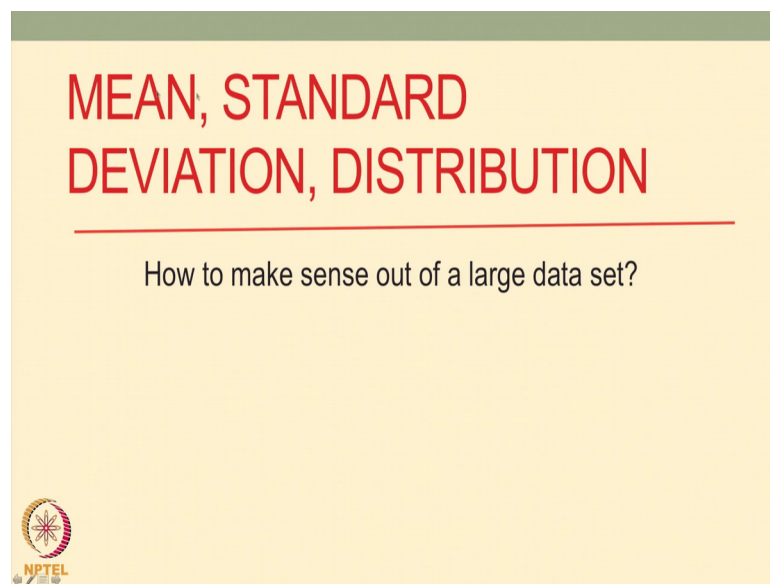


Introductory Mathematical Methods for Biologists
Prof. Ranjith Padinhateeri
Department of Biosciences & Bioengineering
Indian Institute of Technology, Bombay

Lecture - 36
Mean, Standard deviation and Distribution

Hi, welcome to this lecture on mathematical methods for biologists, in this lecture we will continue discussing about statistics.

(Refer Slide Time: 00:29)

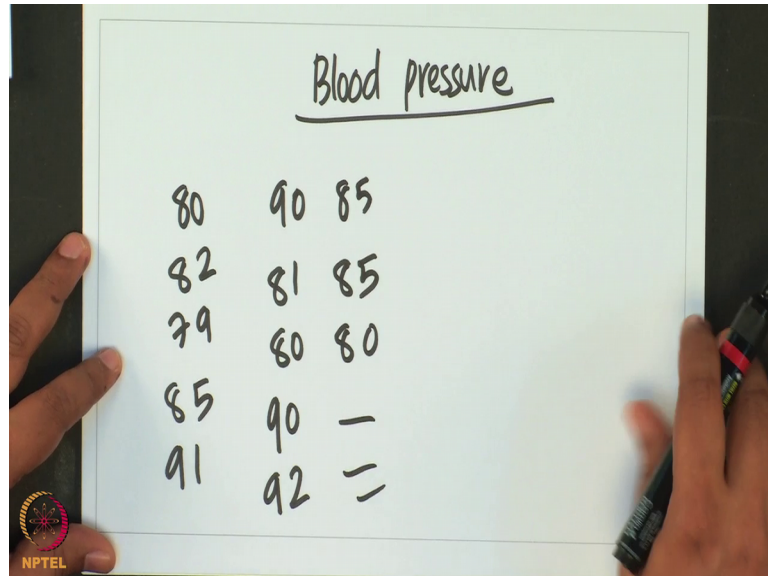


So, the title of this lecture is mean, standard deviation and distribution. And the question we will answer is how to make sense out of a large data set. So, imagine that you have collected large amount of data. So, what is data? The data would be typically you do an experiment and you would get a number. So, this experiment would be concentration of a particular protein in many cells right this is some experiment you would repeat this experiment many times, and you would get concentration of a protein or particular protein of your interest either as a function of time or in many sense. So, this is something that you could get.

Other example could be like if you are doing more of a medical sciences, this could be some quantities like blood pressure or it could be the amount of vitamin d in a population in a particular city or in a particular country, this could be any other quantity that related to health. If we take that and if we write it as numbers it should be a set of numbers.

