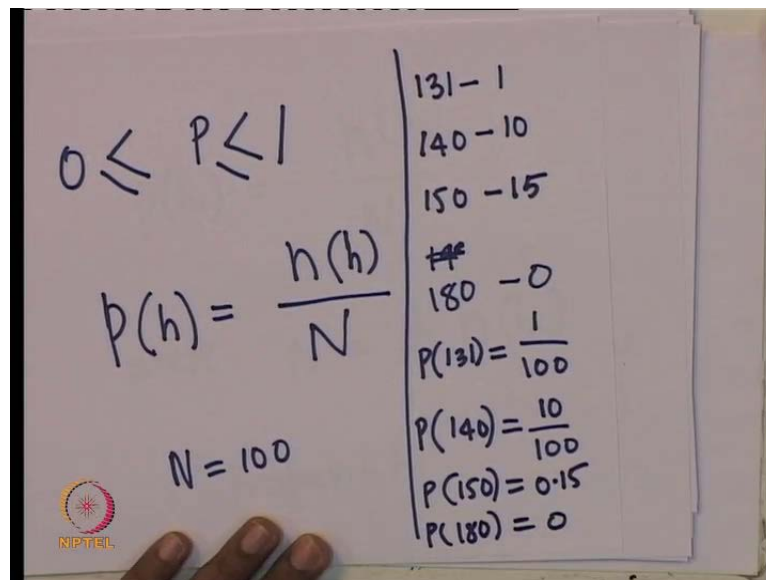
So actually in this slide here it should be n of h I so I can I can write here in this particular way so this is the correct way to write it.

So if this is the case what what what is p of h I so we can call the p of h I as probability so we will later come and understand what is this probability actually mean that let us call this as probability probability is essentially some number between 0 and 1 probability is some number between 0 and 1.

If you just at this moment it is for you to just realize what is the colloquial meaning of probability which you all know when you say something is very highly probable that means very likely that is very very something not very probable that means very unlikely.

So let us have understand only this much at this moment and let us also understand

(Refer Slide Time: 26:19)

the probability p is some number between 0 and 1 so it can be 0 or it can one or anything in between so probability is some number between 0 and 1 and probability that you have a height h that is what we actually defined now probability that you have a height h is number of students having that height h.

Divided by the total number of students so if you have one student so let us say now let us imagine that n is 100 let us take n is 100 then let us calculate so let us say there are one student with height 131 centimeter there are ten students with height 140 centimeter.

Let us say there are 15 students with height 150 centimeter.

And there are 0 students with height 180 centimeter let us say there are 0 students with height 180.

Let us say this is the case if this is a case we can call p of 131 probability that you have students have a height students in a class in your class have a height 131 is one by 100.

So this is the example just by following formula this what you get p of 140 is 10 by 100 so one by 100 is point 00 10 by 100 is point 1 so p of 150 is 15 by 100 which is point 1 5.

P of 180 is 00 by 100 so probability p of h is some number between 0 and 1.

So p of h is some number between 0 and 1. So, how does if you if you typically plot take large number of students in your whole school in your like are many schools and take the calculate p of h and plot it.

How will it look like
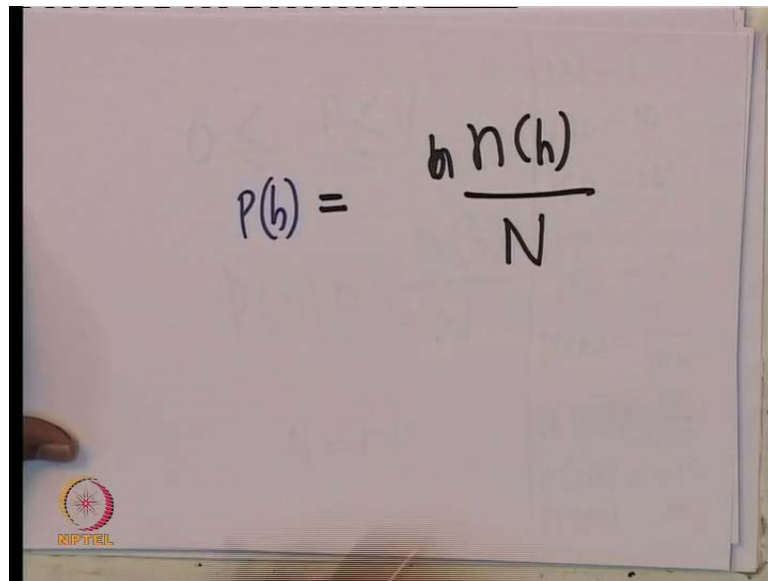
(Refer Slide Time: 29:34)
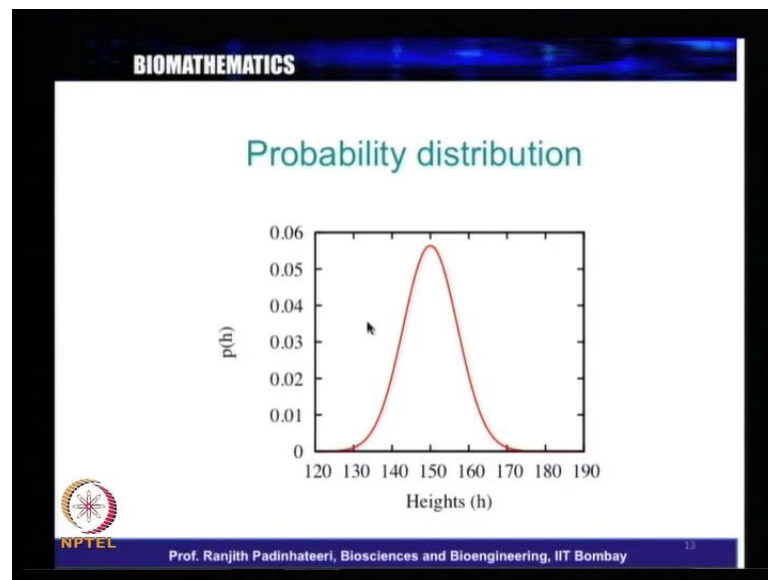
So it might look like this so have a look at this so we call this probability distribution so what is the what is this is the curve which is has a peak somewhere around the 150 and this axis it is height and this axis is probability to.

