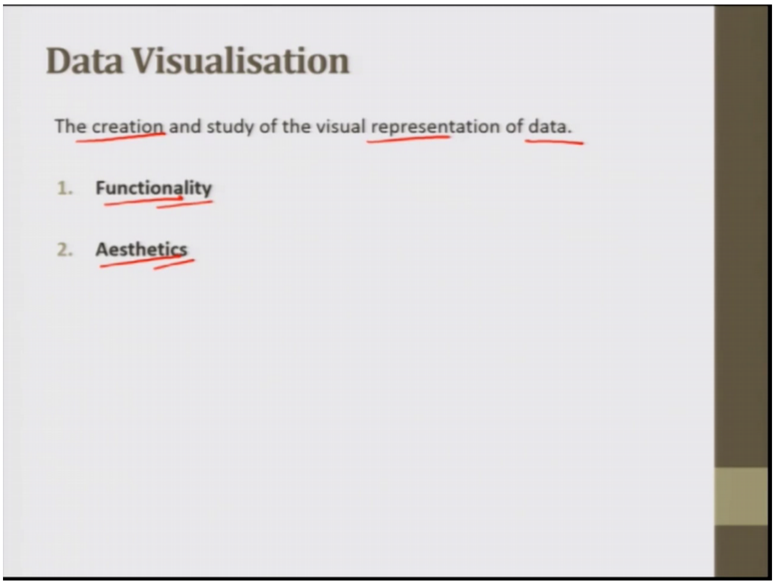
Engineering Metrology

Prof. J. Ramkumar
Dr. Amandeep Singh Oberoi
Department of Mechanical Engineering & Design Programme
Department of Industrial & Production Engineering
Indian Institute of Technology, Kanpur
National Institute of Technology, Jalandhar

Lecture - 40
Statistics for metrology, fundamental concepts (Part 2 of 3)

Next past of this course is Data Visualisation. Now, we discussed what is data? What are data scales? What are the very where some statistics, and how do we describe the data descriptive statistics we discussed? In descriptive statistics we also took one point the forth-point and the last point was data visualisation, which is the graphical representation of data.

(Refer Slide Time: 00:45)



Data Visualisation

The creation and study of the visual representation of data.

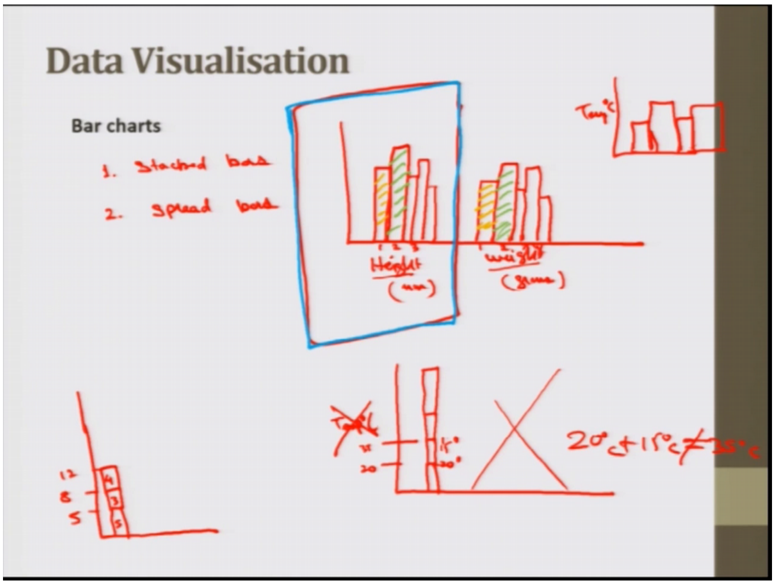
1. Functionality
2. Aesthetics

Why do we need to present the data in a correct form; that certain reasons for this. The creation and study of visual representation of data is known as data visualisation, this is actually definition of data visualisation number one is functionality. There is some function that data is tried is trying to accomplish. The data we need to see the true value or the true value or the true value actually the values are random values which have we have observed, we need to find the average which should be closed true value.

If we having different samples and we have need to find the averages of all this we need to see the variance, there is some function associated with that ok. Functionality is the basic or the fundamental use for what purpose is data being used. We can draw histograms, we can draw scatter plots, what and where to draw, what are the ways applications and which is not recommended at what place we will discuss these things ok.

Now, next is a aesthetics. As, you know in data visualisation we also plot the data in different colours, different sizes, we can show different we can also cheat with scale cheating with scales I will discuss that. So, all those things is aesthetics.

(Refer Slide Time: 02:03)



Data Visualisation

Bar charts

1. Stacked bars

2. Spread bars

Number, first plot is the bar charts. Bar charts are just as we discussed these are the bars like these ok. If, I say if I say let me say height and weight; I can have different colours here I will use it here this is bar 1 ok, this is bar 2, different components are there and I am discussing about the height and weight. I can just see that the height of the components 2 if it is larger than weight is also higher we can find the correlation.

