

Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute of Technology, Bombay

Lecture - 02
Data representation and plotting

Hi. Welcome to today's lecture. So, I hope you have done your; you know assignments and gone through the multiple choice questions which were uploaded. So, today we will start discussing about data and ways of representing data.

(Refer Slide Time: 00:34)

Descriptive Vs Inferential Statistics

Descriptive Statistics: summarizes important characteristics of a set of measurements

Inferential Statistics: procedures for making inferences about population characteristics from a sample drawn from the population



So, broadly speaking there are two components of statistics: one is descriptive statistics which essentially is summarized; that is to basically convert raw data into some numbers. So, that is what descriptive statistics is about. And the other type of statistics is called inferential statistics. Here, we want to develop procedures for finding out or making distinct conclusions from the measures that we have drawn, from the sample, from the population.

(Refer Slide Time: 01:04)

Steps in Inferential Statistics

- Specify question to be asked & identify population
- How to select sample
- Select sample & analyze the information
- Make an inference about the population



Determine reliability of inference

So of course, inferential statistics is the most important thing. So, there are few steps which we need to follow in order to understand, what are the steps in inferential statistics? So, the very beginning the first thing is to identify what is your question right what is your question and who is your population let us say you want to you know you want to make you want to market a soap then and for teenagers. So, what should be the look and feel of the soap? So, has to attract teenagers to using that. So, your population is the teenager the question is basically to make a soap of and identify the essential features of the soap.

So, now you want to have a process of selecting sample; so, you know it is teenagers, but teenagers from where you know what is the proportion of boys versus girls in the sample? So, once you have done that and you know we had discussed in previous lecture that if you are sampling is improper then you might lead to a completely wrong result.

Once you select the sample you have to analyze the information; you select the sample, you ask relevant questions in the course of a questionnaire and you analyze the response is given by you know boys and girls and based on that you want to make an inference that you can apply for the whole population of teenagers. And then finally, you want to determine the reliability of inference, you have come up with pink soap with you know which is more elliptical in nature or oval in nature is what people would want so, but you want to test the reliability of this inference.


So, these are the steps in inferential statistics, but before the; you know the prelude to inferential statistics is descriptive statistics and we want to begin with them descriptive statistics.

(Refer Slide Time: 02:58)

Variable

Variable: characteristic which varies with time and/or different individuals

e.g.: body temperature, height, weight, etc



Student	Gender	Year	Major	# Courses	CGPA
1	F	1st	Maths	5	7.4
2	M	2nd	Physics	9	8.1
3	M	2nd	Biology	10	8.2
4	F	3rd	English	18	6.9
5	F	1st	Chemistry	5	9.0

So, in descriptive statistics; one of the most important things is a variable, what is your variable. So, variable is a characteristic which varies with time and or different individuals. So, our body temperature can be a variable rate. So, you want to figure out whether someone has fever or not fever. So, body temperature is a variable in that case.

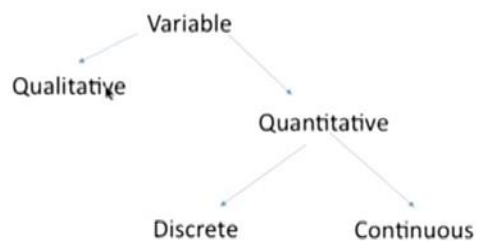
Someone you want to figure out, what is the average height of this population. So, then height; single you weight so on and so forth. This is just an example of you know data of let us say 5 students in a class. So, you have the following categories in other words, you have the following variables what is the gender what is the year in which the student you have selected the students 5 students from you know from the hostel which year their first year; second year so on and so forth, what are they measuring in with maths, with physics, with biology so on and so forth. How many courses have they already done?

So, a first year student would have taken probably taken 5 courses already; that mean, this is second semester of the first year, so on and so forth and what is the GPA of that particular student. So, what you see that the nature of the variable differs a lot. So, in case of gender, it is just a category, you either have male or female in year you have a number 1, 2, 3, 4, major is also categories, you have distinct categories; maths, physics,

biology so on and so forth number of courses is a variable, but it is a discrete variable, you can have only natural numbers which is greater than 0 and in terms of one 2 3 like that, but CGPA is a fraction, it is a number which is depending on what is your you know total CGPA. It can vary anywhere between 0 and 10 let us say, but you can have any variable which is between these numbers.

(Refer Slide Time: 04:56)

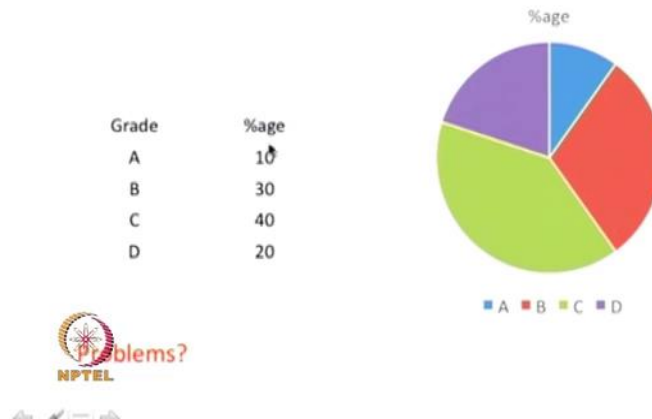
Types of Variables



In other words my variable can be divided into the 4 categories; your type of data can be qualitative. So, qualitative I mean that is a gender is for example, male or female or you can we have a quantitative variable which is essentially like CGPA or which here you are in. So, again which here you are in is a discrete variable and your CGPA is a continuous variable. So, there are various types of variables you have to identify depending on the problem.