So this is the height h and this is probability of having height h. So vary 0 should not have the probability of having one twenty centimeter is nearly 0 there are unlikely that

anybody will have such a very small such a short person. Very high like above one nine and 1 nineties also nearly 0 somewhere in the middle there are many students like 150 there are many students so the probability of finding students having 150 is around point 0 5 in this example. So this is an exercise that you can do you can calculate the probability distribution the way it defined so that the probability distribution is probability of having students height h

(Refer Slide Time: 30: 57)

$$P(h) = \frac{n(h)}{N}$$

Probability of having students height h is n number of students having height h divided by the total number of students.

This will give you the probability now you have this probability how do you find the average can we find the average from this probabilities so it turns out that we can so let us say let us go back to the definition we had so we had del we have we had discussed
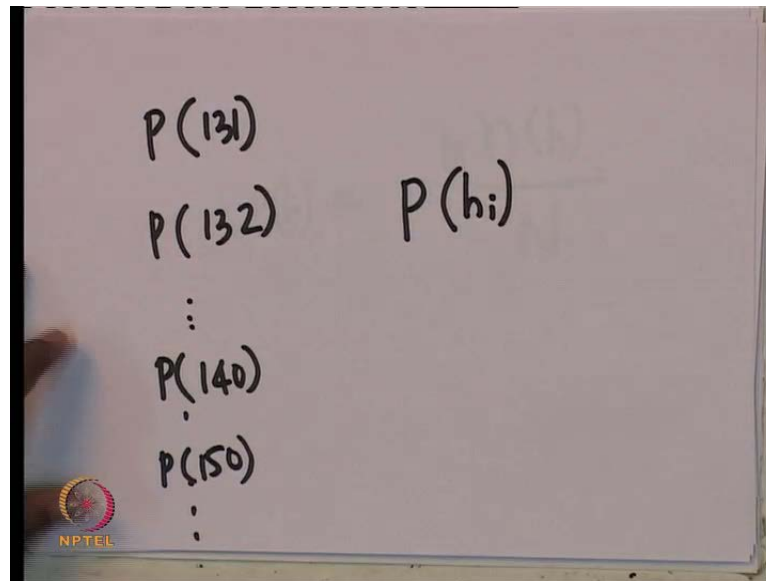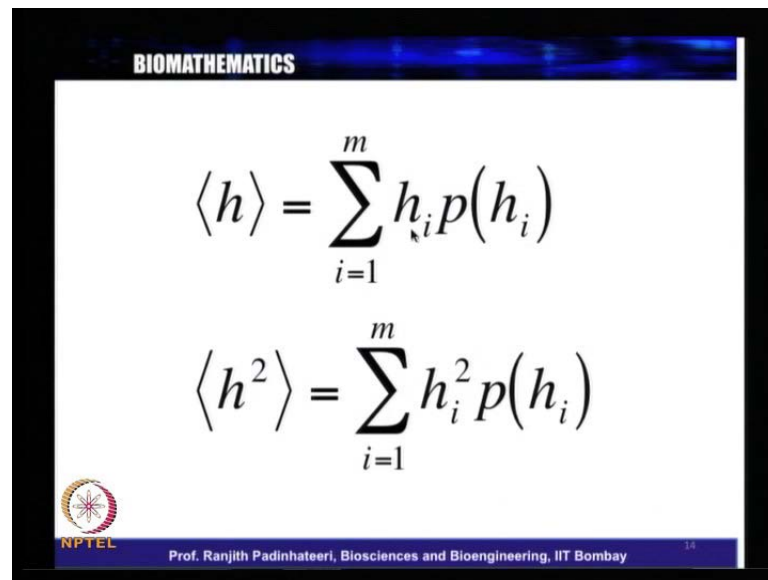
Probability of having student 131 centimeter.

