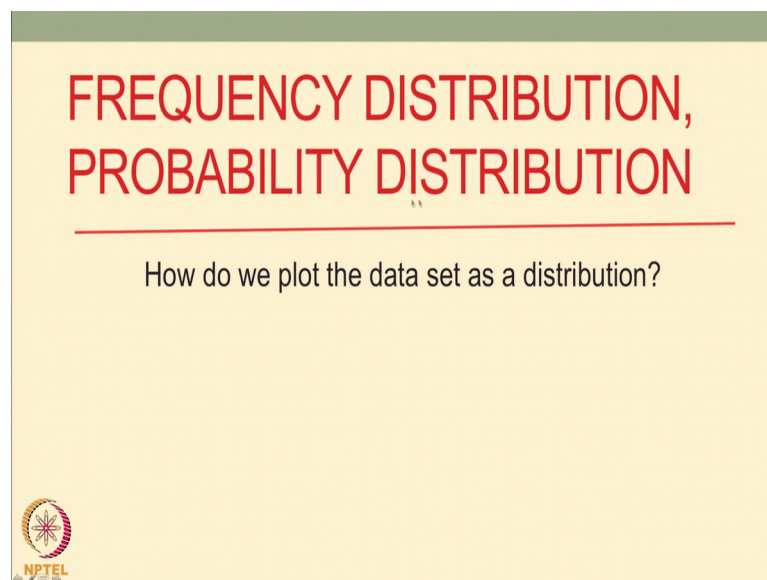


Introductory Mathematical Methods for Biologists
Prof. Ranjith Padinhateeri
Department of Biosciences & Bioengineering
Indian Institute of Technology, Bombay

Lecture – 37
Frequency Distribution and Probability Distribution

Hi, welcome to this lecture on Mathematical Methods for Biologist. We have been discussing statistics and we learned why do we need and when do we use mean or a average and standard deviation and what does it imply. In this lecture we will discuss distribution. So, the title is frequency distribution and probability distribution. And the question we will answer is how do we plot the data set as a distribution.

(Refer Slide Time: 00:40)

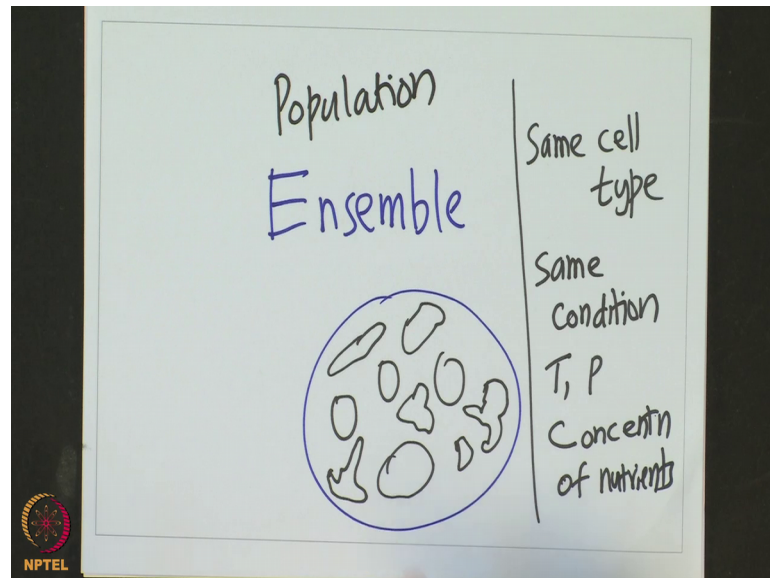


This is a thing that we would answer in this we given a data set how do we plot it as a distribution.

Now, while we discuss this? There is an important thing that one should remember this idea of an ensemble so, or a population. So, this is something that this word very often you might. So, what is an ensemble? Typically imagine that you have an experiment in which you might be looking at many cells for example, under a microscope. So, the all the cells are exactly under the same condition, same temperature, same concentration of nutrients and all that, but they are a population they are not they are individual cells. So,

this is let us think of it, you have a petri dish or you have a culture, you have many many many cells in this.