So, I would take some example some set of numbers. So, let me write a particular example which is in our case let us call it like the blood pressure for example, number.

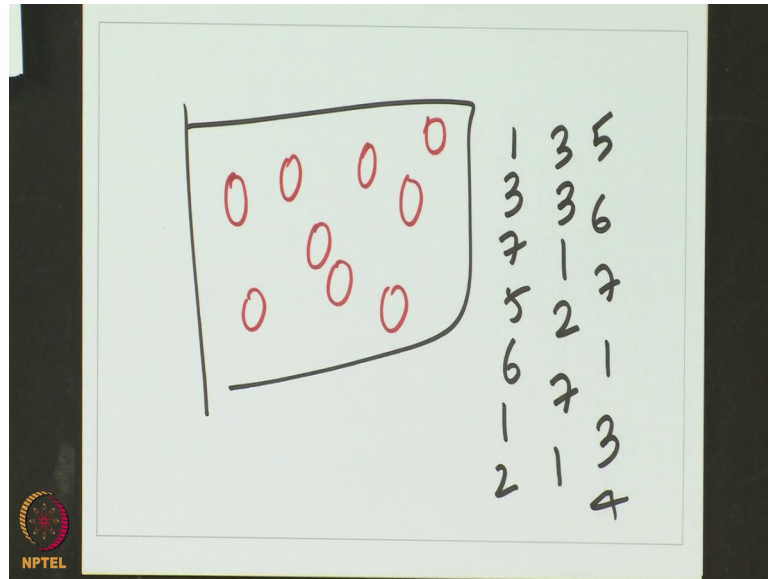
(Refer Slide Time: 01:56)



So, we will write only the smaller number. So, if you take a population and if you look at the blood pressure of the population, and if you write only the smaller number which is around 80. So, we would write 80 we you get some numbers 82, 79, 85, 91, 90, 81, 80, 90, 92, 85, 85. 80 all this numbers different numbers you would get.

So, what do you would this is 1 example if you measure the gene expression level of a particular gene right. If you look at quantify the concentration of a protein in many cells.

(Refer Slide Time: 02:51)



So, you have many cells a population of cells. So, you have a population of cells, in all of the cells if you look at the amount of a particular gene expressed the concentration, you could get some number again that could be some let us say the concentration is in micro molar or nano molar, but it could be like 1 3 7 5 6 1 2 3 3 1 2 7 1 5 6 7 1 3 4.

This could be the same or similar set of numbers you would get from any other health related quantity for example, you measure or any quantity that you measure in a particular experiment that you would do, whatever be the quantity that you measure you would get a set of numbers for a population.

Now, this set of numbers will be huge table, whose data set how do we make sense out of this data. So, we want to convey that we got this numbers, how do we convey this. So, if we have to convey the whole information contained in a huge data set, and if we could use just only just one number, but we have a huge data set, but to convey the meaning of the data I could use only one number, if I could use only one number I would use this number called mean or average. So, that is the importance of this number called mean or average.

(Refer Slide Time: 04:41)

Mean / Average

| | | | | | | | |
|---|-------|-------|-------|-------|-------|-------|-------|
| X | 10 | 15 | 20 | 10 | 10 | 25 | 30 |
| | x_1 | x_2 | x_3 | x_4 | x_5 | x_6 | x_7 |

$$\frac{120}{7} \leftarrow \bar{X} = \frac{x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7}{7}$$

NPTEL

So, you have heard of this which mean or average, what is this? We have many numbers 10 15 20 10 10 25 30 like many many many many numbers we have, this is some measurement that we took in some experiment this could be any quantity of your interest. So, let me call this x. So, this is x_1 , x_2 , x_3 , x_4 , x_5 , x_6 , x_7 this is a quantity that you measure either in 7 different individuals or 7 different tells cells or whatever be the context, we have 7 numbers and we wanted to convey the meaning of this we would calculate this quantity call \bar{x} which is nothing, but x_1 plus x_2 plus x_3 plus x_4 plus x_5 plus x_6 plus x_7 divided by 7 this is called the mean or average.

