Statistics for Experimentalists
Prof. Kannan. A
Department of Chemical Engineering
Indian Institute of Technology – Madras

Lecture – 07A
Exploratory Data Analysis - Part A

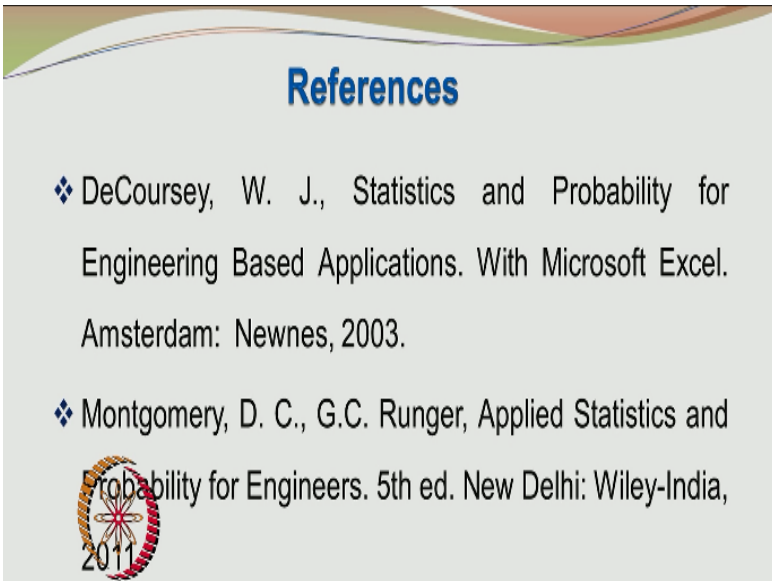
Welcome back. In today's lecture, we will be looking at exploratory data analysis. So far we have studied about random variables both discrete and continuous. We looked at some of the parameters that are encountered when dealing with both continuous and discrete random variables the mean variance standard deviation and so on. We also looked at the moments then we came on to normal probability distributions.

And one of its variance namely the lognormal probability distribution. Before we go further into details of statistics it is worthwhile to take a small break and look at presentation of data. This is also very important to us. Let us say that you have conducted the experiments and you have the data available with you. It will be a good idea to subject the data to preliminary data analysis to get a feel for the data trends.

Between what range of values, you find the data points whether there are some unusual observations and whether the data are looking linear or they are showing a strong curvature and the response is plotted against the main variable or the variables. You also may want to see whether the distribution of the data is following a certain standard distribution. You may want to check whether the distribution of the data is normal.

You may also want to present the data and what are the different effective ways of presentation. What will you actually look for in a data? So these are some of the things we are going to discuss. Now of course this discussion is not exhaustive or complete it is just a starting point. There are many more data analysis procedures which you may definitely understand when you read up after getting exposed to this lecture.

(Refer Slide Time: 03:20)

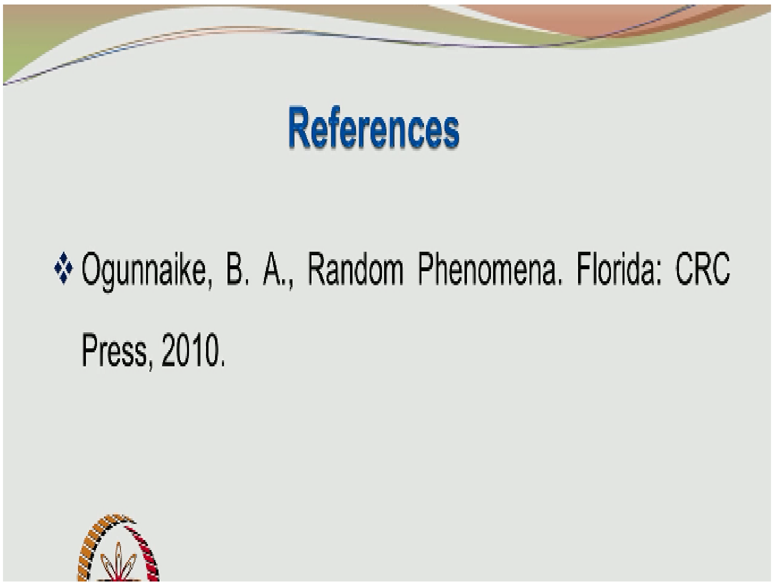


References

- DeCoursey, W. J., Statistics and Probability for Engineering Based Applications. With Microsoft Excel. Amsterdam: Newnes, 2003.
- Montgomery, D. C., G.C. Runger, Applied Statistics and Probability for Engineers. 5th ed. New Delhi: Wiley-India, 2011

Coming to the references there is a nice book by DeCoursey a Statistics and Probability for Engineering Based Applications with Microsoft Excel. We are also going to follow our usual reference books the one written by Montgomery and Runger.