(Refer Slide Time: 01:48)

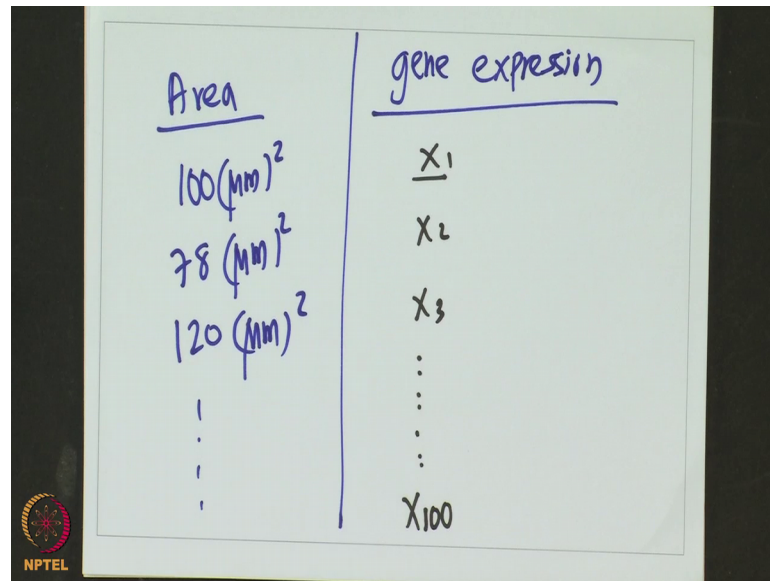


So, you have all of this might have different size a little bit it might have different shape, but these are all same cell type. So, these are all same cell type same condition same cell type. So, let us say they are all e cells or they are all hela cells or same cell type, same condition.

So, the things like temperature pressure and all that there is no these all fixed and the what we have fixed is number of cells is given here and the nutrients the concentrations like the concentration of nutrients, concentration of chemicals nutrients whatever in this we have which all essentially the same for all the cells. So, concentration of nutrients let me write they are all essentially seeing the same condition.

So, you have, this we would call an ensemble of cells or a population. So, these two words would be very often used you have a population or an ensemble. So, this ensemble of cells and in such set of cells you can do various measurements for this is, you could measure for example, its size like you could measure the area of each of these cells and the area of this would be different from the area of this. So, they would all get you will get different numbers. So, you can get a set of numbers let us say you are measuring area. So, you would get a set of numbers like.

(Refer Slide Time: 04:03)



A handwritten table with two columns. The left column is titled 'Area' and the right column is titled 'gene expression'. The 'Area' column lists values: 100 (mm)², 78 (mm)², 120 (mm)², and vertical ellipsis. The 'gene expression' column lists values: X₁, X₂, X₃, vertical ellipsis, and X₁₀₀. An NPTEL logo is visible in the bottom left corner of the slide.

<u>Area</u>	<u>gene expression</u>
100 (mm) ²	X ₁
78 (mm) ²	X ₂
120 (mm) ²	X ₃
⋮	⋮
⋮	⋮
⋮	X ₁₀₀

So, area, you would get some numbers in micron meter square. So, let us say 100 micrometers square or 78 micrometer square or 120 micrometer square. So, these are typical example that you can get many many many numbers. So, if you have 100 cells the more the cells more numbers you can get.

So, area is some example I said you could measure for example, gene expression. So, there is a fluorescence and you can you can measure the fluorescence for example, different cells you can measure the fluorescence like you would get some value in some units. So, value 1, you will get X 1, X 2, X 3 and many many many many values like if there are 100 cells you would get X 100. So, this is the amount of gene expression this could be in micro molar or nano molar or whatever be the unit, but you would get some concentration of proteins expressed in this cells.