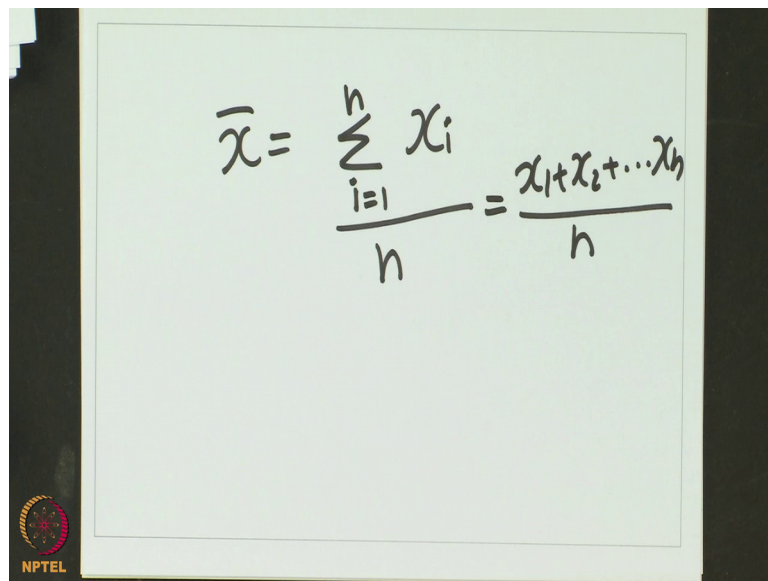
So, we would calculate average of these numbers. So, if I do this we have 25, 45, 55, 65, 95, 120 divided by 7 would be the answer here. So, if I do this I would get some answer which is for these numbers you would get some answer which is 120 divided by 7. So, let me if I do for this for these numbers I would get 120 divided by 7 which is just below 20 So, this is some number below 20. So, something between 15 and 20 you would get a number and this is a mean of all these numbers.

So, when do we use mean we would use mean, when we have to convey the meaning of a data set, but we could only use one number then that is the simplest way of conveying something. The mean gene expression the average gene expression is 3 micro molar; that means, if I add all the gene all the protein concentration of the concentration of mrnas

and divide by the number of cells I would get whatever 3 micro molar that is what it means.

If I just say that the vitamin d level in a population the mean vitamin d level is 25 units if I say that; that means, there is many many there are many many people, some people will have 24 unit, 22 units, 21 units, 26 unit, 27 units, 30 units and if I sum of all them and divided by the number of per person in this you would get 25.

(Refer Slide Time: 07:47)



The image shows a whiteboard with a handwritten mathematical formula for the mean. The formula is written as $\bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n}$. The whiteboard has a small NPTEL logo in the bottom left corner.

So, this is the mean something that we all know and mathematically this is written as \bar{x} is sum over i x_i divided by n . So, this i will from 1 to n ; there are n x_i s this is this is same as saying x_1 plus x_2 plus dot dot dot x_n divided by n . So, this is the thing called mean.

However very often if I convey something only by using one number which is the mean it could be misleading. There are examples where this kind of description could be misleading for example, let us think of an example where we are talking about the average mark in a particular class. So, let us say you have or your own average mark that you got in various subjects.

So, you have various courses course 1, course 2, course 3 and then you got this mark and you want it you are telling somebody the average mark. So, if I just take a person x and that has 3 courses. So, let us take the example. So, this is a person whose name is Raju.

(Refer Slide Time: 09:07)

| | <u>Kishore</u> | <u>Raju</u> | <u>Ajay</u> |
|----------|----------------|-------------|-------------|
| Course 1 | 100 | 70 | 90 |
| Course 2 | 60 | 60 | 60 |
| Course 3 | 20 | 50 | 10 ⇒ Failed |
| Mean | 60 | 60 | 50 : Mean |

It is the person and let us this person had 3 courses course 1, course 2 and course 3 three courses and let us assume that he got marks in this courses as follows. So, let us say the marks 1 got was the first got 70, 60 and 50. So, these are the 3 marks one got and the mean of this 3 numbers is of course, 60; 50 plus 70 120 180 divided by 3. So, the mean the mean is 60.