(Refer Slide Time: 03:43)

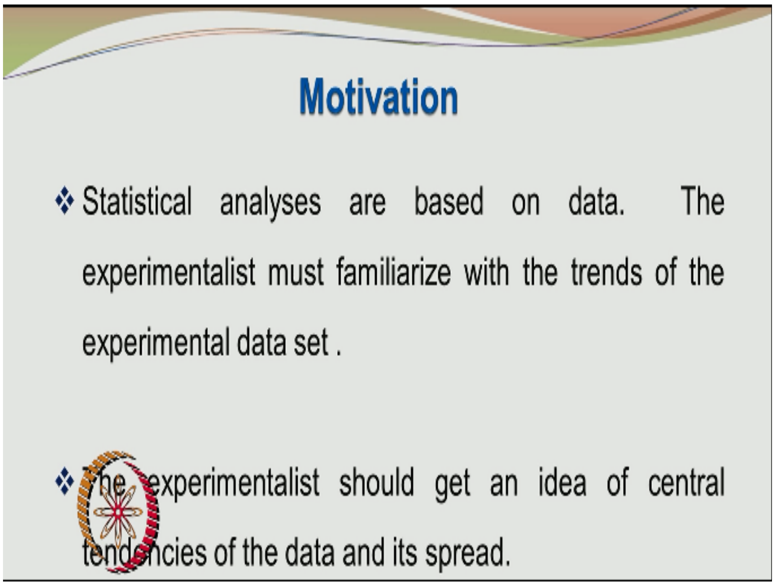


References

- Ogunnaike, B. A., Random Phenomena. Florida: CRC Press, 2010.

Also we will be following Ogunnaike the name of the book is Random Phenomena.

(Refer Slide Time: 03:55)



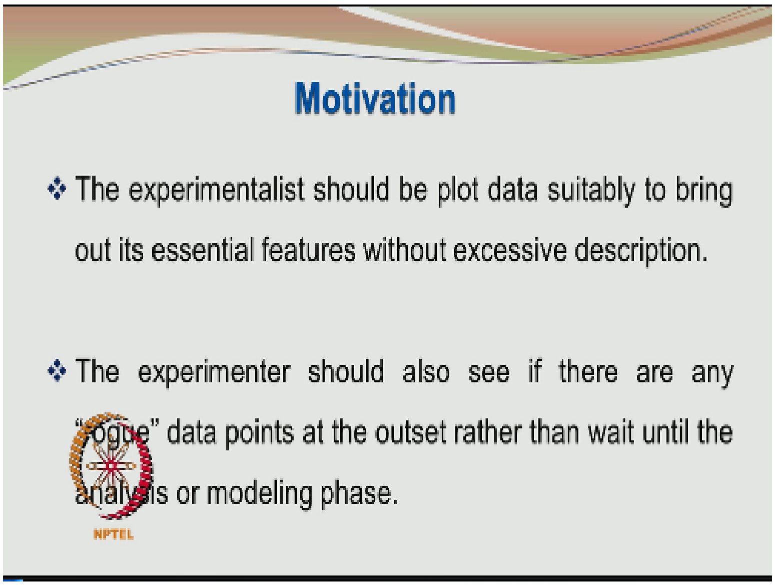
Motivation

- ❖ Statistical analyses are based on data. The experimentalist must familiarize with the trends of the experimental data set .
- ❖ The experimentalist should get an idea of central tendencies of the data and its spread.

Coming to the motivation of the exploratory data analysis as the name implies you are having the data and you are going to do an exploratory analysis of the available data. Statistical analysis are based on data. What is the difference between statistics and mathematics both seem to use for example integration differentiation and many more methods of analysis. Statistics is based on data while mathematics is based on pure numbers.

This kind of interpretation is of course arguable, but this is one way of looking at it. So when you have done the experiment you must familiarize with the trends of the experimental data set. As an experimentalist you should also look for the center point of the data distribution you may want to look at the average value and also quantify the spread of the distribution.

(Refer Slide Time: 05:14)



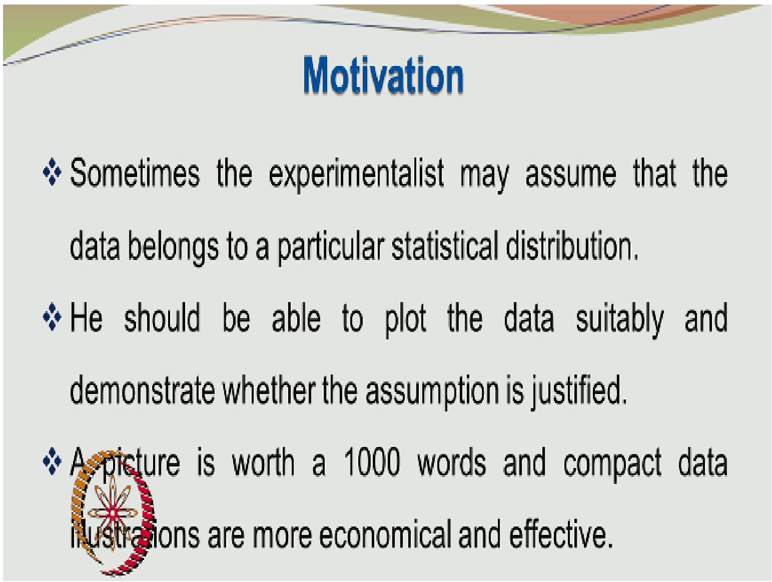
Motivation

- ❖ The experimentalist should be plot data suitably to bring out its essential features without excessive description.
- ❖ The experimenter should also see if there are any "rogue" data points at the outset rather than wait until the analysis or modeling phase.