(Refer Slide Time: 05:27)

Converting categorical data into plots: Pie Charts

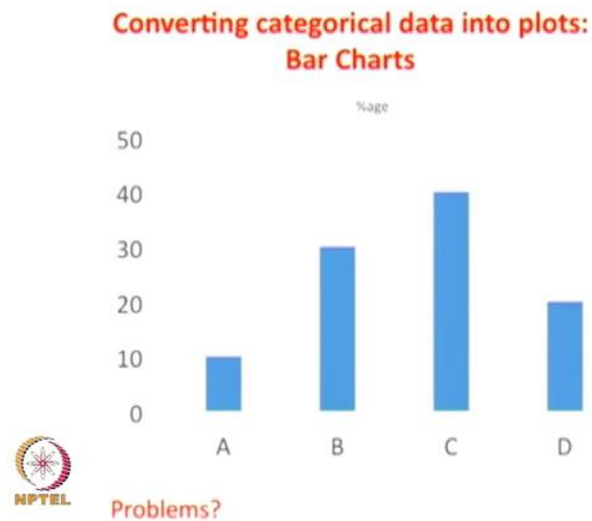


Now, let us say, we you know go back to another plot, where you have a grade. So, you have; you know the mid-sem exam is over and you have graded the students and you want to find out the statistics as to who has gotten what grades. So, there are 10 percent in the population which has gotten grade A; 30 percent in the population, grade B; 40 percent, grade C and 20 percent grade D. So, you can represent it enough what is very you know popularly known and used, it is called a pie chart it is attractive in nature. So, what you clearly see 40 is c and it has the biggest section of the pie chart.

So, the area of this pie chart is proportional to kind of the relative frequency of this number, but so, pie chart is easy to represent easy to understand, but it has its share of problems. So, we need to know, what are these problems? So, imagine in this case, there are only 4 grades. So, there are 4 categories, it is easy to come up with the pie chart imagine a situation where there are 25 different categorized you know category is possible.

So, in other words, each of these percentage areas will keep on shrinking and shrinking. So, imagine you have one case which is 1 percent and the other one which is 41 percent. So, 41 will of course, take a huge chunk of this pie chart, but 1 percent will barely be visible. So, in other words, you it is difficult to represent in pie charts when your volume of data increases such that there are multiple different category is possible.

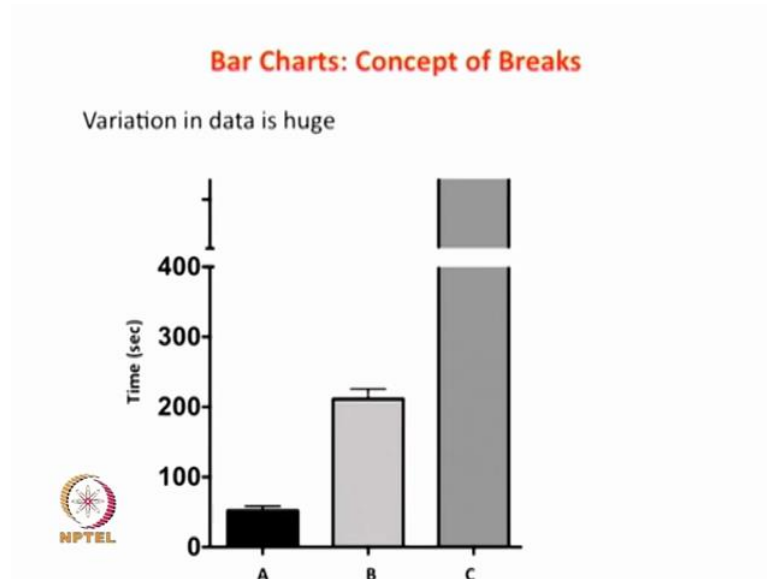
(Refer Slide Time: 07:04)



So, you can express this categorical data into although in something in another thing which is widely used is a bar chart. So, same as before you have the percentage in your y axis and you have the categories A, B, C, D. So, as before one of the weaknesses or deficiencies of these bar charts is that if you have too many bars it looks cluttered if you have few bars there you know it is easy to represent.

So, this is you know coming to the few bars and again the same problem that I mentioned before for pie charts. So, you have a value one which is 2 percent and another value which is 40 percent, how can you represent it in the same bar and still the you know the other person can make sense out of it the 2 percent for all practical purposes will look like 0. So, it is nearly impossible.

(Refer Slide Time: 07:55)



But there is a solution. So, what we do is when the variation in data is huge as is in this particular plot you have 3 categories A, B, C where a value is around 50 and C is maybe even you know 600.


So, what you can do is introduce something called a break. So, you want to show that there is significant difference between A and B. So, whatever is A you know range maximum of c till there you can have a continuous axis in y, but after that you can introduce what is a break. So, let us say this guy is 800. So, you can introduce a break at 400 and then plot again. So, everything still fits into the same thing, but the essential part of the information is there for you together that that this is way smaller than this is also part of the information and this is way smaller than this is also part of the information and you want to capture both these things in the same plot.