If there is some other person other than Raju let us say Ajay some other person the same person this Ajay got some marks let us consider this marks as the following, let us say he got in one course only 10 marks one got right in another course you got 60 marks right. So, and let us say this got 90 marks. So, let us say 90 marks.

So, then this is let us say let us say 80 marks this is fine. So, this is 60 plus 90 150 divided by 3, this is 6. So, here the average would be 50. So, here this mean is 50. So, if I have this mark and if this Ajay says I got 50 percent mark on an average, if I just convey only the mean the mean here is 50 the mean is 50.

So, if Ajay says I got 50 marks on an average in 3 courses that the person who listens who would think that he did reasonably well he got 50 percent marks, but the fact that Ajay has failed in this third course. So, the course number 3 Ajay got only 10 out of 100 so; that means, this course the Ajay is fail, but this does not get conveyed in this quantity called mean. So, the mean is a quantity that conveys something, but there is some information which is not conveyed, the fact that there are some extreme numbers like 10

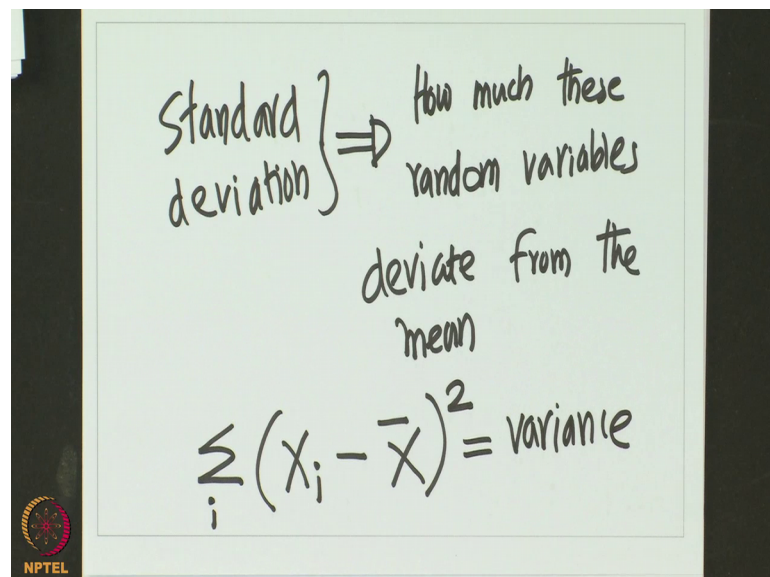
and in another course same Ajay got very high marks like 90 marks. So, he performed very well.

So, this kind of information the same thing like for a fourth person could have a different set of numbers. So, let us say another person. So, another person Kishore could have numbers which is 100 20 and 60 this would have the mean sixty. So, here also the mean of this and the mean of this exactly the same, but even here Kishore has done very badly in course 3 while course 1 he did extremely well

So, just by saying that the mean is sixty. So, both for Kishore and Raju the mean is 60 marks, but Kishore has essentially failed in one of the course. So, in this course the Kishore is while the mean is exactly the same. So, just by saying the mean one does not convey complete information and we need to do something beyond and that's why one uses this thing called standard deviation. So, this quantity called standard deviation which is how numbers how far these numbers are from the mean, this quantity is very useful

So, very often mean may not be necessary or mean may not come convey the complete information.