In order to get proper idea on the spread you may want to first plot the data and after plotting you should be able to discern its essential features without too much of textual description. You should also see whether there are any rogue data points in the data. Data set you have collected and it is better to become aware of these right at the outset because if you can detect the location of these outliers.

You perhaps may repeat the experiments corresponding to the occurrence of these data points and see whether you are still getting the same values. If you otherwise wait until the end of the experimentation, then these rogue data points may stick out like (()) (06:22). Then you would be at a loss as to what to do with them then you will have to attribute some reasons while they may have occurred. So the moral of the story is if you have any outliers you better locate them right at the outset and take suitable action.

(Refer Slide Time: 06:54)



Motivation

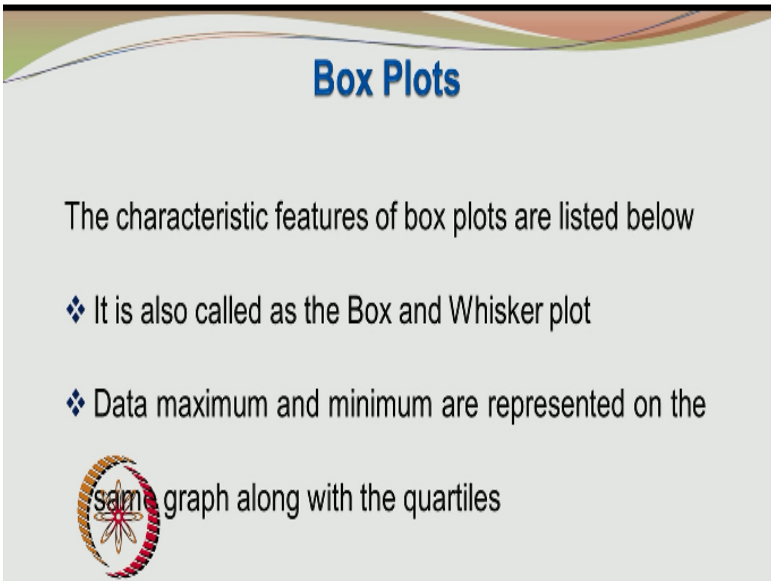
- ❖ Sometimes the experimentalist may assume that the data belongs to a particular statistical distribution.
- ❖ He should be able to plot the data suitably and demonstrate whether the assumption is justified.
- ❖ A picture is worth a 1000 words and compact data illustrations are more economical and effective.

You may want to speculate that the given data belongs to a particular statistical distribution. This may be based on what other people have experienced with this type of data. For example, particle size distributions are commonly expressed in terms of the log normal distribution. So you may want to assume that the particle diameters in your data set can be expressed in terms of (()) (07:31) is the particle diameter.

And then you may want to show them in the form of log normal distribution, but you have to confirm that indeed the data belongs to a particular distribution. So you should be able to plot the data suitably and demonstrate whether this assumption is justified. A picture is worth a 1000 words and hence it is always better to present your data in a compact, economical and effective manner.

In the corporate sector most of the presentations involve data analysis and they have to be presented in a compact manner. A lot of information should be present in 1 or 2 diagrams or presentations. You do not want to show 10 or 20 graphs to drive home your point.

(Refer Slide Time: 08:44)



Box Plots

The characteristic features of box plots are listed below

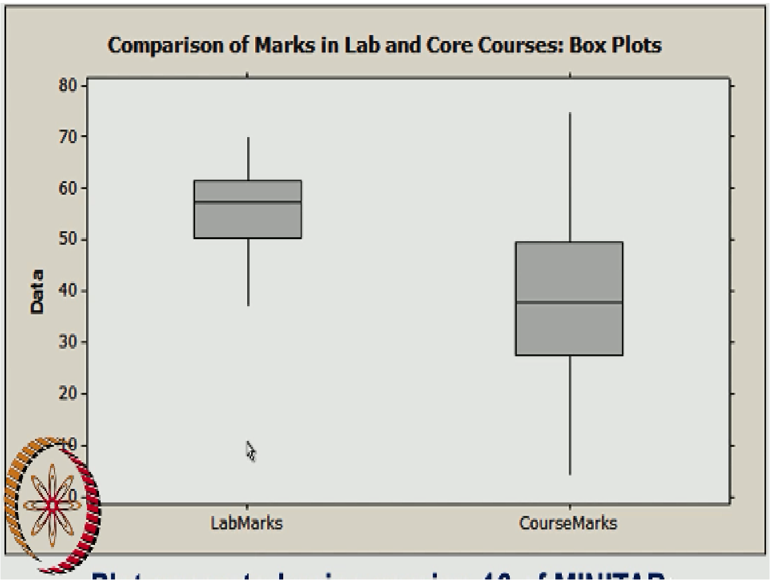
❖ It is also called as the Box and Whisker plot

❖ Data maximum and minimum are represented on the same graph along with the quartiles

Let us look at the box plots. We came across the box plots in the introduction session in the very first lecture if you recall. The box plot is also called as the Box and Whisker plot. In this plot, you are able to show a lot of information. You can show the data maximum and minimum along with the other characteristics like the quartiles, median, the outliers etcetera. The box plots are economical and a lot of information can be presented in a compact manner.