Let us say probability of having students 132 centimeter similarly, let us say probability of having students 140 centimeter and probability of having students 150 centimeter so let us say you have this probabilities.

The average so basically you have probability of having students having some particular height h i.

So in the average is defined in the following way

(Refer Slide Time: 32:14)

The average is defined as sum over h I p of h I so this I goes from 1 to m so if you divide this to m intervals so in in our example we had 50 so 130 131 131 132 130 three 130 four up to 180 we had defined so 50 heights we had defined. If we define 50 height m is 50 so this sum over I 1 to 50 h of I p of h I this will give you the average and if I calculate this will give you the square average. So, we have average and square average h square average and h average; so there is this is essentially the same way we had calculated this. So instead of summing through the data sheet, we can multiply with the probabilities and sum like this; so then you get average and h square average. Do this calculation by taking an example in your case. Do an exercise yourself if you know h average and if you know h square average.

(Refer Slide Time: 33:32)

We can calculate h square average minus h average square which is standard deviation. So the standard deviation that we discussed before can be easily calculated in this particular fashion. So, now we had this h square average and h average square defined in this particular way.
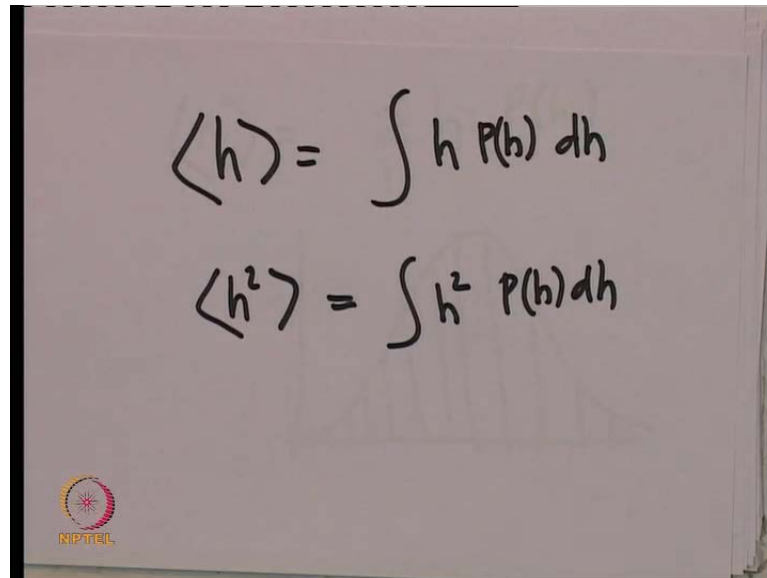
(Refer Slide Time: 34:01)

We had defined h average as sum over I h I p of h i. So now let us this h I we had one centimeter by one centimeter before but, let us say we can we can give h I in a very like 130 centimeter 130 point 1 centimeter in a continuous manner so if h is the continuous

function so height. Then, if we plot it, if you plot it let us say it look like something like this where every value of h you have a p of h i for any value of h you take there is a p of h I has continuous function in that case you can write this sum as an integral. So, just by using the idea but, we learned in the integration this sum can be written as an integral.
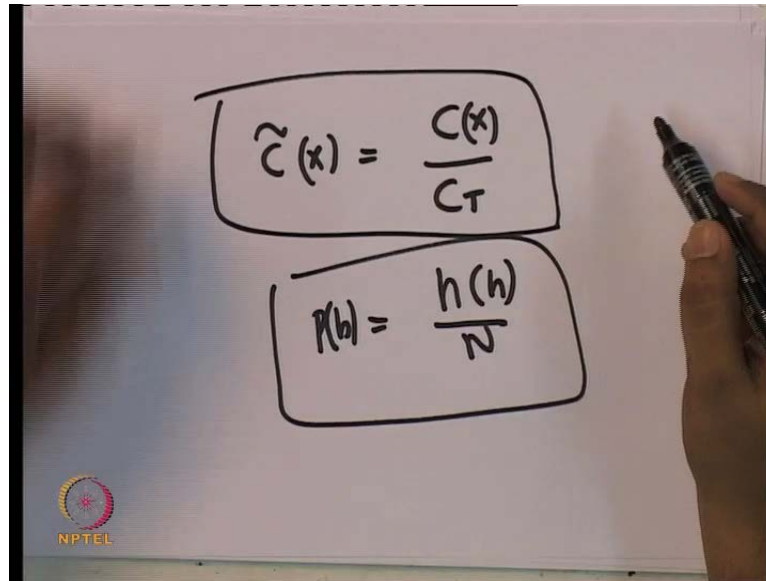
(Refer Slide Time: 35:13)