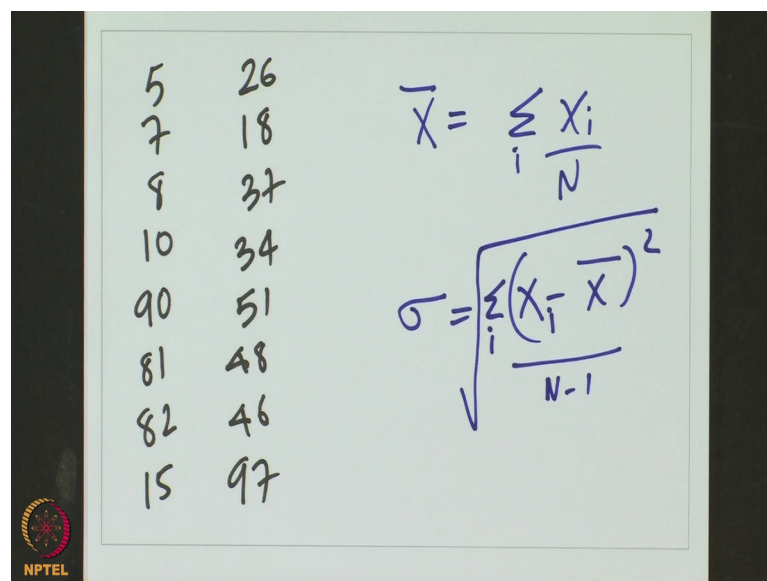
You would want to measure some other quantity if need not be cells always, it could be you have a population of, you have a group of people who whom for whom you would want to measure for example, the vitamin D like people living in a particular city they are in the same city under say similar conditions and you would want to measure their vitamin D level, and something else could be blood pressure and you would get many different values.

So, if you have a population of cells or a population of a group of people or it could be a group of whatever be the organism that you would want to study you can any

measurement you can do a on many individuals many cells this is an ensemble of population. And the data you would get a large number of measurements this data this is one example. Another example could be you could take one individual and repeatedly measure the same thing like you would measure for example, the blood pressure every day like or many times in a day and that also you will get many measurements.

So, typically we let us think of population first and we will come to the measurement as a function of time. So, if you have a population of cells like this or a population of individuals like this you would measure either its size gene expression or you would measure any other quantity like blood measure, blood pressure or vitamin D level or hemoglobin count any quantity that you measure in a population you will have many numbers like this. So, let us say you have numbers between 0 and 100. So, you have like many numbers I am just drawing some example here.

(Refer Slide Time: 07:24)



The slide displays a list of numbers in two columns and two mathematical formulas. The numbers are: 5, 7, 9, 10, 90, 81, 82, 15 in the first column, and 26, 18, 37, 34, 51, 48, 46, 97 in the second column. To the right of the numbers, the mean formula is given as $\bar{X} = \frac{\sum_i X_i}{N}$. Below that, the standard deviation formula is given as $\sigma = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{N-1}}$. In the bottom left corner, there is a small NPTEL logo.

5	26
7	18
9	37
10	34
90	51
81	48
82	46
15	97

$$\bar{X} = \frac{\sum_i X_i}{N}$$

$$\sigma = \sqrt{\frac{\sum_i (X_i - \bar{X})^2}{N-1}}$$

So, let us say the numbers are 5, 7, 8, 10, 90, 81, 82, 15, 26, 18, 37, 34, 51, 48, 46, 97 and so on and so forth, so you have many many many numbers. If you have these are data measured from different a (Refer Time: 07:58) in a from a population ensemble you can of always calculate the mean. So, you can calculate the mean \bar{X} which we said is sum over i x_i divided by N . You can also calculate the standard deviation which we said is sigma which is X_i minus \bar{X} whole square sum over i divided by N minus 1 and root of this. So, this is the two things that we can measure.

But you can the third thing is of distribution which we briefly mentioned distribution how these numbers are distributed. So, from this data we can make a table, from the data sets we are given we can ask a question how many of these numbers are between 0 and 10, so how many of these numbers are between 0 and 10. Let us say, this is some table that we would create.

(Refer Slide Time: 08:53)

How many times numbers fall in a particular range?