(Refer Slide Time: 13:47)



Standard deviation } \Rightarrow How much these random variables deviate from the mean

$$\sum_i (x_i - \bar{x})^2 = \text{variance}$$

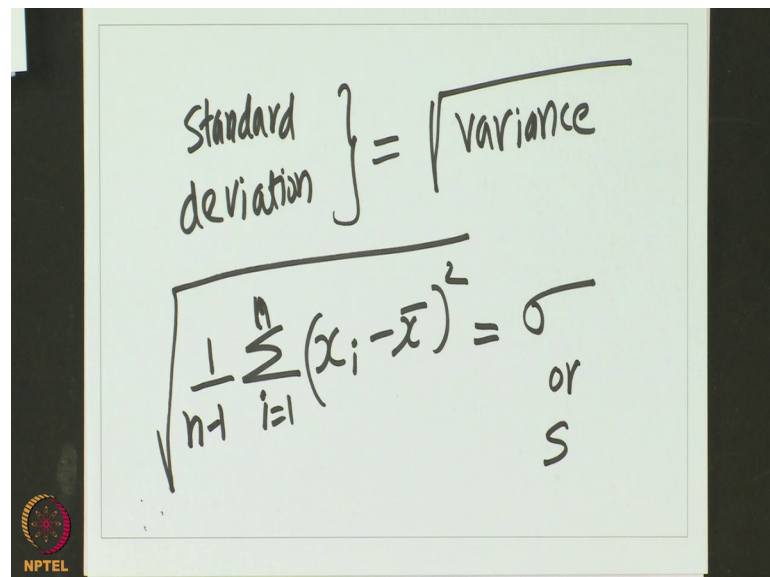
NPTEL

That time I would use this quantity called standard deviation. What is the standard deviation tell us? It tell us a how much these variables this random variables whatever

our marks or deviate from the mean. So, this is what this standard deviation would convey.

So, if I just take the mark of the i th course or i th measurement, and subtract the mean I will get a number if I sum this I will get 0. So, this does not make any sense. So, I have to square and sum. So, you take each number subtract the mean from it, square and sum this is called variance and the square root of this is called a standard deviation. So, this thing is called variance and square root of this variance is called a standard deviation. So, the square root of variance is called standard deviations (Refer Time: 15:15) right.

(Refer Slide Time: 15:19)



The image shows a whiteboard with handwritten mathematical formulas. At the top, it says "Standard deviation } = \sqrt{\text{variance}}". Below this, it shows the formula for standard deviation: $\sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = \sigma$ or s . In the bottom left corner, there is a small logo for NPTEL.

So, standard deviation is the square root of variance, which is nothing, but you take all numbers subtract this square it and sum 1 to n . So, of course, you have to divide in this you have to divided by 1 by n minus 1, there is a you have to divided this by 1 by n minus 1 then only this would become the variance. So, the 1 over 1 by n minus 1 if I divide, and the square root of this is called variance, which sometime would be represented as sigma or sometimes by s standard deviation s , which is sigma square is the variance sigma or s this is what we would use to represent this quantity.

So, let us take some numbers and calculate the standard deviation as an example. So, let us take let us say you measured vitamin d level in a group of people, and we got the following numbers or let us say the numbers.

(Refer Slide Time: 16:47)

The image shows a whiteboard with handwritten mathematical work. On the left, a list of numbers is written vertically: 20, 25, 30, 25, 20. A horizontal line is drawn under the last number, 20. To the right of this list, the number 120 is written. Further right, the formula for the mean is written: $\frac{\sum_{i=1}^5 x_i}{5} = \frac{120}{5} = 24$. Below this, the mean is written as $\bar{x} = 24$. The whiteboard is held by two hands, and a black marker is visible in the bottom right corner. An NPTEL logo is in the bottom left corner.

We got from a group of people are 20, 25, 30, 25, 20 these are the numbers vitamin d levels at 1 2 3 4 5 different individuals, and we know how to calculate the mean. So, this is 120; 1 2 3 4 5 divided by 5. So, this was basically 120 is the sum. So, sum over I 1 to 5. So, this is my x_i x 1 x 2 etcetera. So, this is my x_i sum over I x I divided by n 1 over n which is 5. So, which is 5 here. So, this is 120 divided by 5 which is 24 So, the mean is 24. So, \bar{x} is 24.

Now, what we need to do is to subtract \bar{x} from each of these numbers. So, let us do this little carefully.