(Refer Slide Time: 08:53)

Working with quantitative data

Body Mass Index (BMI)

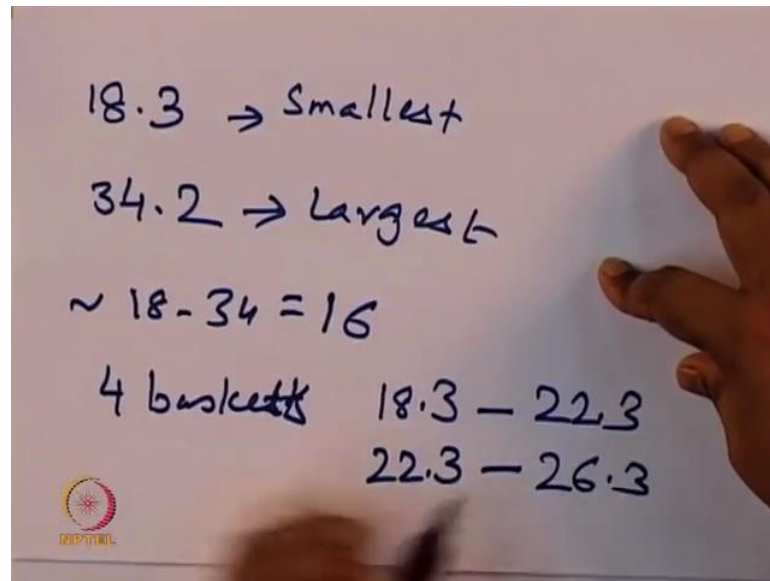
18.3	21.9	23.0	24.3	25.4
26.6	27.5	28.8	34.2	31.0
19.2	21.0	24.5	25.5	27.8
28.2	31.0	29.1	28.1	24.2
25.6	20.0	20.0	25.0	25.2



So, let us have a simple example, we are talking about working with quantitative data. So, this is the body mass indices of you know 25 people in a class, let us say you have this entire you know of course, these values are continuous variables so that you can have all these values.

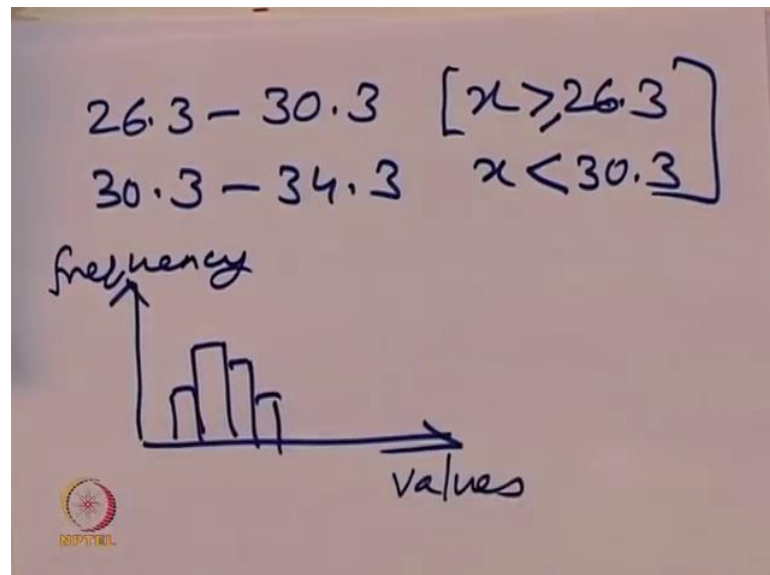
Now, we want to know, how can we convert it into a way of representing it? So, identifying categories A, B is perhaps not the good; you know good way because it is not a discrete quantity, but a continuous quantity, but what you can do is you can identify what is the range. So, in order to identify the range we want to know, what is the smallest value in this population? So, I can go through this list and I think the smallest value is 18.3. So, 18.3 is the smallest value and the largest value largest value is 28.8; 34.2.

(Refer Slide Time: 09:56)



So, 34.2 is the largest value this is smallest, this is largest. So, we can divide it. So, 18 to 34 is roughly 18 to 34 is equal to 30; you know 16. So, we can have a range of 4 baskets. So, we can identify 4 baskets, let us say 1 is 18.3 to 22.3 another is 22.3 to 26.3.

(Refer Slide Time: 10:37)



We can have another 1 which is 26.3 to 30.3 and 3.3 to 34.3.


Now, each of these numbers would mean that you in this basket, something will come in if let us say that number x is greater than 26.3 greater equal to 26.3 and x is less than

30.3. So, this would make sure the same point x does not go into multiple basket baskets. So, this way what we can generate is called a histogram.

(Refer Slide Time: 11:10)

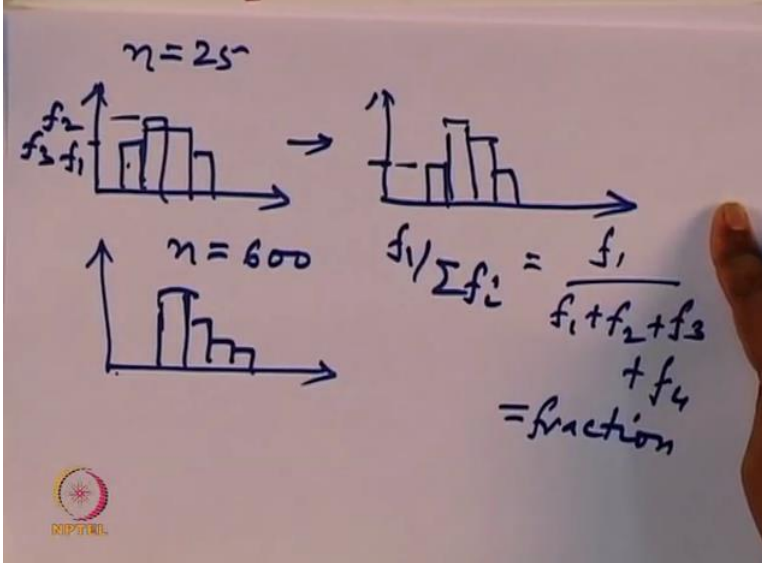
Histogram

- Convert data into frequency
- Can be plotted as numbers, or %age
- Unimodal Versus Biomodal Versus Multimodal distribution



So, you convert the data into frequency you can then plot them as numbers or percentage and then you can have multiple distributions depending on the nature of the data. So, your histogram looks something like this. So, you can have these bars. So, in our case we have 4 bars. So, we will have these distributions. So, these are values and this axis is frequency or the number of them. So, it is possible. So, it is possible to convert this data.