I can just think of the some correlation between them, but bar charts is the very commonly used graphical representation. And we can have 2 types of bar charts, what those are stacked bars and spread bars; and spread bars ok. These are actually stacked bars, we can have sample number 1 2 3 by; we call it item number item number 1 2 3 4. These are known as spread bars I can have number of items from 1 2 3 number of observation 1 2 3 this is spread here ok. This is spread here like this, this is spread here like this weight is also proportional ok.

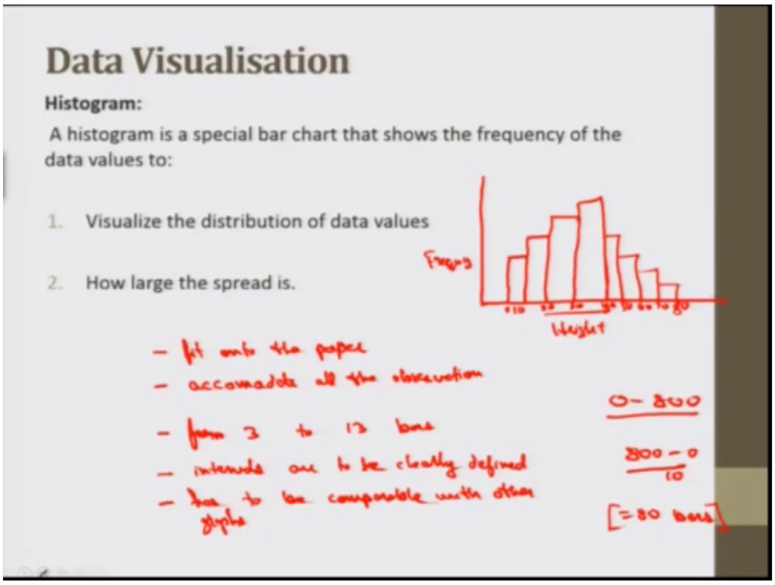
So, I can use a height may be is in mm weight is in grams. However, this is not a very proper way of representation of 2 different entities on 1 scale but yes, if I even use this only, if I even use this only ok, it is also a bar chart. This is also a bar chart ok. And, this is a spread bar. Stacked bar is when we are stacking 1 bar over other, when we are stacking like this ok.

Now, we can just plot the temperatures different temperatures, when we send if the scale here is temperature like if this is temperature higher temperature degree centigrade I will use a spread bar like this ok. That to show that a temperature is varying with a more appropriate for plot for temperature would be the line diagram, but the temperatures cannot be like this. If, I put temperature degree centigrade here this is not recommend we cannot say that the temperature 20 degree plus temperature another 15 degree. This is 20, let me say this 35, this is 20 degree total this is 15 degree that we makes it to 35 degree. Now, 20 degrees of temperature plus 15 degrees of temperature is not equal to 35 degree centigrades ok.

So, this is not recommended here. However, if the heights or I will say if the number of pieces that we have inspected, and we are keep note we are just making the note of the pieces those are those are being inspected or those are being checked in each hour ok. In first hour I will let me say inspected 5 pieces. In the next few hours I selected I inspected further 3 pieces that makes it to 8. In the next few hours I inspected 4 pieces that makes it to 12 ok, this is 4, this is 3, this is 5.

In that case stacked bar can be used ok. So, we have to be very careful that which kind of chart which kind of graphical tool always using for our data. We just cannot pick anything and put any data over there.

(Refer Slide Time: 06:44)



Data Visualisation

Histogram:

A histogram is a special bar chart that shows the frequency of the data values to:

1. Visualize the distribution of data values
2. How large the spread is.

- fit onto the paper
- accomodate all the observation
- from 3 to 13 bars
- intervals are to be clearly defined
- has to be comparable with other graphs

0 – 800

$\frac{800-0}{10}$

[= 80 bars]

So, next is this histogram. Histogram a special type of bar chart that shows the frequency of the data values: number 1 to visualise the data distribution of the data values. And number 2 to know that how large the spread is histogram is kind of a bar chart as we discussed in the frequency distribution, we have some interval here ok.

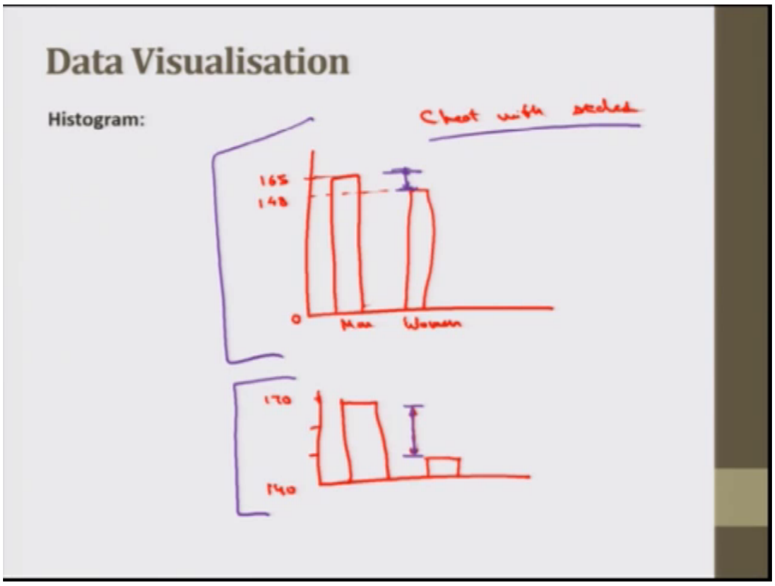
Let me say if this is again height, height in this direction and we having frequency in this direction is kind of a frequency distribution again ok. Let me say heights from 0 to 10 ok. This is let me say 0 to 10, again I am having from 10 to 20, then put 20 to 30 and so on. If, I am having such kind of plot, this is known as histogram, 70-80, let me say the value is 0 maximum value is 80 here ok. This might be the height of any components that we are trying to see, they are such a guidelines to draw histogram; we cannot just divide the intervals into any number we like ok.