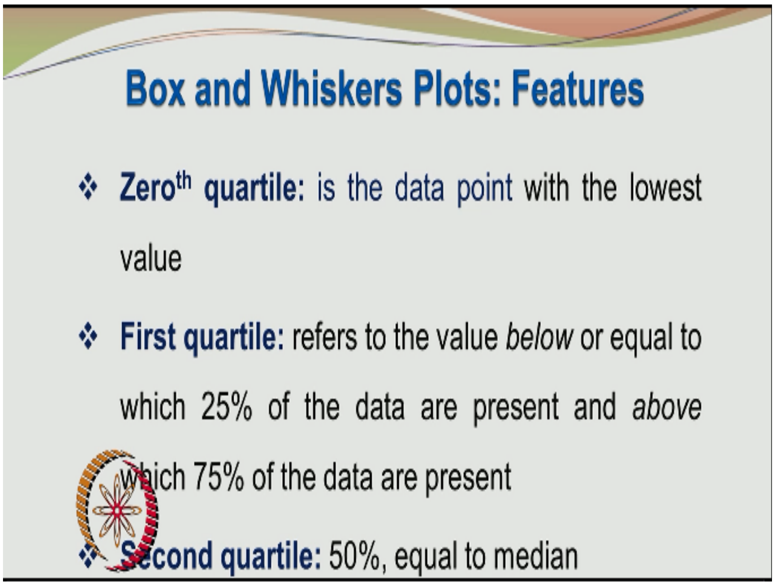
If you have different sets of data and you want to compare them box plots are quite useful. Let us look at the features of the box plot.

(Refer Slide Time: 09:50)



I will show you a diagram of the box plot. Here we are comparing 2 sets of data. This is the so called Box and these are referred to as Whiskers. So this is referred to as the first quartile. The second line is the second quartile, the third line is the third quartile. Quartile you may relate it to quarter or one-fourth. So in this diagram we have shown the Whiskers we have shown the box, we have shown the first, second and third quartiles.

(Refer Slide Time: 10:41)



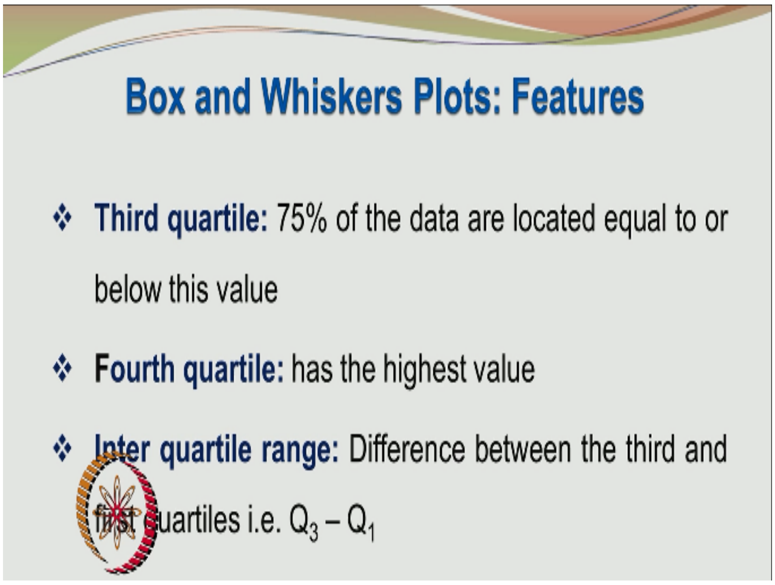
Box and Whiskers Plots: Features

- **Zeroth quartile:** is the data point with the lowest value
- **First quartile:** refers to the value *below* or equal to which 25% of the data are present and *above* which 75% of the data are present
- **Second quartile:** 50%, equal to median

Now let us go back and see the definitions for this quartiles. The 0th quartile is the data point with the lowest value. For example, in a class where the marks are distributed the teacher may want to arrange the marks from the lowest mark to the highest mark. Usually this is not done and the marks are distributed in the random manner, but some teachers want to present the papers in the ascending order of marks.

So coming to the 0th quartile it is a data point with the lowest value. The first quartile it refers to the value below or equal to which 25% of the data are present and above which 75% of the data are present. The second quartile refers to the data points below or equal to which 50% of the data are present and above which obviously the remaining 50% are present. The second quartile is also equal to the median.