$$\langle h \rangle = \int h \, P(h) \, dh$$

$$\langle h^2 \rangle = \int h^2 \, P(h) \, dh$$

So then in that case the h average can be written as if h is the continuous function you can write this as h p of h d h so this can be converted to an integral if h is the continuous function similarly, similarly, if you had h square average can be written as integral h square p of h d h.
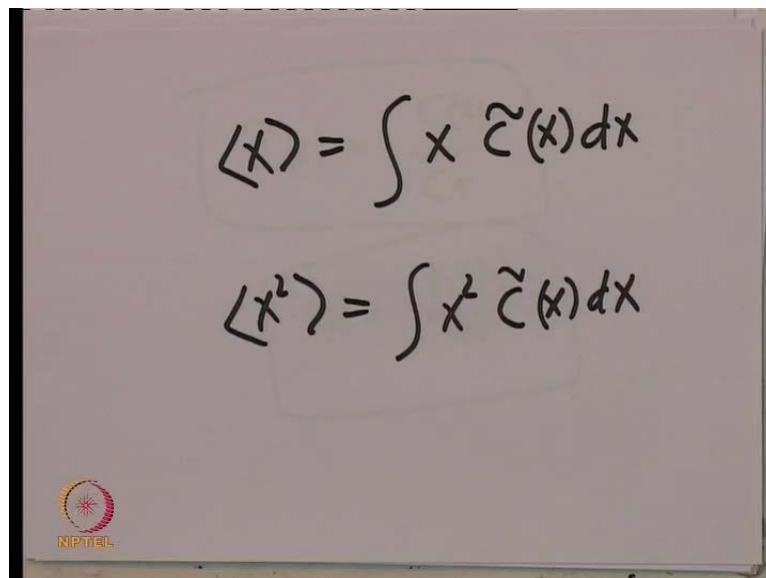
If h is the continuous function we can define the averages and square average in this particular way in fact when we did the case of diffusion concentration we had then precisely this. If you remember we had defined c tilde there is something called c tilde of x as c of x y total concentration.

This is just like we define to today N of x by m we define the p of h this is exactly the same way we have define concentration. So this is this appear like a probability so here this is probability we said that this is probability so this is also like a probability.

We had defined x average and x square average has x c tilde of x d x and x square average has x square c tilde of x d x.

If you go back to the lecture where we discussed diffusion we had discussed, we have defined x average and x square average in this particular way. So, the reason for defining

this is as we understand in today's lecture. If you have a distribution, c of x or c tilde of x was the probability distribution for concentration so the probability of having concentration at a particular distance x so what is the probability or the concentration at the particular distance x can be defined in some kind of a probability in this particular way and I can define the averages in this way. Ok, now let us go back to the distribution that we had.

(Refer Slide Time: 37:59)
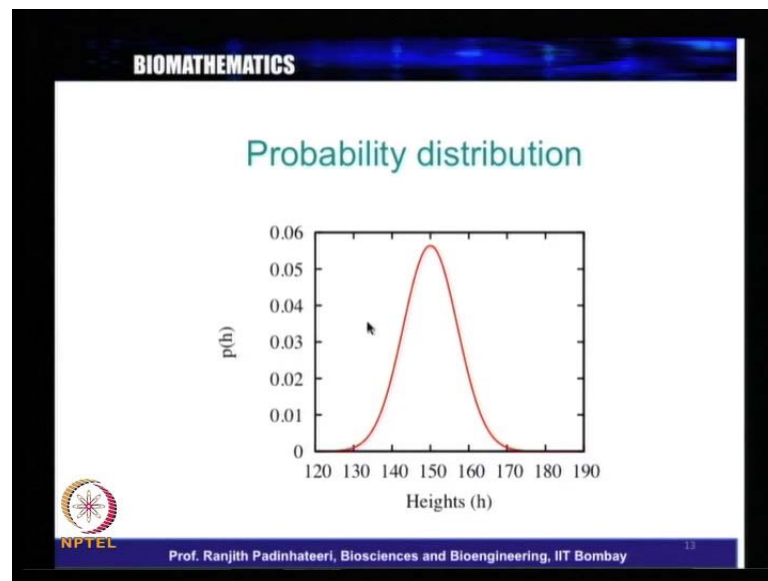
We had a particular distribution in this particular fashion so we had a curve which looks like this. Now, what is the name of this curve which is having in this particular kind of a distribution? So rough typically most of the things in nature have this bell shape curve. So this bell shape curve or this bell shape distribution is called normal distribution.