(Refer Slide Time: 11:47)



$n = 25$

f_2
 f_3
 f_1

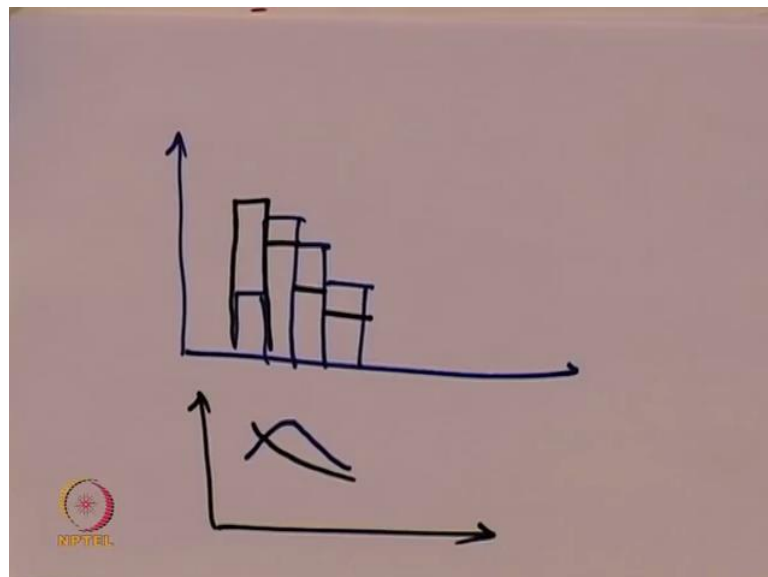
$n = 600$

$$\frac{f_1}{\sum f_i} = \frac{f_1}{f_1 + f_2 + f_3 + f_4} = \text{fraction}$$


Now, let us say you are going through this exercise you have this distribution, in one case where the total number of observations were 25 and another distribution where n is equal to 600. Is it possible to put both of these data on the same plot and this is where you have to do what is called as a normalization exercise. So, what is n equal to 25's total and each of these values frequencies?

So, you convert it you normalize the curve in other words you divide every; if let us say this is my f_1 , this is my f_2 , this is my f_3 so on and so forth, I convert them into fractions. So, the nature of the curve will not change. So, this value this value is now, f_1 by summation f_i . So, it is equal to f_1 by f_1 plus f_2 plus f_3 plus f_4 .

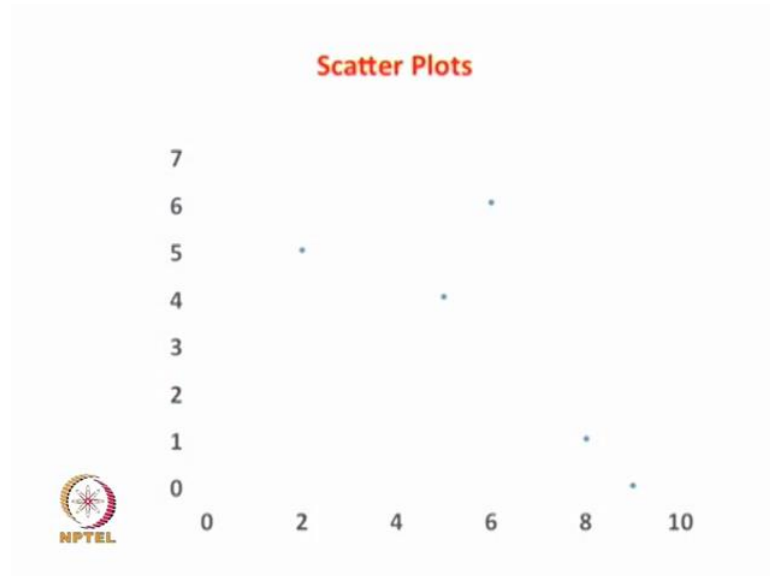
(Refer Slide Time: 13:04)



So, this you will get a fraction is a fraction. So, once we have done this then it is theoretically possible to generate the following plot I have the same thing and another one let us just say hypothetically. So, the way I drew is. So, if I just; if I were to draw the outlines of this curve these curve would look like this. So, you have one curve like this and the other curve which is like this. So, it is possible to plot both of them at the same time, but you have to do is normalized.

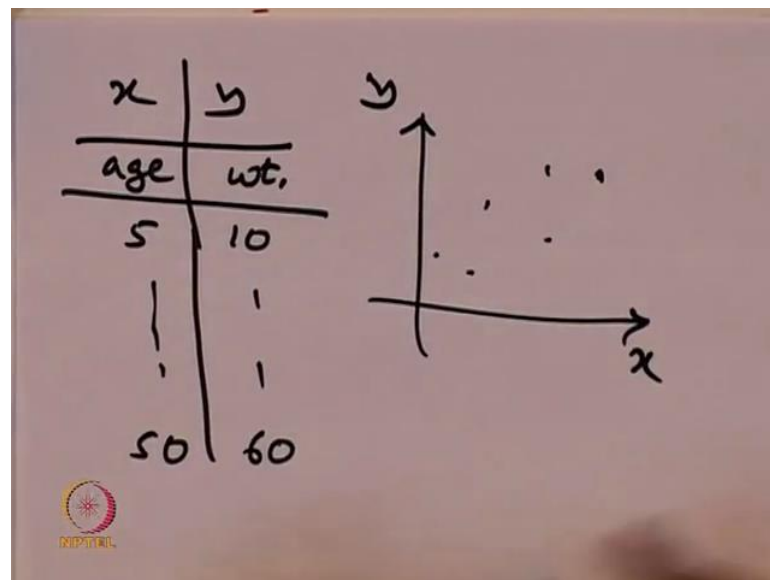
But another caveat of this is you must ensure that the data is from a similar distribution. So, any of course, there is greater certainty when you have sampled 600 individual measurements, but when you are you know plotting the same thing with n equal to 25, this is the great possibility that the nature of that distribution will shift.