(Refer Slide Time: 12:07)

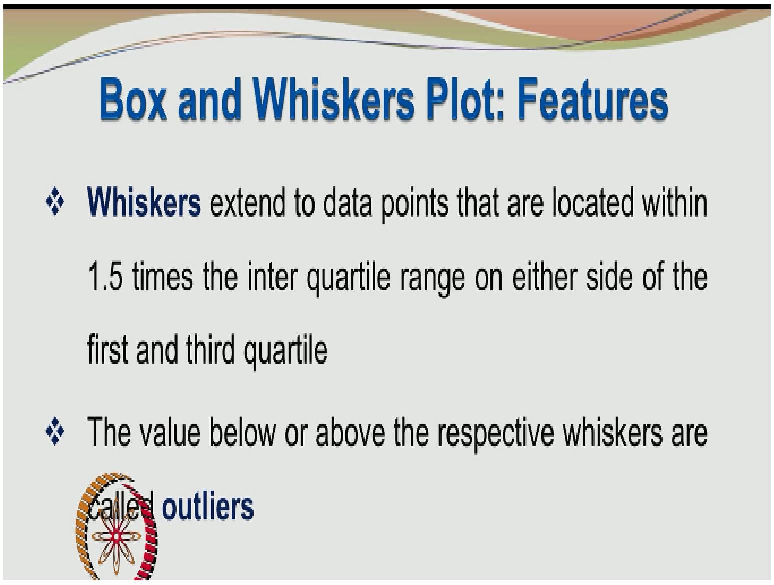


Box and Whiskers Plots: Features

- **Third quartile:** 75% of the data are located equal to or below this value
- **Fourth quartile:** has the highest value
- **Inter quartile range:** Difference between the third and first quartiles i.e. $Q_3 – Q_1$

The third quartile by now you will be familiar with it. The 75% of the data are located equal to below this value and 25% of the data are located above this value highest number. The inter quartile range is the difference between the third and the first quartiles that is Q3-Q. Q3 is the third quartile and Q1 is the first quartile. What about the Whiskers if you recollect I told that these vertical lines shooting out of the boxes on either side are termed as Whiskers.

(Refer Slide Time: 12:51)



Box and Whiskers Plot: Features

- **Whiskers** extend to data points that are located within 1.5 times the inter quartile range on either side of the first and third quartile
- The value below or above the respective whiskers are called **outliers**

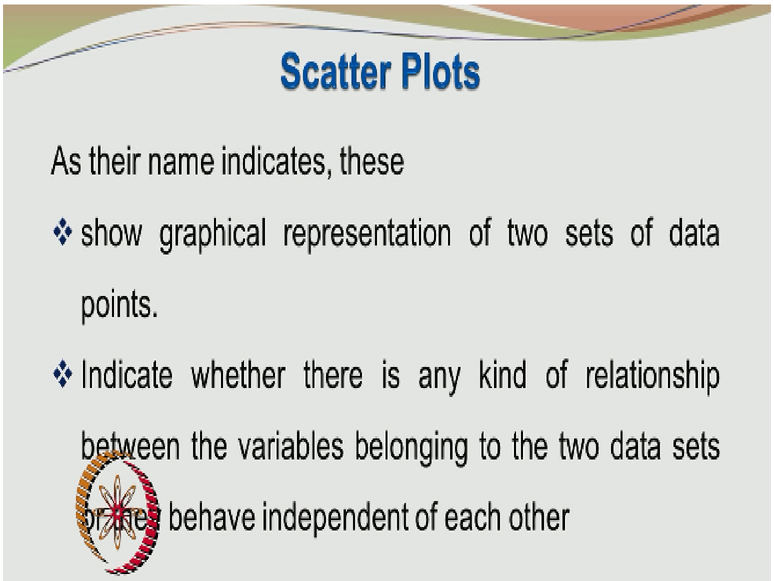
The Whiskers are drawn from the edge of the box to the data point that is located within 1.5 times the inter quartile range. So you have the first quartile and the third quartile. Then you want to indentify data points that are located at 1.5 times the inter quartile range from these 2 quartile. The Whiskers need not be of equal length. So this value is the one which is lying within 1.5 times the inter quartile range.

These 2 distances are not the same, but however this data point at the very end of the whisker is falling within 1.5 inter quartile range. Similarly, here also you have whiskers, but it appears that the data point at the edge of this whisker is lying at the same distance from the third quartile. As this point was lying from the first quartile so it depends upon the data set. We already discussed about the performance by the students in the lab and in the course.

Lab is more of a group activity and so the marks are sort of closer to each other when compared to the core course performance. If there are any data points below or above the respective Whiskers, then they are referred to as outliers. What we are trying to do here is whatever data point is falling within the quartiles or close to the quartiles are considered to be the kind of expected points.

And if you are having any points which are lying beyond the 1.5 times the inter quartile range from the first quarter or the third quarter is considered to be a rogue point or a outlier. This box plot can be generated in different ways I have used the version 16 of MINITAB to generate this box plot diagram.