If, we there are 80. So, it seem instance let me say if they are 800 height varies from 0 to 800, then what do you think, should have the interval of 10 if it is from 0 to 800 and if I make an interval of 10 ok. The total would be actually 800 minus 0 by 10 this is equal to 80 bars. So, it is not very legitimate to draw 80 bars in 1 chart. So, there certain guidelines in this number 1, I can put here is a histogram has to fit onto the paper, fit on to the paper if I drawing it on the paper.

If I am even using software’s, then also the 8 bars is not a very great idea to use. Then next point I can put here is it should accommodate all the observation. It should accommodate all the observations ok. It should accommodate all the observation means all the 800 values from 0 to 800 has to be accommodated. Then we need to fix the number of bars, that we that we used to draw the histogram. That accommodate number of bars is from 3 to 13 bars, this is ideal, this is ideal 3 bars to 13, 13 bars we can even move go beyond 30 bars if we like. But, 3 to 13 is an ideal or I can say the best selection of the histogram as a graphic tool. Now, the intervals here must be defined very clearly, internals are to be clearly defined ok.

So, we should be able to at least compare the histogram with the other graphs. So, it has to be comparable has to be comparable with other graphs. So, more we have the number of observations the more bars, we can draw in general it is 3 to 13 bars are the ideal ones, but also we can have more number of bars if we if we need to select only the small interval, if there is no other way to do that sometimes those cases also occur.

(Refer Slide Time: 11:14)



Data Visualisation

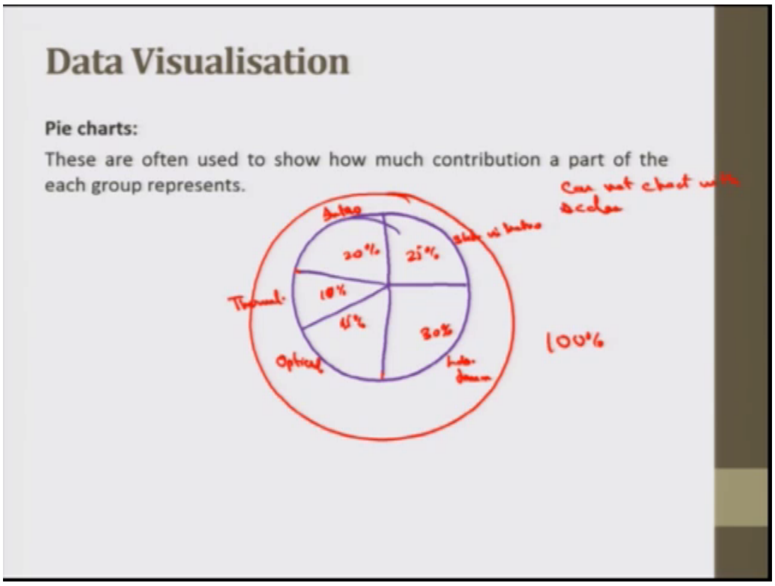
Histogram:

Now, interesting thing I like to discuss here is with histogram over with bar chart, we can cheat with scales. For instance if let me say I need to represent that the height is height from 0 to let me say 150 ok, you call it one fifty centimetres ok. Let me call the height of the human beings height of the workers who are working in my machine to lamp ok. It the height of the men let me this curve and height of the women, if I show this the maximum height of the men here is let me say at 150 a small number.

If, I took the Indian standard I put I might put 165 for women it has to be 100 and 48 ok. Let me say this is the height 165 148, this is one way. Another way is if I do not start with started with 0, I start it with just 140 to 170. I have 150 and 160 here then I can say that one height for the women would be close his to here 148 and height for the men would be here. So, it looks like that this value is high, this difference is high we can cheat with scales in case of bar diagrams ok.

By just changing the scale here ok. This is known as cheating with scales. So, it depends sometimes in marketing, if it is very deceptive sometime when they need to show the cost savings they will show ok, this much big saving is there ok. When they need to show that defects are maybe reduced they will show the big difference, where they need to show that there is a small difference, where they say cost raised they may show this, this small cost raise only ok. These things are deceptive sometimes.