(Refer Slide Time: 14:19)



So, another way of; another type of plot which is widely used is called scatter plot. So, scatter plot is just x and y values.

(Refer Slide Time: 14:24)



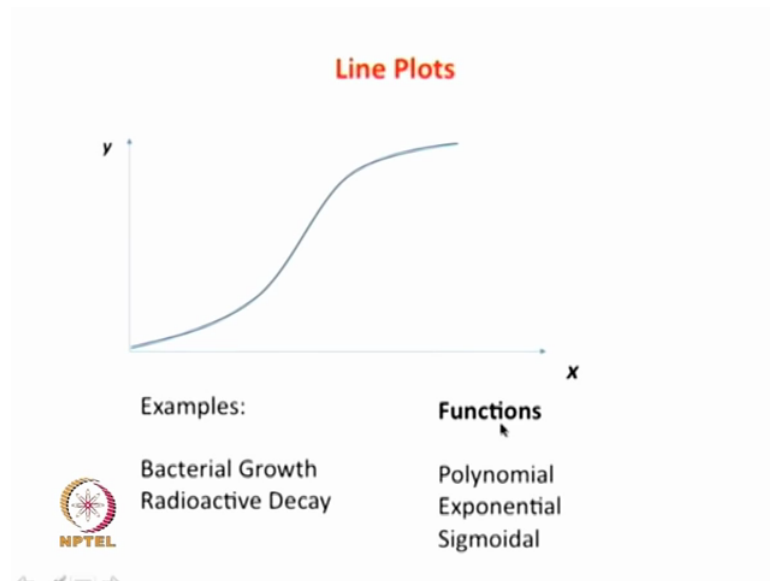
Let us say I have x versus y I have age as 1 variable and the other variable is let weight. So, I can have this generation age is varied in all, it is a 5 years weight is 10 kg's so on and so forth and 50. For 50 years age is you know 60 kg's. So, you have a range.

Now, depending on the nature of this data you might have a, you know points which look like this. So, this is my x, this is my y. So, you might have data which looks like this

or as I have plotted in this particular curve you have a kind of a reverse such association where you have greater the increase in x the y value decreases with a notable exception. So, this is where your; you know data analysis. So, you know in this case do you call it a negative association or do you want to have a much more non-linear nature of this curve.

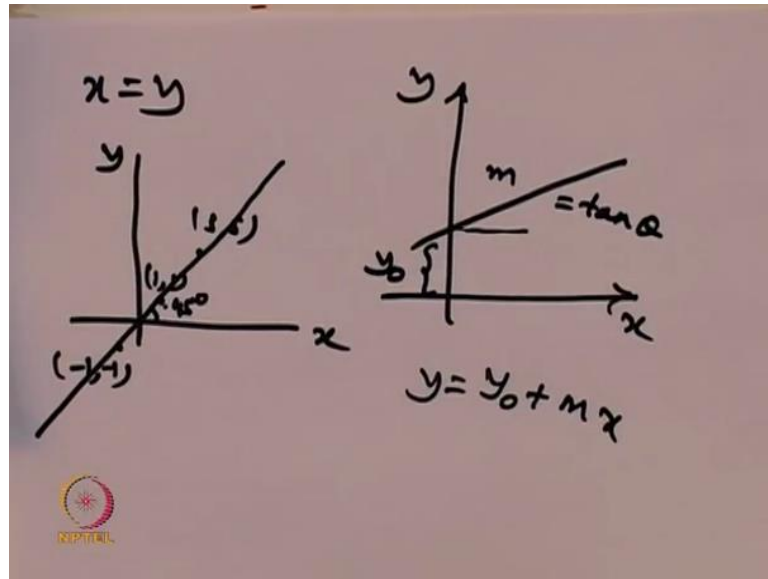
So, scatter plots are widely used. So, again here it is better to plot these points as scattered as oppose to connect them then it is much difficult to make sense out of this data.

(Refer Slide Time: 15:35)



But you can mix if you were to connect it then you can generate what are typically called as line plots. So, this is an example of a line plot where x and y , I have plotted it in a slightly which looks like a; you know s in some way. So, these are reminiscent of bacterial growth curves, but you can have various functions which describe these line plots. So, it makes sense to connect them with line when you know that the underlying phenomena is actually a physical process which has a given time constant associated with it or a given you know mechanism by the way in which it happens. So, there is it is under control it is not a completely random association.