(Refer Slide Time: 15:36)



Scatter Plots

As their name indicates, these

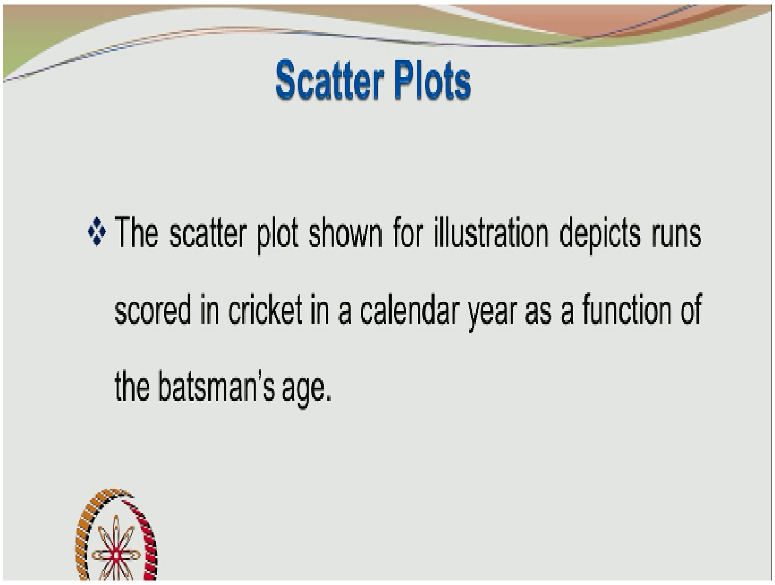
❖ show graphical representation of two sets of data points.

❖ Indicate whether there is any kind of relationship between the variables belonging to the two data sets or they behave independent of each other

Now let us look at another kind of plot namely the Scatter plot. It shows the scatter on the data to put it simply. It shows the 2 data set on a regular graph sheet. It compares 2 data sets you may want to plot the first data set along the x axis on the second data set along the y axis and then see if there is any correspondence between the 2. We want to see whether there is dependency between the 2 data sets.

For making the comparison of course the length of the 2 data set should be the same. The first data set has 20 points the second data set should also have 20 points.

(Refer Slide Time: 16:33)

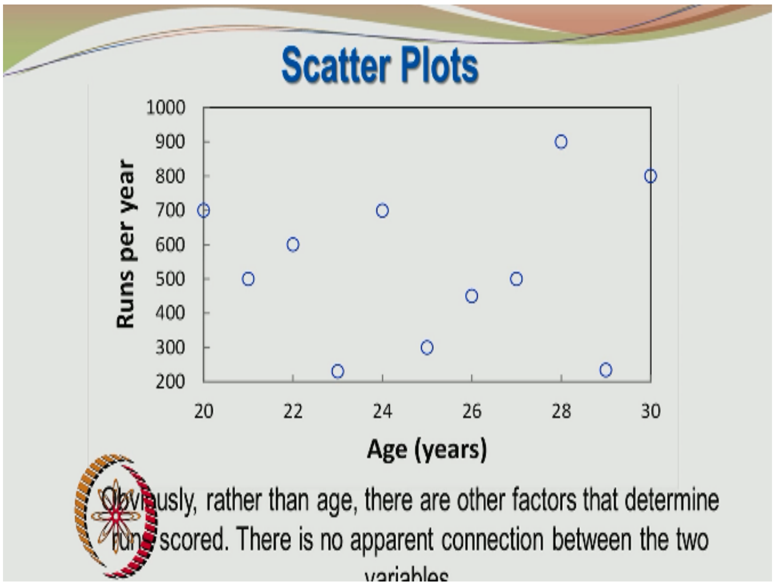


Scatter Plots

- The scatter plot shown for illustration depicts runs scored in cricket in a calendar year as a function of the batsman's age.

I will demonstrate the scatter plot with the help of an example. Let us say that we are looking at a batsman performance over the years and we want to show runs scored in a calendar year as a function of the batsman's age.

(Refer Slide Time: 16:59)



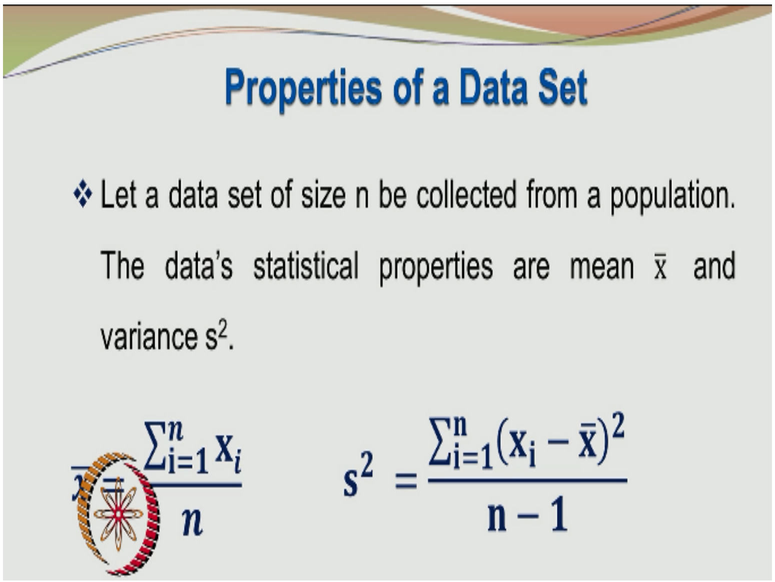
Scatter Plots

Obviously, rather than age, there are other factors that determine runs scored. There is no apparent connection between the two variables

Whether the batsman is getting better with the age or he is getting worse with the age or he goes through an optimum phase before beginning to fade out. Well when we look at this particular scatter plot for this particular batsman if you look at that graph the runs scored per calendar year is shown on the y axis and his age is shown on the x axis. From this diagram it can be seen that there is no apparent relation or dependency between the run scored per year and age.