(Refer Slide Time: 38:34)
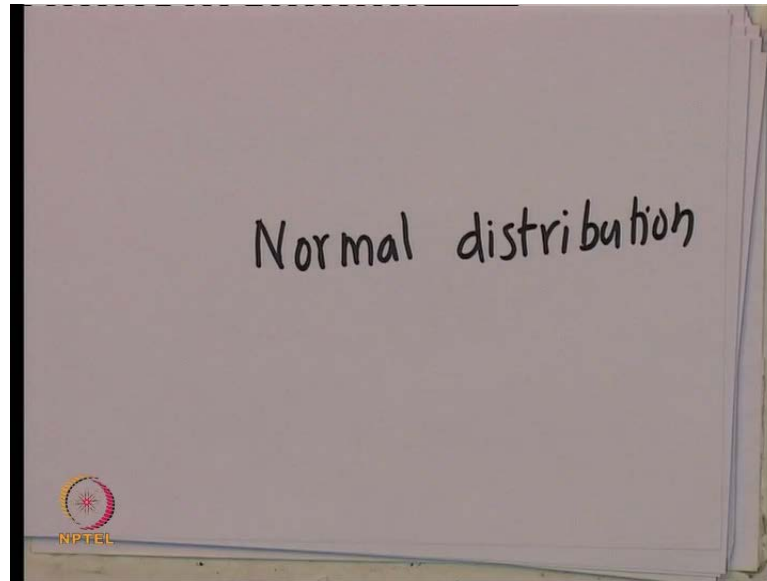
Let us write it so the bell shape curve is called normal distribution so many things in nature as it might be the case with height of students or the mark of students or there are many examples that we will come along as we go along and we will discuss as we go along so all these examples in all these examples the distribution might look like bell shape curve.

So, then this distribution called the normal distribution. What is a mathematical property of this normal distribution? How does it look like so the shape of this normal distribution can be written mathematically in this particular form?
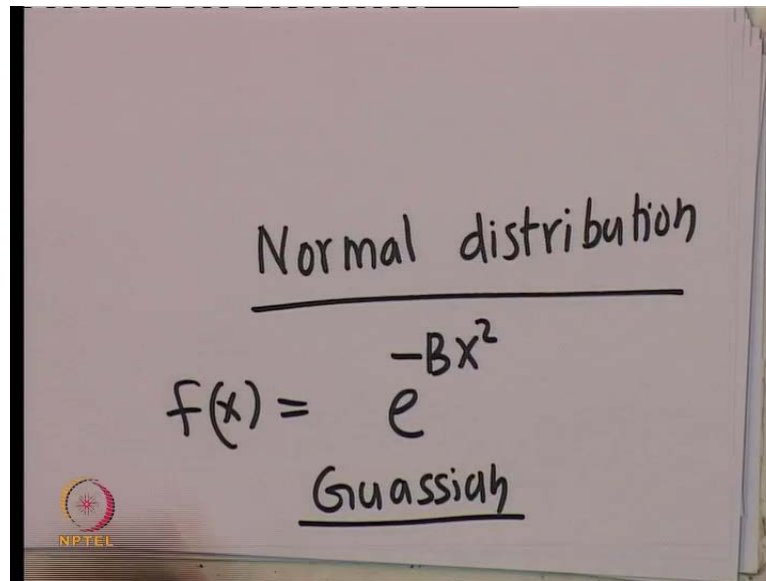
(Refer Slide Time: 39:24)

Look at here p of x is equal to a exponential some constant b into h minus h average whole square. This is the mathematical formula for a normal distribution where a and b are some constants so we will clearly understand in the coming classes and what is a and b stands for.

What is so there is like some constants if you wish in a simpler form we can write it e power minus b x square. So, this is simplest you can write it as e power minus b x square if you wish in an much more simpler manner. So this has a particular, if this kind of a, if you have a function this kind e power minus b x square so that is called a Gaussian function.

(Refer Slide Time: 40:28)