(Refer Slide Time: 18:03)

| x_i | $x_i - \bar{x}$ | $(x_i - \bar{x})^2$ |
|---------------|-----------------|---------------------|
| 20 | -4 | 16 |
| 25 | 1 | 1 |
| 30 | 6 | 36 |
| 25 | 1 | 1 |
| 20 | -4 | 16 |
| $\bar{x}: 24$ | | 70 |

$\sigma^2 = \frac{70}{n-1}$
 $= \frac{70}{4}$
 $\sigma = \sqrt{\frac{70}{4}}$
 $= \sqrt{17.5}$

And if I look at this quantities. So, this would be 20, 25, 30, 25, 20 the mean of this would be \bar{x} which would be 24. So, the mean of which is 40 plus 50 plus 90 plus 120 divided by 5 which would be 24.

Now, the next thing to calculate standard deviation you should calculate is x_i minus \bar{x} . So, that is I should subtract this \bar{x} from each of this. So, 20 minus 24 is minus 4, 25 minus 24 is 1, 30 minus 24 is 6, 25 minus 24 is 1, 20 minus 24 is minus 4. So, sum of this would be 0. So, that is not what we want this is 8 in the minus 8 and plus 8 and this sum would be 0, but what we want is x_i minus \bar{x} whole square, which is 4 square which is 16, this is 1 and 36 and 1 square is 1 and 4 square is 16. So, this the 36 plus 16 is 42 plus 52, 53, 54. So, this would be 70, and sigma square is 70 divided by $n - 1$ there are 5 of this. So, we would do 70 divided by 4 and this would give us an approximately. So, this is going to be 17.5 roughly approximately close to 70.

So, whatever be this quantity 70 divided by 4 would be the sigma square and root of this 70 divided by 4. So, sigma would be this root of 70 divided by 4, which is essentially root of 70 divided by 4 is 17.5. So, there will be like root of 17.5. So, this answer the standard deviation is root of 17.5. So, which is approximately 4 something 4 point something right.

(Refer Slide Time: 20:38)

Handwritten calculations on a whiteboard:

$$20, 25, 30, 25, 20$$
$$\bar{x} = 24 \checkmark$$
$$\sigma = 4 \cdot () \leftarrow \sqrt{17.5}$$
$$\approx 4.2 \quad \underline{\underline{24 \pm 4.2}}$$

NPTEL logo is visible in the bottom left corner of the whiteboard image.

So, if we have this number. So, if we had this number 20, 25, 30, 25, 20 these were the numbers we had, and we had mean which is 24, and the standard deviation sigma was 4 point something some quantity which was to be precise which was root of 17.5.

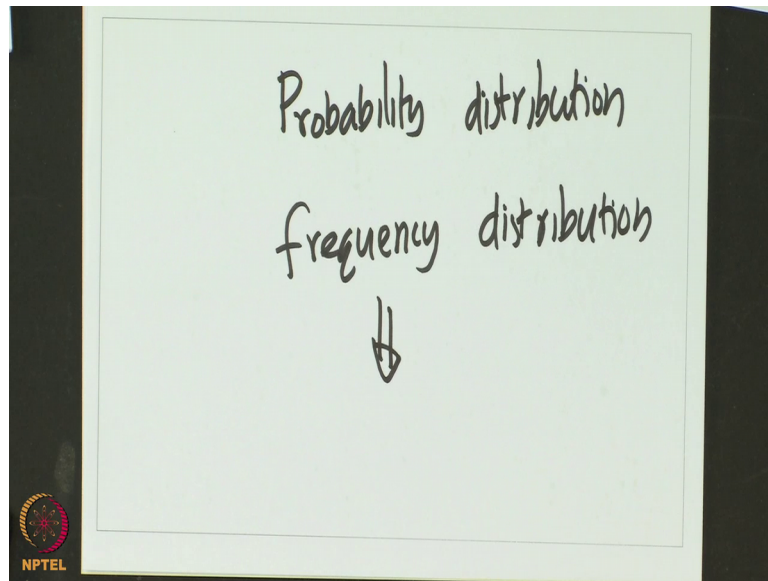
So, approximately 4 is the standard deviation you would get, if I little above 4 is what you would get which. So, we have 2 numbers now to convey which is one is 24 which is the mean, and the standard deviation let us say this is approximately 4.1. So, let us say approximately this is let us take it as 4.1 or 4.2. So, let me take it as 4.2.