The runs scored per year may have fluctuated based on other reasons they might not have been due to the aging of the batsman. So this clearly shows that there is no dependency on the runs scored with the age of the batsman in the range of 20 years to 30 years. Normally when you do an experiment and collect the data it is referred to as the sample.

(Refer Slide Time: 18:30)



Properties of a Data Set

- Let a data set of size n be collected from a population. The data's statistical properties are mean $\bar{x}$ and variance s^2.

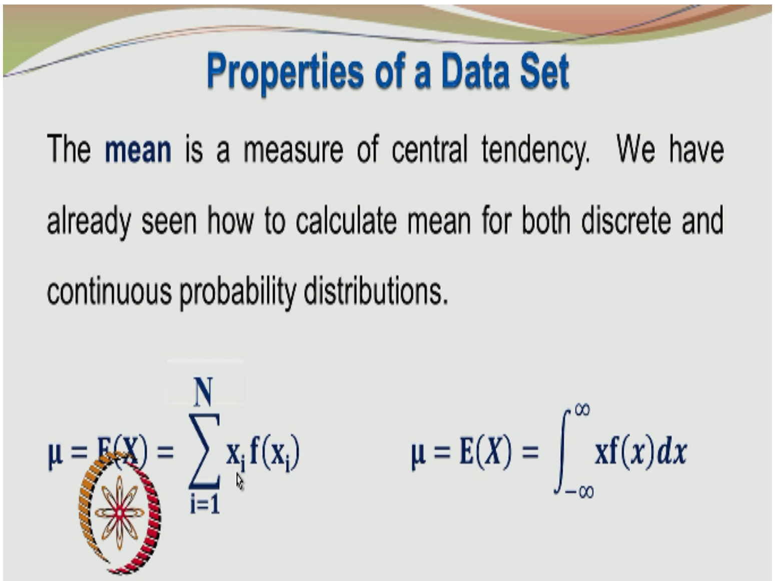
$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$$

So we would like to look at the sample properties and the most common one would be the sample mean. The sample mean is denoted as x bar and the variance is denoted by s square. We have earlier seen the mean being represented by mu and variance being represented by sigma square. For example, in the normal distribution the mean was given as mu and the standard deviation was given as sigma square, but remember here we are talking about the sample. Earlier we were talking about the population.

The population parameters were given in terms of mu and sigma square for the mean and variance respectively. Here we are talking about the sample and we denote it by x bar and s square. The sample mean is the arithmetic mean and we just sum all the data point values and divided by the number of data points. The S squared is the sample variance which is defined in terms of the deviation from the mean.

The deviation of each and every sample data point from the mean and this squared and then added and after that we divided it by n-1.

(Refer Slide Time: 20:10)



Properties of a Data Set

The **mean** is a measure of central tendency. We have already seen how to calculate mean for both discrete and continuous probability distributions.

$$\mu = E(X) = \sum_{i=1}^{N} x_i f(x_i) \qquad \mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

So what we saw was for a discrete data set the mean is a measure of central tendency. For a discrete distribution the mean is also referred to as the expected value of x and that is given by sigma i=1 to N Xi f of Xi and in the case of continuous distributions we have mu= expected value of X that is=-infinity to + infinity x f of x dx. So looking at the properties of the data set we are defining for a discrete distribution we say that that the mean mu is expected value of x and that is given by sigma xi f of xi.

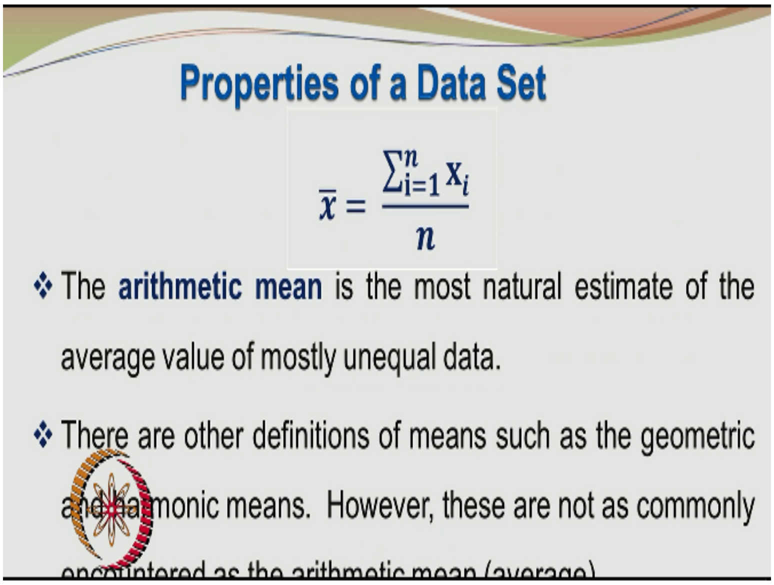
We are also defining the arithmetic mean as sigma xi/n. So are we having 2 different definitions what is then the basis for the arithmetic mean. It is actually quite simple. If each of the xi values have these identical probabilities of occurrence, then f of xi will be simply 1/n. So you put 1/n here that is independent of the index i so we have mu equals sigma xi f of xi that will become sigma xi/n.