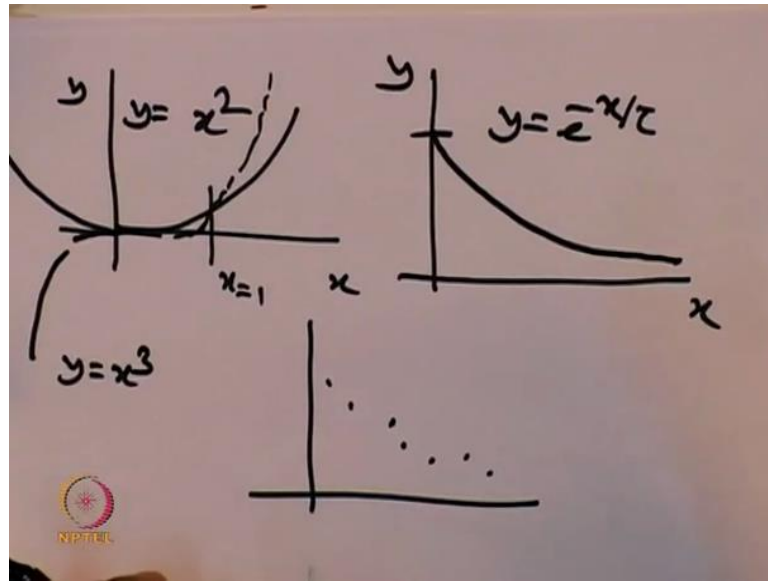
(Refer Slide Time: 16:25)



So, that is when you can have very nice linear plots. So, just a small detour on the type of plots you know you are all well conversant with linear plots x equal to y is a very simple plot and you know how to plot it you have x you have y you take these points and you know you take these points. So, let us say this is 1 comma, you have minus 1 comma minus 1 you know 5 comma 5 so on and so forth, you have a line which goes like this and this is 45 degrees.

In general if you have a line which kind of shifts up. So, in general why you know you can have a line which is like this in this case. So, there is an intercept a nonzero intercept on the y axis which you can call at y_0 and it has a given slope. So, you can have m as the slope or m is nothing, but $\tan \theta$. So, in this case y is equal to y_0 plus $m x$ is your equation. So, you can have multiple types of you know functions these are all linear functions.

(Refer Slide Time: 17:28)



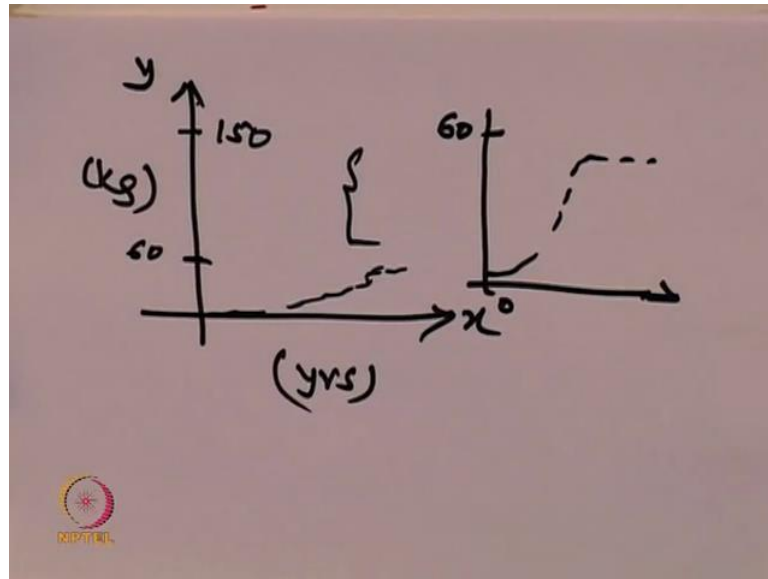
That I have drawn you can have something like this let us say this is an example of a parabola. So, this is x this is y and y is equal to let us say x square.

So, the far higher you are you have a non you know non-linear nature of the curve. So, these are. So, y is equal to x cubed will look similar, but it will it have much sharper peak be before x equal to one and lower peak before this, but y is equal to x equal to x square is symmetric, but y is equal to x x cubed looks like. So, y is equal to x cubed looks like this when x is negative your y values are negative.

So, these are some of the simple curves in polynomial you can have exponential curves which are let us say an exponential decay curve will looks like this let us say x this is y at. So, if it is y is equal to e to the power minus x in terms of decay you have at x equal to 0 you have y equal to one and then you have a characteristic time. So, in the most general case you have x by tau which you know which represents the time constant of the (Refer Time: 18:29).

So, when you are trying to fit data let us say you have a data which looks like this then it should immediately occur to you that this has something it, might look like an exponential it might be a pair you know it might look like a polynomial. So, polynomials are easy to fit because they have multiple dimensions, but it is not you know why is to always fit every function with a polynomial.

(Refer Slide Time: 19:01)

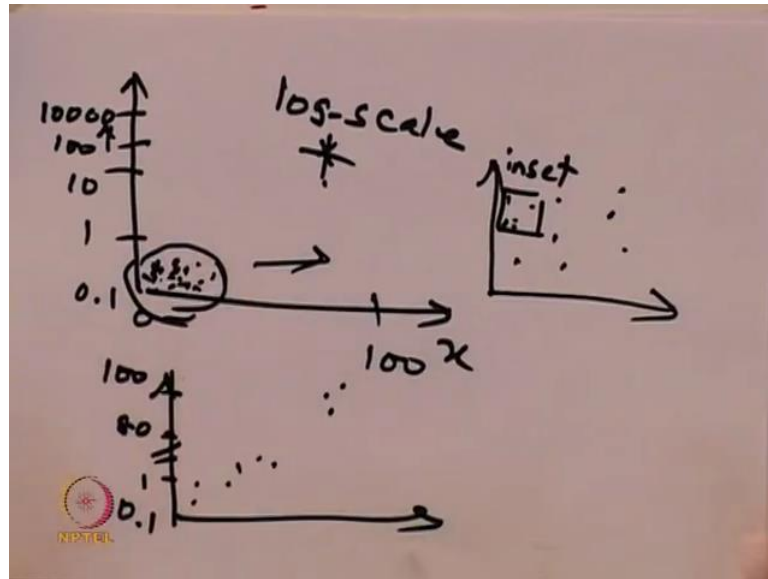


Now, what are the things that you need to keep in mind while doing these plots let us go over them one by one. So, of course, when you make a plot first thing after label your variables x and y you have to put their units ideally. So, let us say this is if is age then I can have years in my; if this is weight I can have kg. So, I need to know what is the what are my axis and what are my units and I need to choose the appropriate range let us say for example, we are we I want to make a plot of population expansion. So, if I plot like this right it gives me the impression. So, sorry; this is kg now let us say it is a weight itself weight which is increasing as a function of years which will also probably be a linear you know increase and then some saturation after point.