(Refer Slide Time: 13:45)



Data Visualisation

Pie charts:

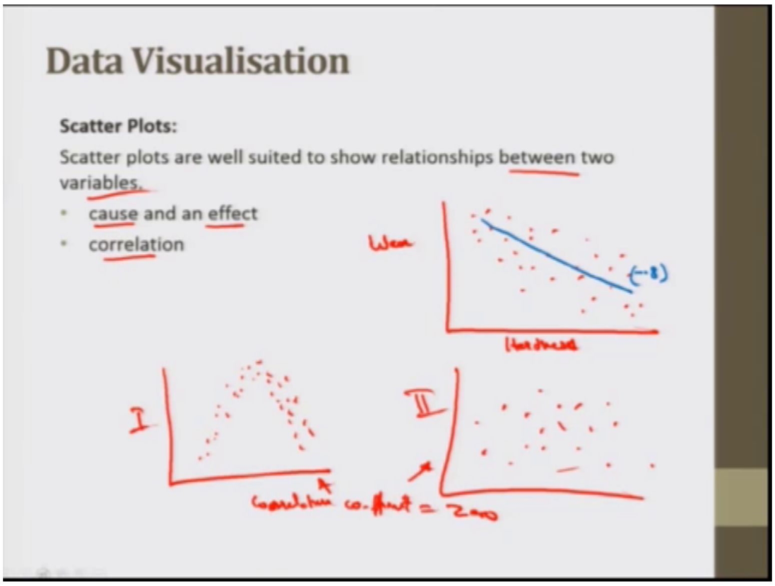
These are often used to show how much contribution a part of the each group represents.

Now, next is the pie charts; pie charts you people might be knowing it is like your pizza pie we can divide the data into various sectors here ok. And, we cannot cheat ok; we are cannot cheat with scales here, because this whole data this whole pie is 100 percent. So, there is often used to show how much contribution a part of each group represents. So, the pie charts typical example I can put here is a you divide this course a 30-30 hours course, we can divided into sections how much time is depth is spent on the specific limit say linear measurements, linear angular measurements.

How much time is depend is spent on laboratory demonstration. We can say let me say 33 percent could laboratory demonstrations. Let me say 30 percent on this ok. Let me say then 25 percent on statistics stats in metro, then let me say 30 25 is 55 20 percent on maybe introduction, introduction of the course. Then I can divide that is to a 30 25 is 55 plus 20 75 plus 15 plus 10 plus 10 plus 15 ok.

I can say this is on the thermal instrument this may be on optical instruments, but the thing is that the whole pie chart is containing the complete data set. It is 100 percent in total we can also put numbers, but pie we cannot cheat here it is just representing the whole figure. So, this is used to represent the contribution of each sector to the overall area ok.

(Refer Slide Time: 16:19)



Data Visualisation

Scatter Plots:

Scatter plots are well suited to show relationships between two variables.

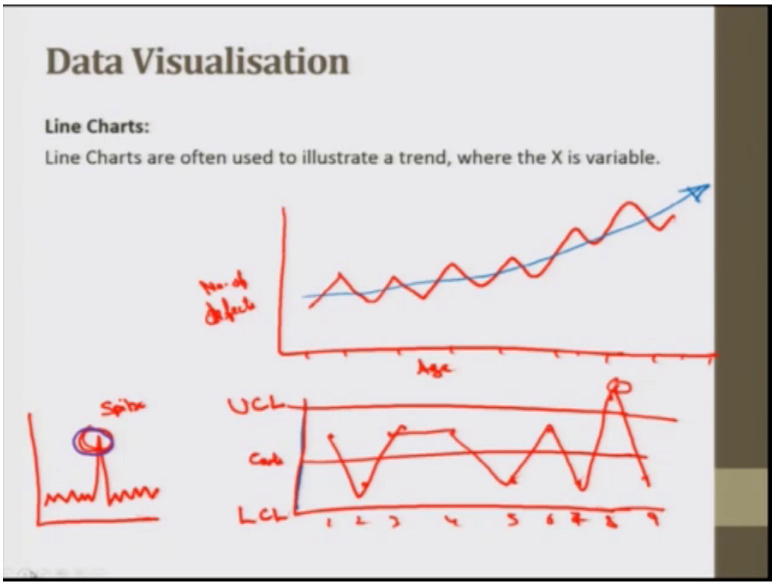
- cause and an effect
- correlation

So, next I have here is Scatter Plots. Scatter plots are very important scatter plots are well suited to show the relationship between 2 variables. This can because and an effect and they can be correlation the cause and effect in correlation. We can have scatter plot like this ok. Let me see, if I am having a in my metrology lab I am trying to measure the force or I am trying to measure the hardness. Hardness and I have to put it in the other way then, hardness and wear. When I am testing some specimen for wear, the hardness is increasing for a more hardness I will see the wear would be less. And for less hardness the wear would be higher we can see this kind of scatter plot here ok. When a data set is having more than 20 observations typically so, it can be like this number can vary, but they are we have a number of observations.