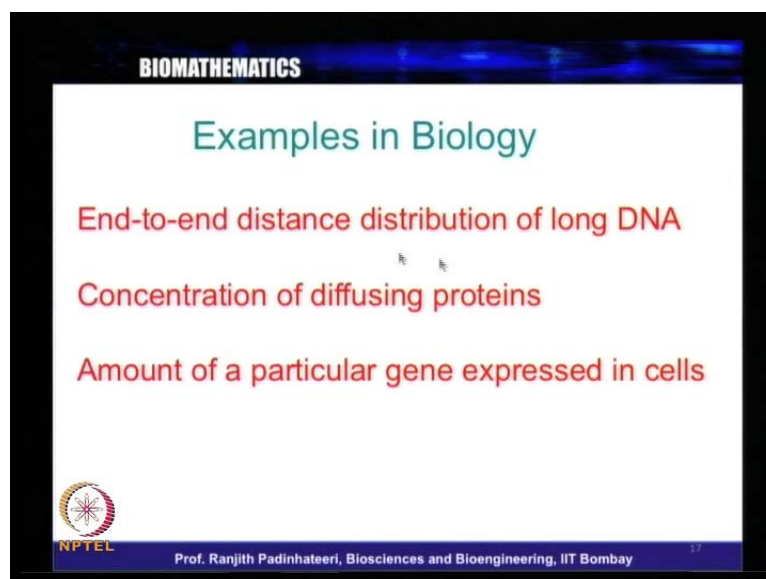If you have a function f of x which is e power minus b x square so this is called a Gaussian. The normal distribution also has a name called Gaussian distribution and which has a bell shape curve, now what is the meaning of this distribution? That we will come when we will discuss in the coming lectures; but, just realize that there are these. Examples are many; examples from nature fall into this category. So, let us look at some examples from biology; some examples from biology include End-to-end distance distribution of long DNA; what does this mean?
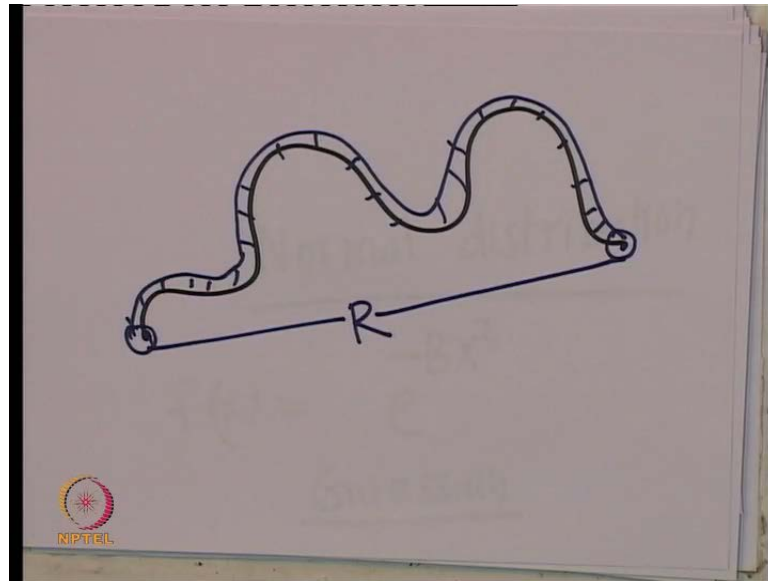
(Refer Slide Time: 41:14)

So, let us say <mark>you take a let us say</mark> you look at the DNA. Let us say, DNA has some particular shape and you ask the question what is the distance of the DNA from one end to the other end. So this is like a double standard DNA if you wish like it will have all those.

I am just showing is double standard DNA as just so let us say very long DNA and you can ask the question from this end to this end what is this distance let me call this distance is r now let us say that you have let us say that you have you have in a pastry dish you have a million DNA Avogadro number of DNA or large amount of DNA let us say you have a particular concentration of DNA and imagine that you have this amazing property that you can just take photograph of this DNA you have a let us imagine amazing device where you can take the photograph of DNA. You freeze the DNA at a particular moment and take the photograph. So if you do that, what do you expect?
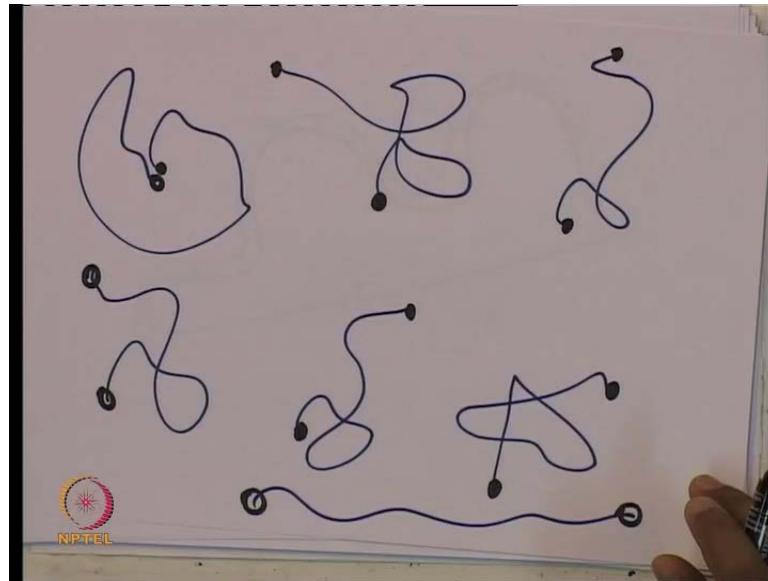
Sometimes, some DNA will be like this some DNA will be like this some DNA will be like this some DNA will have this shape some other DNA will have this shape some other DNA will have this shape so here some other DNA might have this kind of a shape now here the distance between the two ends is very small here the distance between the two end is very large.