Range	Frequency
0-10	2
10-20	5
20-30	11
30-40	15
90-100	3

Range Frequency

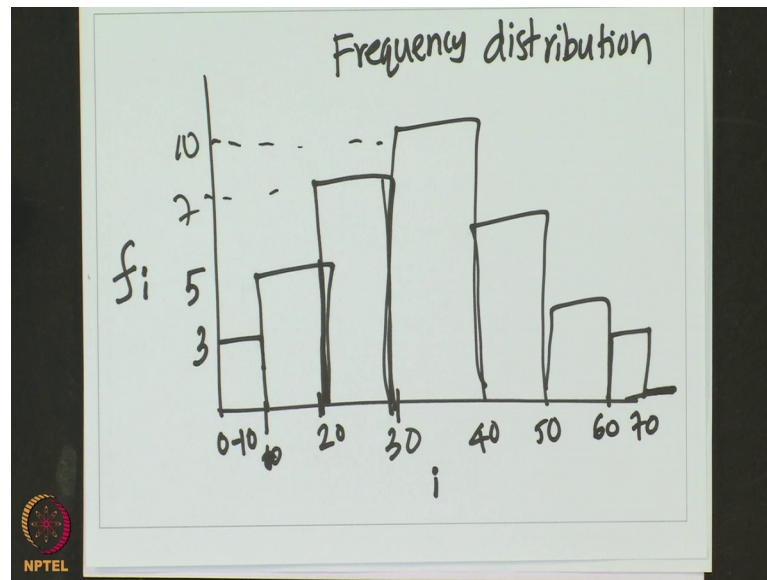
The image shows a handwritten table on a piece of paper. The table has two columns: 'Range' and 'Frequency'. The rows are: 0-10 with frequency 2, 10-20 with frequency 5, 20-30 with frequency 11, 30-40 with frequency 15, and 90-100 with frequency 3. To the right of the table, a handwritten question asks 'How many times numbers fall in a particular range?'. Below the table, the words 'Range' and 'Frequency' are written with arrows pointing to their respective columns. An NPTEL logo is visible in the bottom left corner of the slide.

So, let us say two of those numbers are between 0 and 10 and how many of they are between 10 and 20. So, let us say 5 then you could ask the question how many of them are between 20 and 30. So, there are many numbers between 20 and 30, 11 numbers out of all those numbers and how many of them are between 30 and 40 and, there is 15 and so on and so forth. So, you could just finally, you will have ask the question how many numbers are between 90 and 100, you would get 3. So, you got some table like this where you have a range here, you have a range and this is frequency how often numbers fall in this range.

So, this is frequency this is this is range and this is frequency. So, this is frequency this is the range, what is frequency, how often, how many times, how many times or how often numbers fall or appear in a particular range in a particular range of our interest. So, this is the frequency, how many times numbers fall. So, 5 times in this whole population, 5 times the area is between 10 and 20, 11 times the area is between 20 and 30, 15 times the area was between 30 and 40 and so on and so forth.

So, if you have a table like this you can plot a histogram. So, basically this versus this we can make a plot. So, we would this is the plot that we would make. So, we would have between 0 and 10.

(Refer Slide Time: 11:14)



So, this is between the first range how many 3 times, this is 3 times, this is 0 to 10 range and between 10 and 20, between 10 and 20 the numbers in this range was 5 times. So, between 10 and 10 to 20 range, this is 20 in the 10 to 20 range you had 5 times; 20 to 30 range you had 7 times, in the 20 to 30 range the numbers were 7 times and so on and so forth you could ask 40 to 50 times, 50 to 60 times and you would get a distribution like this what does. So, this is 7 and this is 10 and so on and so forth some numbers. What is that? So this is called a frequency distribution.

So, I am going to plot here f_i versus i . So, i is the range range 1 2 3 and this is your frequency for that particular range. So, 0 to 10, 3 times; 10 to 20, 5 times; 30 to whatever 40 like 10 times and so on and so forth. So, this immediately tells us if you have a graph like this, this immediately tells that most of the area fell between 30 and 40, most of the area fell between 30 and 40. Very few fell between 0 and 10 and very few fell between like whatever the large number. So, this is 50 60 70, between 60 and 70 very few fell and beyond that there was nothing there is no number above this for example. This is one example I am plotting this tells us how numbers are distributed, how many numbers are between 0 and 10, how many numbers are between 20 and 30. So, this graph will tell us

how the numbers are distributed in this range. So, this is the frequency distribution which most of you know.

Now, given this distribution this range this smaller the range of course, the better it would be the more data we have we could take the small smaller range and once you have this number the what we can calculate. So, let us have this distribution once more.