So, we can which is not recommended to draw one line like this, such a big all these points in one line. So, scatter plot is giving you the broad idea that what is the major trend. I can say this is a kind of a negative correlation, I can put the correlation coefficient here close to minus 0.8 ok, minus 0.8. So, this is the use of scatter plot correlation sometimes cause and effect sometimes the correlation is 0 and we just have this scatter plot like this, this is correlation 0.

But, it is interesting that sometimes the non-linear curves are there and if the scatter plot is like this. Here, also the correlation coefficient correlation coefficient is equal to 0, in both the cases, in this case and in this case. Here which is actually there is no correlation, but here in case 1. This is case 2 in case one we have the correlation it is first increasing then decreasing, but it is non-linear it is having some curve here ok. So, correlation coefficient does not work, but yes scatter plot can be used to find the rough idea what kind of data do we have ok.

(Refer Slide Time: 19:07)



Data Visualisation

Line Charts:

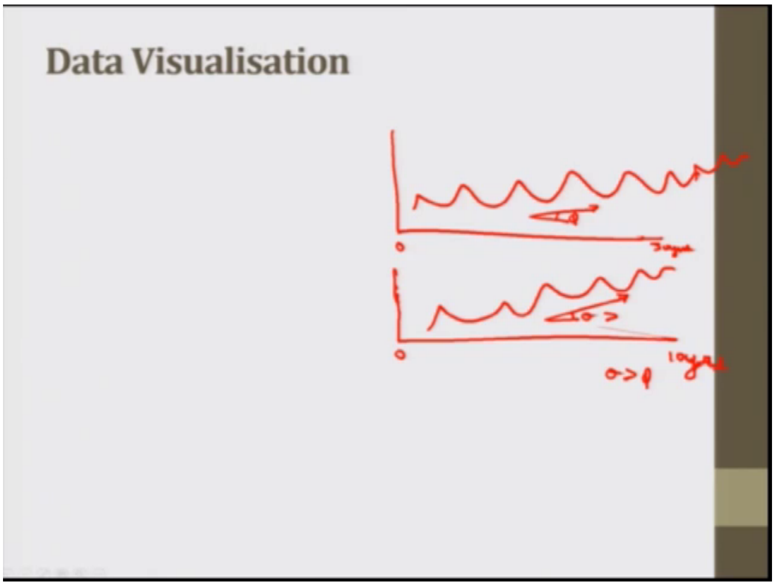
Line Charts are often used to illustrate a trend, where the X is variable.

Next is line charts. Line charts are often used to illustrated trend, where X is variable. Now, line charts are just drawn simply like this one instance I am having age of the instrument and number of defects ok. It can be like this, first point, second point, third point, fourth point, fifth point, sixth point it can be like this ok. We can see with age the number of defects in any instrument if instrument is used for let me say 40 years. The I can say the number of defects or the number of effective observations those are being taken from then, this is a increasing trend it is showing an increasing trend here ok.

So, line diagram is also like we when we plot the control charts or X bar chart we draw the line diagram to see whether do we have outliers or not. So, we will see that in control chart we have this central line here with us have this upper control limit here we have the central line ok. Then we at each observation that we are having in a day will just put the value here ok. We pet value of 1 2 3 4 5 6 7 8 9 and so on. Then, if we draw this line diagram, we can see that and if we draw this line diagram, we can see that it is having this kind of trend this is one outlier that has come 1 2 3 4 5 6 7 and 8; that has come at eighth location ok.

Also, sometimes in line diagram we can see that if this data is like this, sometimes there is a peak like this, what is this why this peak has occurred we can work on this as well ok. This is a peak where you can say a betrayal call it spike, this is spike why this spike has come ok.

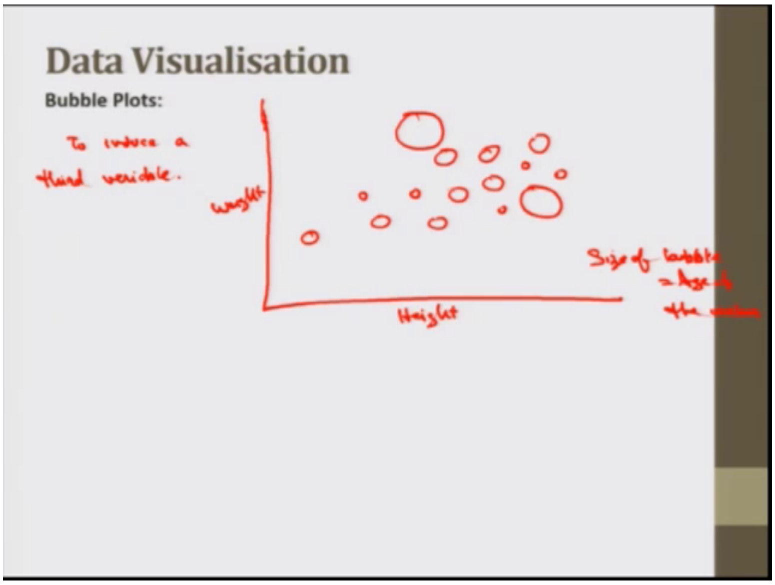
(Refer Slide Time: 21:55)


Data Visualisation

In line diagram also we can cheat with the scales. If, we need to show less variation or more variation, we can just cheat with the scales instance if I need to like show the age. That, that trend I need to show that the trend is small or the slope of the trend is lesser, if need to show that the trend is lesser I can widen this thing ok. I can widen the horizontal scale here.