Here the distance between the two ends is somewhere in between so here the distance between two ends is this here you have another distance here the distance is something else here the distance between the two ends is something else. So you can just like we did the experiment of measuring height you can do an experiment of measuring the end to end distance of the d n a and you can write make a histogram and plot it then it might look like it might look like a Gaussian distribution if you wish it a, it may look like a Gaussian distribution if you wish so.
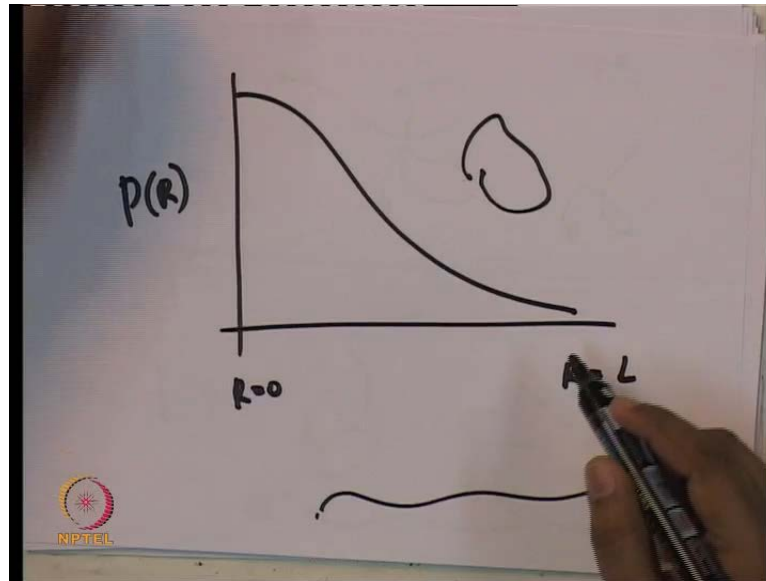
So it might look something like this when there can be many of them where the two ends are very close so this is r equal to 0 and this is r is r equal to l and this is p of r so there can be many of the DNA have this height which is very large and very small of them having may the probability that you will find the which is very short end to end distance that means there ends are very close to each other could be large and the probability you will see their ends far apart could be small if this is the case this will this might be look like a normal distribution so this is the half of a normal distribution the other half here which is the negative part which is actually meaningless; which we are not plotting so only one half of the probability distribution of the Gaussian r greater than 0 if you wish we can plot this is a vector.

But we will not come to the now that you can you can get such a distribution this is one half of a Gaussian distribution that is if you plot e power minus a r square between r equal to 0 and l this will look like this so if you plot this function e power some constant we will be define this as b. R square for some value of b between 0 and l this will look like this. So this is end to end distribution of a long DNA so look like at the slide here so this is the first example end to end distribution of a long d n a.
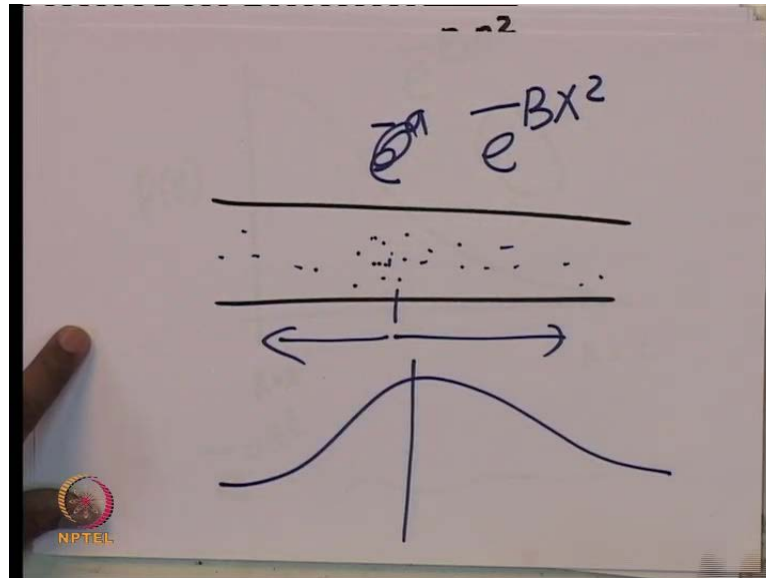
We already saw that concentration of diffusing proteins can have a Gaussian distribution. So, when we discuss the concentration it diffusing example we said that if you have a tube and if you look at a particular time. How many proteins are here as a distance if you go from x equal to 0 to either way there are large number of proteins at the middle of the tube and the fewer proteins at the end of the tube so this might have this kind of a Gaussian like a shape so if you plot this might also have a shape this will be like e power minus e power minus some b x square.

So this is another example; you could also think of another example which is let us say amount of a particular a gene expressed in cells. So let us say let us imagine that you can measure the amount of the gene expression and cell so you have a have a bunch of cells in a pastry dish and each of the cell you take a particular gene and then see how much of this gene is expressed. So you can you can count let us say how much m r n a's produced. Or how much gene expression has happened? So, then what you might get is something like this; let us plot here.