(Refer Slide Time: 14:07)

The image shows a handwritten frequency distribution table and calculations on a whiteboard. The table has three columns: 'Range', 'f_i', and 'N = Σ f_i'. The 'Range' column lists intervals from 0-1 to 9-10. The 'f_i' column lists frequencies: 7, 10, 15, 20, 5, and 2. The 'N = Σ f_i' column shows the cumulative frequency calculation. To the right of the table, there are handwritten calculations for the midpoints of each range: 0-1 ⇒ 0.5, 1-2 ⇒ 1.5, 2-3 ⇒ 2.5, 3-4 ⇒ 3.5, 4-5 ⇒ 4.5, 5-6 ⇒ 5.5, 6-7 ⇒ 6.5, 7-8 ⇒ 7.5, 8-9 ⇒ 8.5, and 9-10 ⇒ 9.5.

Range	f _i	N = Σ f _i
0-1	7	7
1-2	10	17
2-3	15	32
3-4	20	52
4-5	5	57
5-6	2	59
6-7		
7-8		
8-9		
9-10		

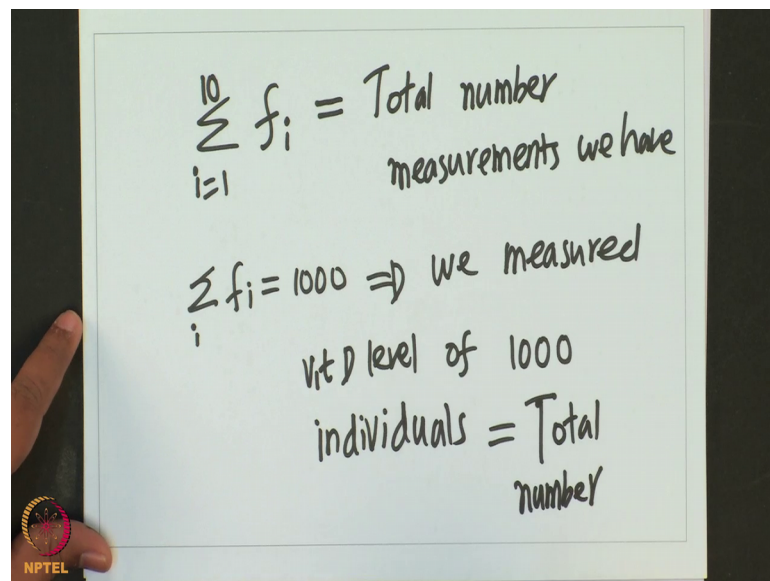
So, let us say, the range you could even take, let me take some measurement which comes between 0 and 10. So, the frequency is f. So, let us take here whatever between 0 and 1. So, the range would be 0 to 1, 1 to 2, 2 to 3, 3 to 4, 4 to 5, 5 to 6, 6 to 7, 7 to 8, 9, 10 these are the range and corresponding to this range I would have a different numbers. So, now, this I would represent by a number here I would for example, represent by the middle value of this let me call this 0.5, 1.5, 2.5, 3.5 and so on and so forth 9.5.

So, I have instead of 0 to 1 range I just put a value 0.5 here that is what I put here instead of 1 to 2 I could represent this by a middle value which is 1.5. This is one way of representing 2 to 3 I could put a middle value 2.5, 2.5, which means around 2.5 how many numbers are there within this range 0.5. Around 3.5 how many numbers are there within this range of 3 to 4 which is half 0.5 up or down.

So, now you could say there are 7 numbers between 0 and 1, there are 10 numbers, there are 15 numbers, there are 20 numbers and then there will be like 5 numbers and there

will be like 2. So, you could get a frequency like this for every range f_i versus i . Once you have this frequency distribution what 2 let us understand what are the two things that is meaningful. So, sum over i f_i what does that mean? So, f_i is the frequency how many numbers are between 0 and 1, how many numbers are between around 1.5, how many numbers around 2.5, how many numbers around 3.5, when I say around here I precisely mean between 3 and 4, 3.5 plus 0.5 or minus 0.5 in that range of 0.5 around 3.5 how many numbers are there. That is the thing that I can represent for example, here.