So, if I approximately take it as 4.2 this number would be the standard deviation. So, a typical number here I would write it as, 24 plus or minus 4.2 that would be if I just now convey 2 numbers, which would be 24 plus or minus 4.2 this is just for an example.

So, I am now conveying instead of conveying all this numbers, I could just convey 1 number or I could convey 2 numbers which is 24 plus or minus 4.2. So, there are 2 numbers standard deviation and the mean. So, we could convey datasets by 2 numbers mean and standard deviation.

Now, what else we can we do of course, we can do a few other things, and the ultimate thing is to plot this data as a distribution. So, how the numbers are distributed according to the ranges that we take. So, this is called a distribution how. So, very often these are called like probability distribution or frequency distribution.

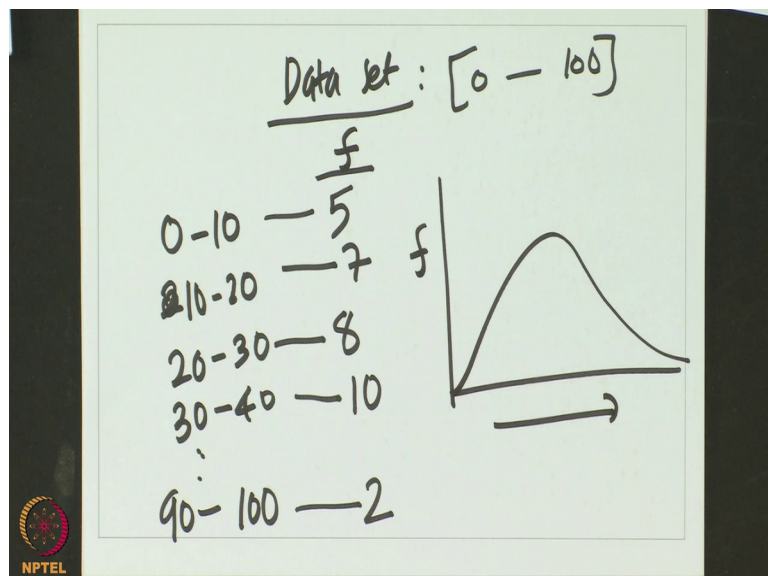
(Refer Slide Time: 22:51)



So, we will talk about this how this variable frequency distribution. So, this is another thing which we would discuss how these numbers are distributed between some range of our interest. So, this is something that we would also discuss. So, we will imagine that you have marks or any quantity, between some numbers let us say between 0 and 100.

So, we would ask a question suppose we would have lot of numbers between say all possible numbers between 0 and 100, we could ask a question how many numbers are there between 0 and 50; 10; so in the data set given to us.

(Refer Slide Time: 23:46)



So, you have a huge data set, and those data set is between 0 and 100. All numbers are between 0 and 100, we could ask the following question how many numbers are between 0 and 10, how many numbers are between 10 and 20, how many are between 20 and 30, how many are between how many numbers are between 30 and 40 and so on and so forth finally, we will ask how many numbers are between 90 and 100.

So, let us say there are 5 numbers here, 7 numbers here 8 numbers between 20 and 30, 10 numbers between 30 and 40, 2 numbers between 90 and 100 this will tell us a distribution of this mark. So, this I would call f and this is the mark range. So, I can plot f versus the range of marks and I would get some curve.

So, this is called a distribution about which we will discuss in detail later. So, to summarize what we have been saying, to describe a huge data set we would use 2 or 3 things which we discussed. One is the mean or average which we know how to calculate the second thing is called the standard deviation, which is how much a quantity is deviated from the mean, and the third quantity a third way is to take the whole thing and draw as a distribution about which we will discuss in detail later, but we could do a we could compute the distribution from the data, which would have maximum information about this data and very easy to convey. Something about which we will discuss with this we will stop this lecture and continue in the next lecture. Bye.