So, in this case if I want to show it is linearly increasing. So, let us say this maximum value is around 60. So, I need to make sure and I am plotting till one 150. So, this portion of my plot is completely destroyed, because I am not using the space I am I am visually trying to convey that the weight is not changing much with years, but in reality the weight is changing with years. So, I should actually redraw this plot that this is from 0 to sixty and my curve should look something like this.

So, if it was like sixty then I can clearly see there is a non-linear increase initially which means that initially when kids are growing their weight increases drastically, but once they reach a certain age it starts to kind of plateau off.

(Refer Slide Time: 20:44)



Again the other point of breaks as I said, in whatever you have done in bar graph you can have the break here itself again let us say you have a variable x which goes from 0 to hundred and a variable y which goes from point one to 10,000.

So, here if you if you put a linear value. So, also all values which are very small will look like this just look like a mess here, but what you can do is you can either plot it in log scale. So, if you plot it in log scale then accordingly every point. So, this will be one this will be 10000 so on and so forth. So, the points will be well separated out and you can see them. So, it is important to choose appropriate range and all again as before let us say this is from point one to hundred what I can do is I can introduce a break. So, let us say I can have point one to one and then 80 to 100 if all the data is just here and then remaining is here. So, this is how I can really make use of the whole plot and still plot my axis. So, that everything is clearly visible.

One more thing is let us just say that you have all your data is here and there is one outlier which is here all your data is reasoning is essentially concentrated in this portion of the curve, but there is one point which is way out which is an outlier. So, do would you bother to plot the entire range or would you just bother to point this plot the scenes I think it is it makes sense to plot the centre then and then blot it up. So, you make it big. So, then you have all the scattered and as an inset you can have this higher value where

all these points are looking the same. So, make this whole curve as the inset. So, this is called a inset to handle outliers.

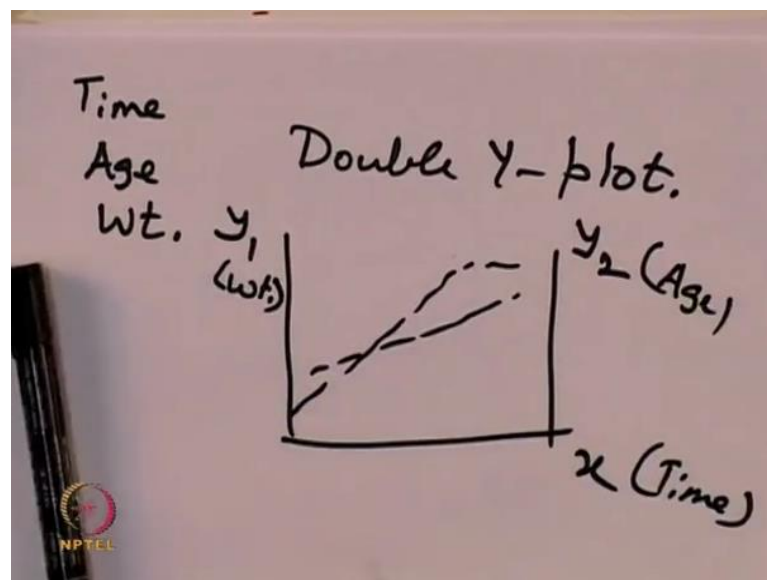
(Refer Slide Time: 22:44)

Things to keep in mind

- Label axes
- Units (choice of units, e.g. cell speed)
- Choose appropriate range
- Large variations in data – breaks & insets

 • Outliers

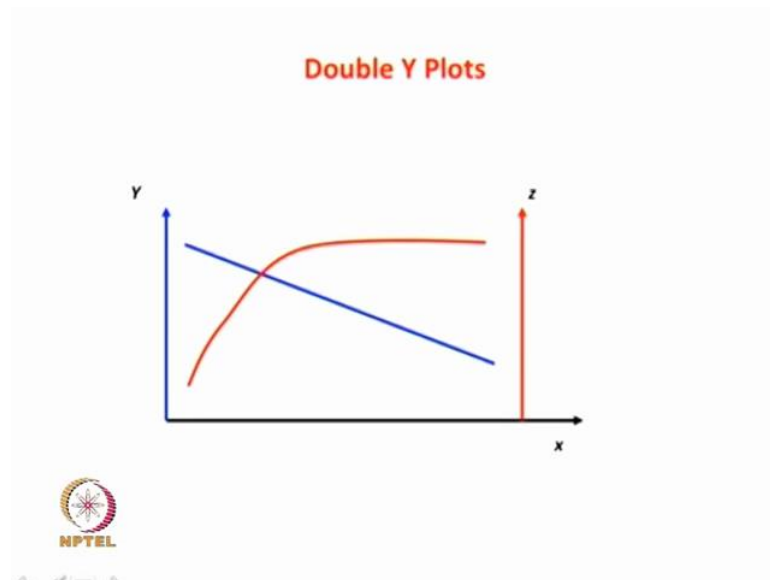
(Refer Slide Time: 22:53)



So, now these are some single y plots again let us just say that I have 3 variables right let us say I have 3 variables time age and weight 3 variables and I want to understand and I want to make a single plot of putting all of them together. So, this is where you can make use what is called as a double y plot. So, you can have 2 axis. So, this is you can label this as y one axis this is as y 2 axis this is x and you can plot them let us say with you

know weight with time might saturate and age with time has a; you know linear relationship. So, this if x is my time this is my weight and this is my age then this guy will have a linearly increasing curve.