(Refer Slide Time: 16:55)



Handwritten notes on a whiteboard:

$$\sum_{i=1}^{10} f_i = \text{Total number measurements we have}$$

$$\sum_i f_i = 1000 \Rightarrow \text{we measured vit D level of 1000 individuals} = \text{Total number}$$

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

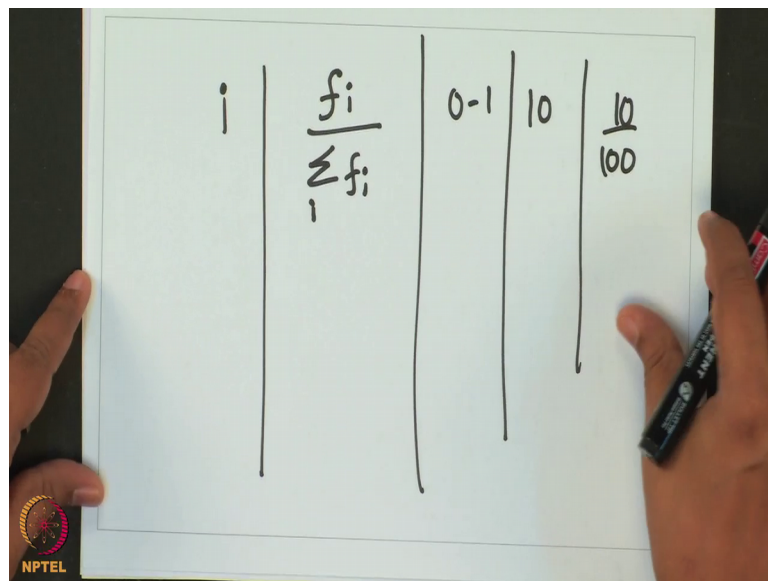
Now, let us think of a meaningful quantity here which is sum over i 1 to 10 in this case there are 10 range values. So, if there are 10 of them sum over i f_i this gives us the total number of measurements we have. If you have 100 cells sum over f_i will be 100, if there are 1000 cells sum over f_i will be 1000, if there are 1000 individuals we are measuring blood pressure of 1000 individuals some over a f_i will be 1000.

So, let us say sum over i f_i is 1000 this implies, we measured some quantity let us say vitamin D level of whatever the measurement we do 1000 individuals if they whatever be the measurement we this what it means. So, this is first meaningful quantity sum over i f_i is 1000 which is the total number total number. So, there are two numbers here the range which is how many, there is how many the serial number 1 2 3, 1 2 3 so on and so forth up to whatever be the number 10 or 11. So it will be let us say 10 here then this is

let me call this as my little n , little n is 10 here and there is capital N which is sum over i f_i . So, there are two n s here little n is this 10 and some capital N sum over f_i .

Now I can divide this whole thing by the capital N . So, I can plot i in the X axis, i verses f_i divided by sum over i f_i . So, that is I instead of saying between 0 and 1 I had 10 numbers this is my f_i and let us say I am doing experiment on 100 individuals I could divide this by 100.

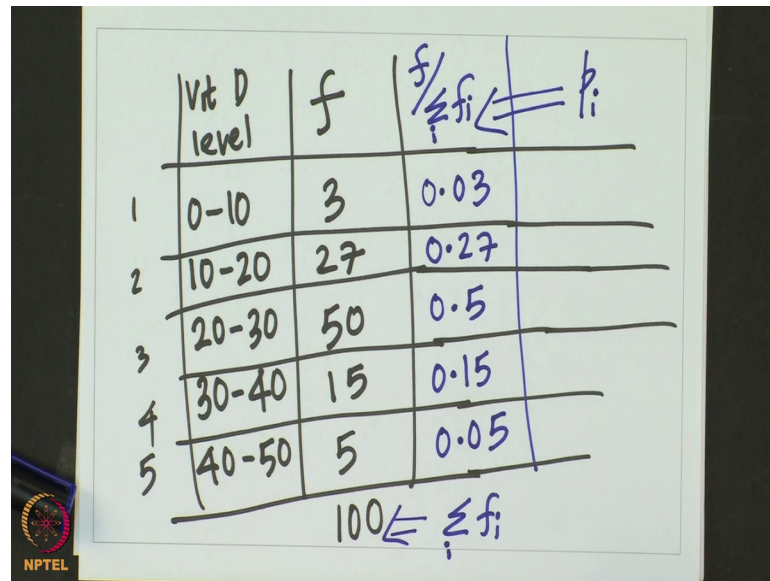
(Refer Slide Time: 19:15)



i	f_i	$\sum_i f_i$	$\frac{f_i}{100}$
0-1	10		

So, let us take an example to do this. So, let us consider 100 individuals and we measured the vitamin D level for this 100 individuals and we got some value. So, let us take an example of measurement of some medical let us say vitamin D level. So, we are going to plot vitamin D level for a population.