I can show age from 0 to 50 years ok. Here I can show the age from 0 to 10 years, it from 0 to 15 years it is like 0 to 50 year, it is like this there is a slope that is increasing like this. For 0 to 10 the slope would be seen a little higher ok. This angle would be greater this theta value is greater than this I call it phi value ok. The theta is greater than phi and I can show that the slope is high. So, this is always there cheating of scales can happen.

(Refer Slide Time: 22:58)



Data Visualisation

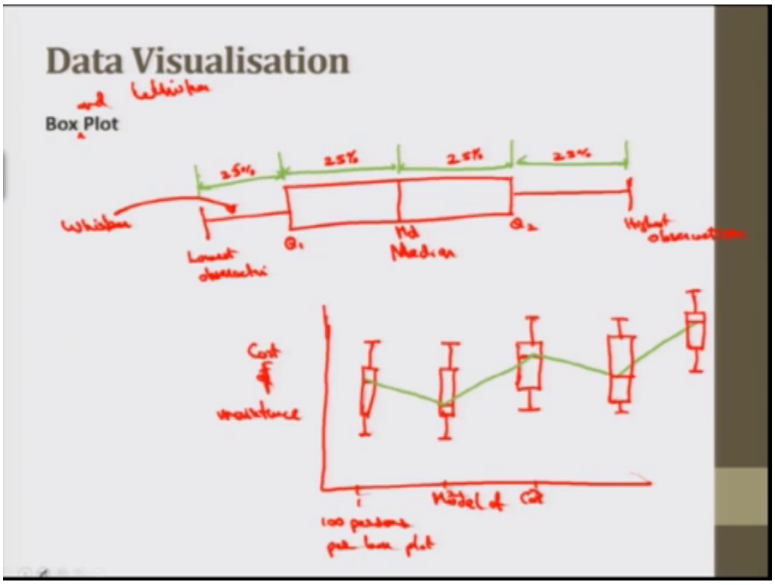
Bubble Plots:

To induce a third variable.

Next is Bubble Plot. Bubble plot it is a kind of a scatter plot, but we can also introduce another dimension into it, another dimension why that another dimension means another variable in. For instance I am talking about the weight and height of maybe workers. And, also I am putting I am in place of the dots, those were in scatter plots I am putting bubbles ok. Where these bubbles represent, where is bubble the size of the bubble represent the maybe age of the person ok. I can put size of bubble is equal to age of the worker. So, bubble plots are used to induce a third variable.

However, we can also have 3 D plots like surface diagrams ok. And, the certain other ways to represent the third variable bubble plot is one that in 2 dimensions we can just see the value of or the field of the data here.

(Refer Slide Time: 24:28)



Data Visualisation

Box Plot

Next, plot is Box Plot. From statistical viewpoint box plot is very important, but is box plot we divide the data into quartiles. This is my box plot ok, I divide my data into quartiles this is my lowest observation, and this is my highest observation. Hence, as I discussed this is my median and this is Q 1 and Q 2. That means, I am having 25 percent of data here 25 percent here 25 percent here and 25 percent here.

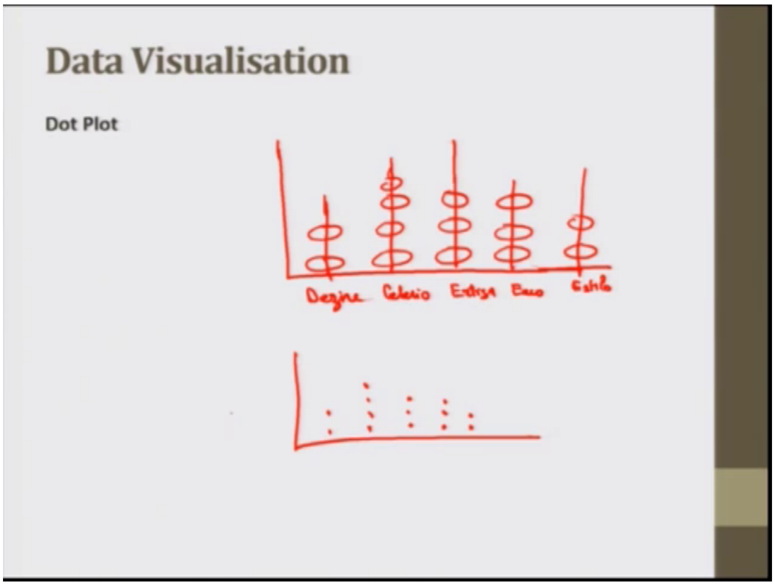
This is my box plot just I am putting the dimensions here, this is 25 percent this is 25 percent 25 percent of data means 25 percent determines of values line between, if there are 30 values; 30 value means if I have 40 values 40 values means 10 values in each of these. Now, this line is known as whisker this is known as whisker, like a animals have whisker which is a long hair like in the moustache. So, this is that kind of this whisker it is a long extended line ok.