(Refer Slide Time: 23:48)



So, this is just another example of a double y plot. So, in this case I have had a reverse weight in which variable y exhibits a decrease a linear decrease with x as a function of time and variable x actually exhibits a saturation profile. So, beyond a certain value of x it reaches the saturation.

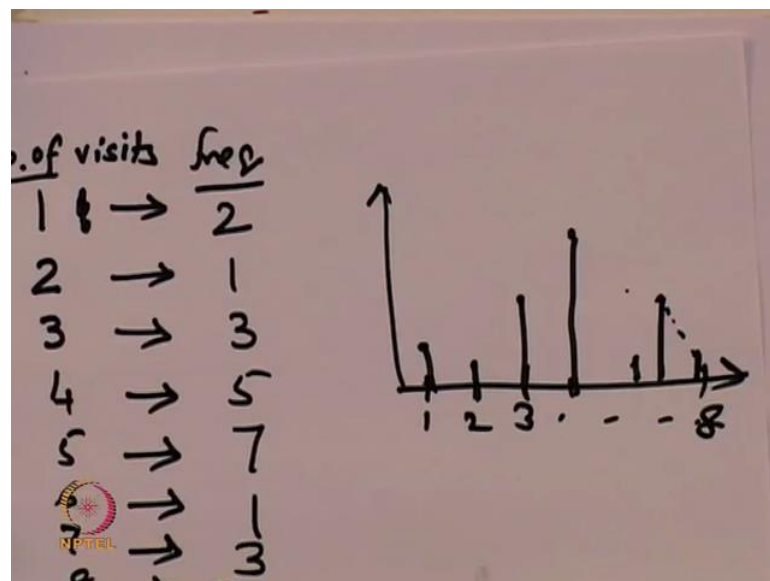
(Refer Slide Time: 24:08)

The figure is a table titled "Example 1" with the caption "Number of visits to a dental clinic in a typical week". The data is arranged in a grid with 5 rows and 5 columns. The NPTEL logo is located in the bottom left corner of the slide.

Number of visits to a dental clinic in a typical week				
6	7	5	1	8
4	9	3	3	4
7	2	1	4	5
5	5	5	5	7
3	4	4	5	8

So, now let us solve few examples. So, we have the following example where you have number of visits to a dental clinic in a typical week. So, as you can clearly see. So, these numbers are all discrete numbers you do not have a fraction because the number of visits is of course, a discrete number, but and you want to know what is the best way of plotting it. So, first you see what is the range. So, we have all the way from 1 to 8. And I think when you have this kind of data it is good to sort it. So, if I were to write the same data together in a sorted form.

(Refer Slide Time: 24:45)



I have 1, the frequency of 1; I can make the frequency of 1. So, I have 1, 2, 3, 4, 5, 6, 7, 8 and this is my frequency axis we have the number of visits and the frequency axis. So, for number 1 the frequency is 2 the number 2. So, the frequency of 2 is only 1 frequency of 3 is 1, 2, 3, frequency of 4 is 1, 2, 3, 4, 5, frequency of 5 is 1, 2, 3, 4, 5, 6, 7, frequency of 6 is only 1 frequency of 7 is 3 8 is 1 2.

So, you have the number of visits because these numbers are small there is absolute. So, of course, this axis has to be 1, 2 like that 3 and because these numbers are small there is absolutely no necessity to make it into a relative score you can just have these values. So, for example, for 1 it is 2 for 2, it is 1 for 3, it is 3 for 4, it is for 4 it is 5, 6, it is 17, it is 38, it is 2. So, you have if I if I connect that actually I should have them as bars. So, what you see is, almost they it is not a uni modal distribution there is a reasonable amount of

variation in the data. So, if this was if this 5 was slightly higher then you have a nice histogram like shape, but this is different.


So, this is of course, the discrete variable and then you can I think histogram would be the easiest way to plot it. So, let us and you know histogram is the way to plot it.

(Refer Slide Time: 27:12)

Example 2

Test Scores of 20 students

61	49
93	74
87	66
42	45
55	68
67	88
82	59
50	71
29	21
55	50

 Test Scores of 10 students in 2 exams

Let us take another example. So, I have test course of you know 20 students, I have test scores of 20 students and I want to know that what is the average test score as before. And what is the way of plotting it as before? I think the histogram is the best way of plotting it. So, we can again go through the same process we know what is our lowest number which is around 29 which is our highest number which is around 93 and we can make it into a histogram.

So, again where histogram is a good way of representing this; the last one the last one is an example. So, imagine. So, the above data is not test scores of 20 students whether it is test scores of 10 students in 2 exams; so 10 students: exam 1, exam 2. So now, we have to plot this you know. So, you can make them as 2 separate histograms, but if you want to plot it in the same plot maybe it is best to put it for the for each student the x and the y and that might give a some correlation between how they performed in each of the exams.

So, I guess that brings us to the end of this. Just a brief recap we discussed you know the nature of variables either qualitative or quantitative, we discussed some of the common ways of representation which is pie, chart bar, chart histograms, line or scatter plot and then double y. So, depending on the nature of the data the range the size of the data you might choose to you know use the histogram or the scatter plot as the case. Maybe when you are trying to look for some correlation you want to preferably use thoughts like scatter plot.

With that I thank you for today's lecture. And I hope that you attempt the questions which we upload for the multiple choice questions.

Thank you.