(Refer Slide Time: 20:10)



The image shows a handwritten table on a whiteboard. The table has three columns: 'Vit D level', 'f', and ' $\frac{f}{\sum f_i} = p_i$ '. There are five rows of data. Below the table, the sum of frequencies is calculated as $\sum f_i = 100$.

	Vit D level	f	$\frac{f}{\sum f_i} = p_i$
1	0-10	3	0.03
2	10-20	27	0.27
3	20-30	50	0.5
4	30-40	15	0.15
5	40-50	5	0.05

$\sum f_i = 100$

So, we have a population and the first one is vitamin D level and this would be in some typical unit that one would get from a lab. So, you in the some unit typically this could be around 25 let us say vitamin D level is around 25 units. You would ask the question how many people have vitamin D level between 0 and 10 units, how many have vitamin D level around 10 and 20 units, how many have vitamin D levels around 20 and 30 units, how many have between 30 and 40 units. So, and you could ask up to 50 let us how many have between 40 and 50.

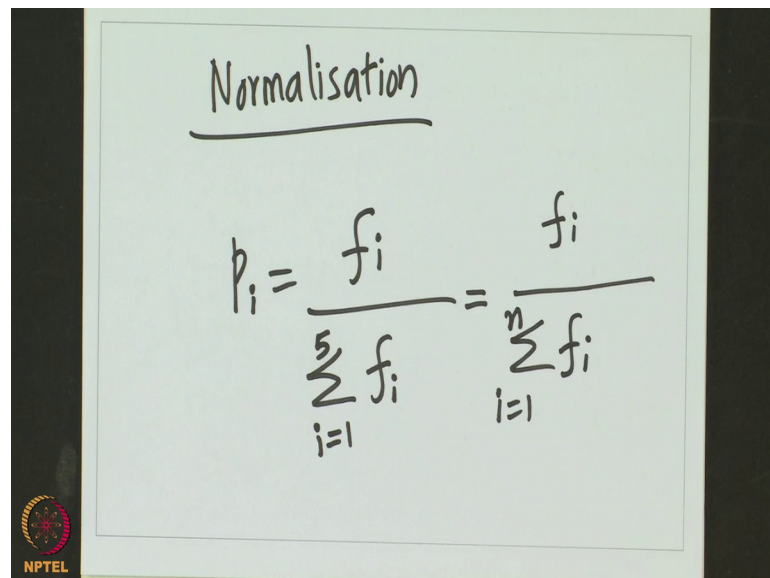
So, this is a range of vitamin D level in appropriate units and the frequency f for each of this range would be given by this, given by the second column. So, the frequency for 0 to 10 let us say 3 people have between 0 and 10, and let us say 27 people have between 20 and 30, 10 and 20 and let us say 50 people have between 20 and 30, so this is 77 plus 3, 80, and let us say 15 people have between 30 and 40 and 5 of them have between 40 and 50. So, let us say this is a measurement that we got from a survey from a population.

So, that sum of this 5 plus 15, 20, 70, 73 plus 70 plus 27, 97. So, this is 100. So, this is 100 and there are 1 2 3 4 5 ranges we have and we have 100 people. So, we are doing experiment on 100 people. Now we can divide. So, sum over i f_i is 100, sum over i f_i . So, what is this? This is sum over i f_i , sum over i f_i this is 5 plus 15 plus 50 plus 27 plus 3 is 100.

Now, I can make a new column here which is f divided by sum over i f_i you know it was f divided by 100. So, we can divide each of this number by 100. So, what is that mean? So, this means 3 by 100 which is 0.03, 27; this is 0.27 this is 50 divided by 100 which is 0.5, 15 divided by 100, 0.15 and 5 divided by 100, 0.05. So, this column this thing we would call as probability, this thing we would call as p_i which we would call as probability. This is for all this purpose of this course we will define probability in this way there is some other mathematical concepts go into this which we would not discuss for this particular course.