(Refer Slide Time: 48:13)
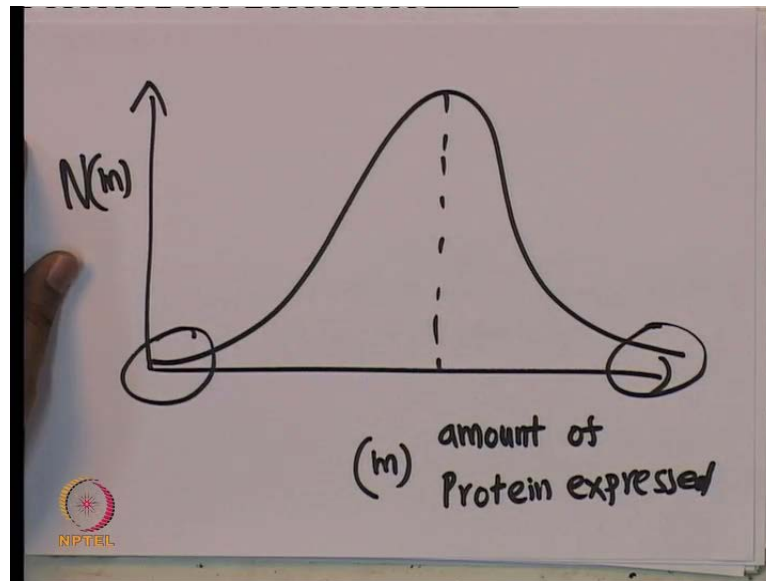
Let us plot in this axis so let us plot something like this and lets plot some function like this it should be little more symmetric so when I plot just an symmetric let us say it look like a Gaussian let this is not any like a Gaussian but, it should be very symmetric. Gaussian should be symmetric but, in the case of gene expression we do not what should be symmetric or not but, let us say you have such a curve now here in this axis we plot amount of protein expressed or it could be called as number of m r n a and amount of protein expressed and how many cells express this amount so number of cells so amount let me call this amount as m and this is n of m. How many cells expressed very little of this gene very few cells number is very small very few cells amount very little gene small amount large amount is also expressed like only few number of cells expressed large amount of protein maximum number of cells expressed something intermediate amount of protein.

If you have this example, this might also look like a normal distribution if we wish but, it will surely have distribution of this roughly this shape is the same shape which is peak somewhere in the middle and dying down to the both ends so here this is a amount of protein expressed verses number of cells having that particular amount of protein. So we had many examples in biology. So, to summarize we learned a few things; what all did we learn?
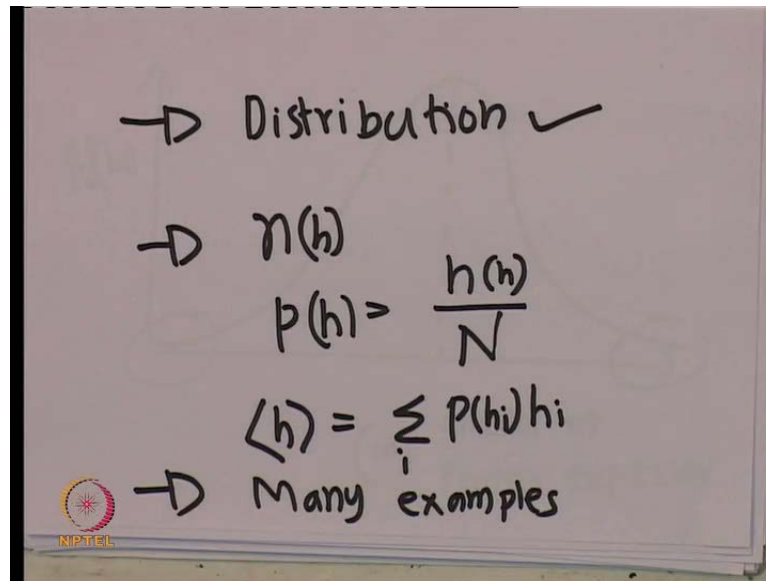
We learned that probability distribution, the distribution you can show, you can present the data in the form of distribution so that it makes much more sense like a histogram. So, the distribution is one thing we learned and how do we find averages etcetera from the distribution so if you know this n of h the number of students having height we can define probability distribution p of h has n of h divided by the total number. We can define averages by sum over I p of h I times h I we can also define standard deviation in this particular fashion and there are many examples in biology.

So, these are the things that we learned today. We learned about distribution the height distribution and how do we convert this distribution to probability distribution and how do we calculated have how do we calculate average and standard deviation from this probability distribution and many examples; so this is the summary of today's lecture.

We will discuss various other distributions and properties of distributions in the coming lectures and we will discuss many more biological examples. So with this we will stop today's lecture with this discussion of distribution or interaction of distributions we will stop today's lecture and we should you should remember this idea of distribution carefully think about this carefully because, this is some idea that we will be needing to learn statistics in a better way and this is very useful to present data and to analyze data. So, just introducing distribution we will stop today's lecture; thank you.