So, this is also known as box and whisker plot. So, what is the significance of box plot actually this is median here, we can see the location of the data, where is our data trying to being located. For instance I can have boxes here in my data ok. Let me say this is like this ok. So, in example can be this is cost of maintenance from different customers and this can be the model of car, for a specific company with same for Maruti receptor models it is it can be Maruti desire it can be Maruti, Eeco, Baleno, Celerio, Ertiga like those.

So, what is the cost for the customer (Refer Time: 27:18) 400 customer for each data there are 400 customers for 1 2 3 1 2 3; 100 percents per box plot. So, we can see that in this case the median the central data is quite towards the lower side. It is into a quiet for the first case it is quite to on the upper side, it is also in the upper side in the third case. So, this kind of box plot can give you rough indication even I can draw a can join the median points here straight lines. So, this is if I having number of plots like this here.

This is an extension of my line charts ok. The line chart is the green colour and box and whisker plots are giving the more detail of each point here each data here data point here. So, this is box and whisker plot and we will use this when we will do further other discussions.

(Refer Slide Time: 28:33)



Data Visualisation

Dot Plot

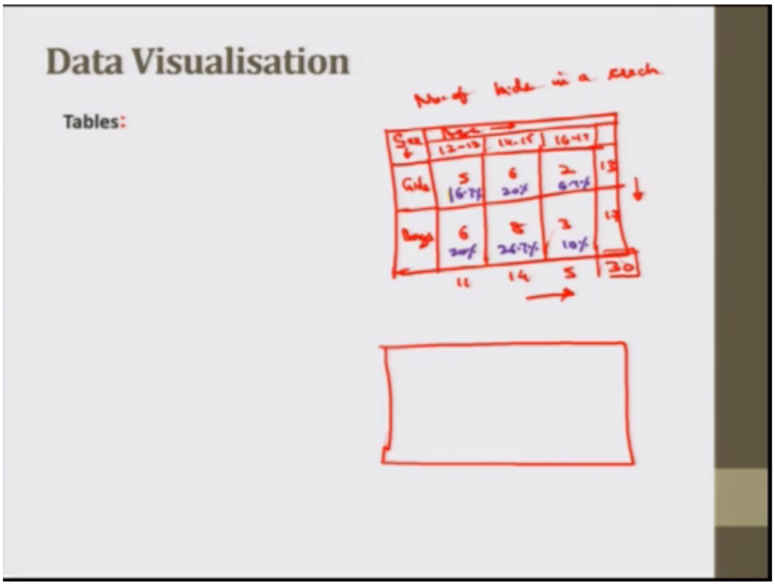
And next is Dot Plot. Dot plot is recording the data as and when it is happening, like a I will took the example of the cars only a cars are coming to the service stations to get serviced and each car they get gets service just need they are just need notice that, how many cars are there.

How many different models are there being these are servicing each day. So, they can put here different models ok. We can put that these into different beans ok. I can even have some pins here and keep putting rings over here ok. Let me say this is these are motors of Maruti car, Maruti desire, this is Celerio, this is you call it a Maruti Ertiga, then Maruti

Eeco you can put any model like those this is Estilo. So, each car they are servicing they are putting a ring here.

They putting a ring here for in a day then get getting each car they are putting a ring over here again ok. They getting one Celerio putting ring they getting one Ertiga then putting a ring just according the data. This kind of plot can also be put like this 1 2 1 2 I am putting rings here 1 2 3 4 1 2 3 1 2 3 1 2 ok. So, this is dot plot.

(Refer Slide Time: 30:17)



Data Visualisation

Tables:

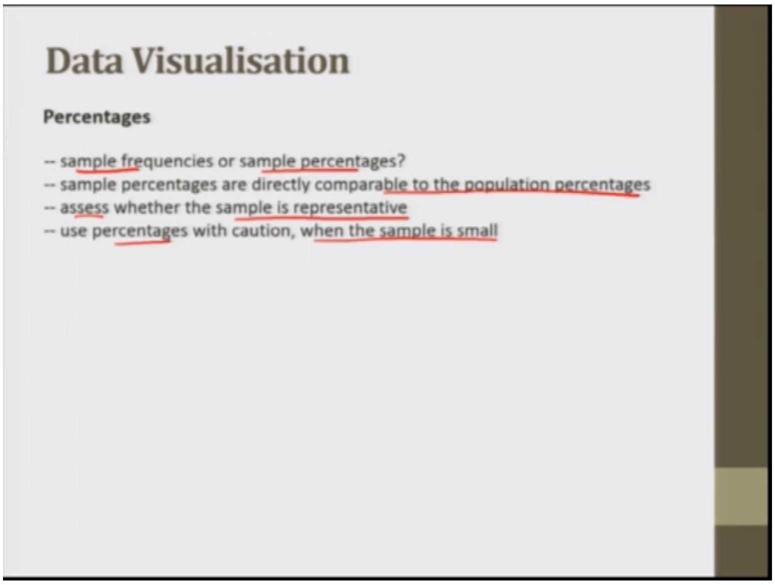
No. of kids in a coach

Sex	Age 12-13	14-15	16-17	
Girls	5 16.7%	6 20%	2 6.7%	13
Boys	6 20%	8 26.7%	3 10%	17
	11	14	5	30