But for the practical purposes the frequency divided by sum over f_i f divided by sum over f_i is p_i . This is an important thing to remember this thing will call normalization, this process that we did this a this is a very important thing normalization which is basically say calculating p_i is equal to f_i divided by the total. So, here in this case i is equal to 1 to 5 f_i . So, this was in other words if there are N range, if there is N range you should calculate f_i divided by sum over i 1 to small n which is the range f_i this quantity is the probability.

(Refer Slide Time: 24:04)



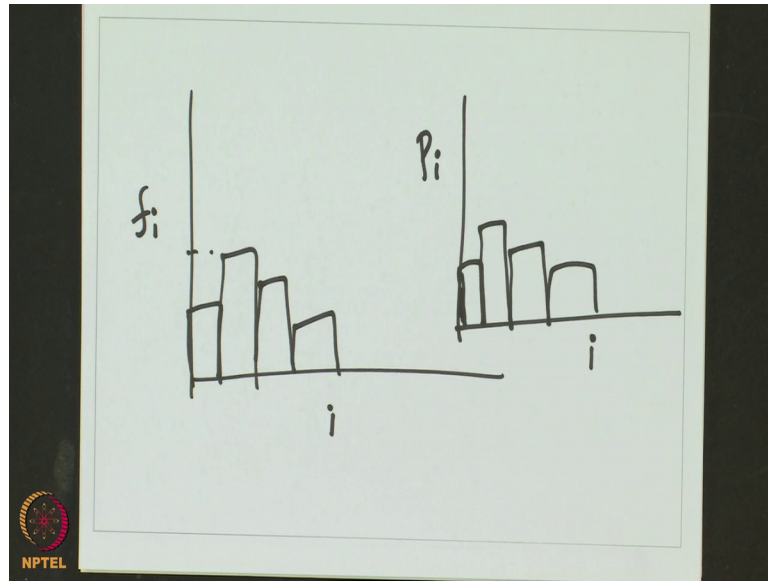
The image shows a handwritten formula on a light green background. At the top, the word "Normalisation" is written and underlined. Below it, the formula for probability p_i is given as:

$$p_i = \frac{f_i}{\sum_{i=1}^n f_i} = \frac{f_i}{\sum_{i=1}^n f_i}$$

In the bottom left corner, there is a small circular logo with the text "NPTEL" below it.

Now I can instead of plotting f_i I could also plot p_i . So, I could plot p_i versus i just like, either I could plot f_i versus i , and you would get some graph similarly I could also plot p_i versus i . So, you would get some distribution like this. Similar it would look very similar here to, this would look very similar just that the y axis numbers have changed.

(Refer Slide Time: 24:55)



So, here if this was, if whatever be the number here this would be divided by some constant, this is the total. So, this would look very similar, but the y value is changed. But this is very important to do this normalization because let me just come to this let us come to this and think about how many people have vitamin D level between 20 to 30 and the answer here is 50. And when I say 50 people have vitamin D level between 20 to 30 when I say this, this number alone does not make much sense unless I know how many total people are there, how many measurements did we do, how many people did we have in our sample and unless we know that the number 50 alone does not make sense. So, this normalization is always very important.

In some other a context if you think of it, let us say you took some examination and if I just say that 1000 people got, 1000 people failed when I say 1000 people failed it depends on how many people we have if they have 1 million people and 1000 people fail that is very different from having 2000 people and out of that 1000 people failed. So, this is important to know how many people are there in total and divide that by that. So, this normalization process is very very important in statistics. If you do not normalize it may not have meaning always. So, always is good to plot p_i as supposed to f_i , always good to plot p_i and not really f_i because normalization is very important, p_i will have more meaningful; p_i will have some more meaning in the sense that we could interpret it more accurately the interpretation will be more accurate.

So, I urge all of you to plot always p_i as much as possible and interpret things from p_i and in the next lecture we will discuss how to interpret things from the p_i and of course, f_i and p_i are essentially the same thing, but if you report is always report p_i and f_i can be sometime the typical value plugged out of f_i can be misleading therefore, always do plot p_i .

Now, with this I would summarize probability distribution tells us how the numbers are distributed. It is always important to remember what is ensemble population and for what population are we drawing this probability distribution. Think about it and we will discuss more about this in the coming lectures, but always compute p_i and plot p_i which will have meaningful quantity we can interpret it and compute some meaningful quantities from p_i .

With this we will stop this lecture and continue in the next lecture. Bye.