Now, next in data visualisation next comes is tables. Tables as we know those are very common we have been using those table is a kind of a matrix in which you have rows and columns, it is important to have the tiles for the columns, and the tiles for the rows as well and overall title of the table.

So, in tables also we can represent the percentage is as well; for instance this is some table we having a title for row we having titles for the columns here, and the title for rows here ok. We having some data we can put that data into percentages as well. So, all those examples can also be taken into account.

(Refer Slide Time: 30:58)



Data Visualisation

Percentages

-- sample frequencies or sample percentages?
-- sample percentages are directly comparable to the population percentages
-- assess whether the sample is representative
-- use percentages with caution, when the sample is small

So, percentage is as these are samples we consider sample percentage is, sample percentages are directly comparable to the population purpose percentages. And assess whether the sample is representative or not. Use percentages with caution when sample is small.

Now, percentages can also be used to represent data to represent the data in the form of a table. In the for instance if I put the numbers here let me see let me I put the age group here if I put age varying from maybe 12 to 13, then 14 to 15, then 16 to 17. Let me say this is 5 6 2. And, these are 2 things here 2 points in the category girls and boys ok, 5 6 2 girls 5 girls are from age 12 to 13 6 from 14 to 15 and 16 from 2 form 16 to 17.

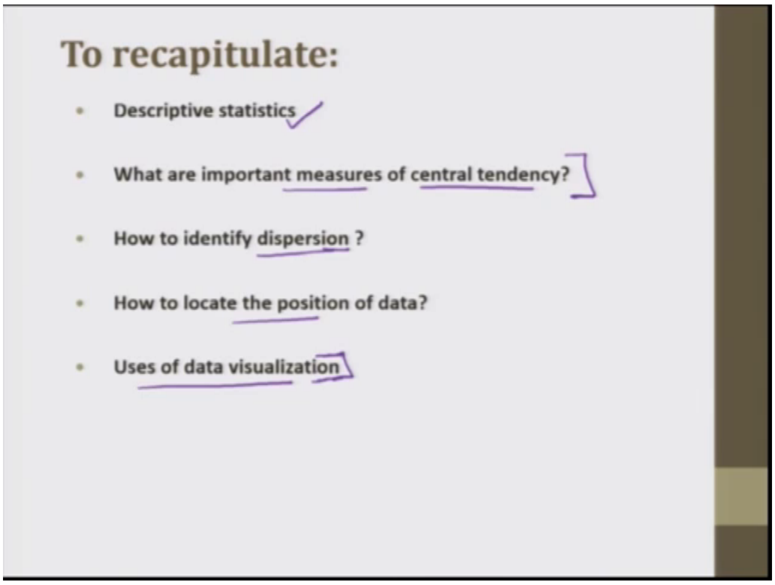
Similarly, I can have boys 6 8 3 total here is 5 6 2 is 13 17 overall total here is 30 ok. And, here also I can have the total 5 and 6 is 11 6 and 8 is 14 2 and 3 is 5. So, overall total is from this side or this side. This table has to have the heading this is number of kids you can say call it crach ok. Then we have the age here this is age and this is age this side we having age then we are having their gender here. So, I call it their sex here. This is the table.

This data can also be represented in the form of percentages the percentages, if the data is like this I can in place of 5, I can put a representation in the same table is actually 5 by 35 by 30 is 16.7 percent 6 by 30 is 20 percent, 2 by 30 is about 6.7 percent, this is again

20 percent, this again this is 8 by 30 is 26.7 percent. And, this is tributary is 3 by 30 is 10 percent. Now, this kind of tabular form formulation is used for goodness of fit ok.

So, this is data visualisation. Next, I will take you to the demonstration part of this lecture we will discuss how to plot these graphics that we have just discussed.

(Refer Slide Time: 34:06)



To recapitulate:

- Descriptive statistics
- What are important measures of central tendency?
- How to identify dispersion ?
- How to locate the position of data?
- Uses of data visualization

So, just recapitulate we discussed this descriptive statistics in the last 2 lectures and we discussed the important measures of central tendency ok. That was the send the mean, median, mode, then measure of dispersion ok; then, we discussed locate the position we discussed quartiles and percentiles, measure of dispersion were then range and standard deviation variance. And also we discussed the uses of the data visualization and different data visualisation tools or graphs that we can use.

Now, we will meet in the next part of the lecture, we will plot the graphs and then we will discuss about the; we will take a hour to consider education forward to the distribution of the data.

Thank you.