So that will be the capital N where N is the total number of entities in the population. Now we do the same thing for the sample mean when you have the sample we have used the same formula, but a sample is a subset of the population the number of entities in the sample will be much smaller than the population. We may not be finding it practical to take the data from each and every entity in the population.

So we take a representative sample from the population and get the important characteristics. So when we do the mean that is denoted by not mu, but it is denoted by x bar and that is given by sigma i= 1 to n xi/n. Here n is small n it should not be confused with capital N. Capital N is meant for the overall number of entities in a population it may be even going into lakhs or million.

So it can be a huge population, but a sample is usually in the order of let say 30 it can even be as low as 5 it can go up to 30 or 40. The sample need not be larger than that. So the sample mean is defined as sigma i=1 to n xi/n. We did have f of xi, but since the probability of occurrence of each of the item in the sample was identical it became 1/n and so we have x bar= sigma i= 1 to n xi/n.

(Refer Slide Time: 24:08)



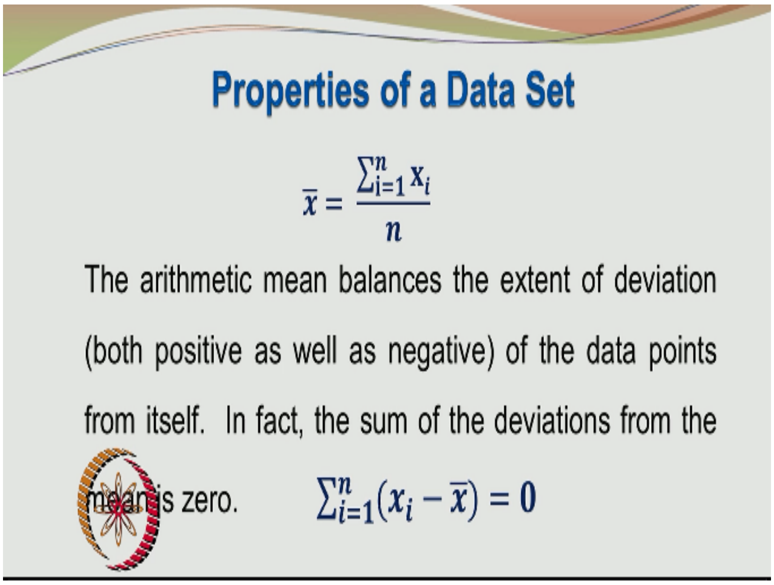
Properties of a Data Set

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- ❖ The **arithmetic mean** is the most natural estimate of the average value of mostly unequal data.
- ❖ There are other definitions of means such as the geometric and harmonic means. However, these are not as commonly encountered as the arithmetic mean (average).

There is the most natural way we take the average of a finite data set. You have other definition such as the geometric mean and the harmonic mean. However, these are not commonly used as that of the arithmetic mean.

(Refer Slide Time: 24:40)



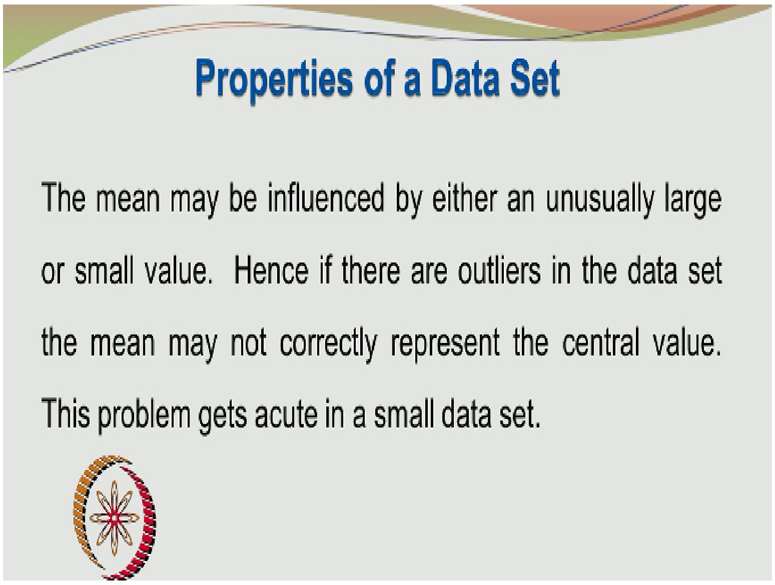
Properties of a Data Set

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

The arithmetic mean balances the extent of deviation (both positive as well as negative) of the data points from itself. In fact, the sum of the deviations from the mean is zero. $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$

The arithmetic mean balances the extent of deviation both positive as well as negative of the data points from itself. In fact, you identify the mean in such a way that the positive deviations from that will balance out the negative deviations such that the total sum of these deviations will be= 0.

(Refer Slide Time: 25:15)



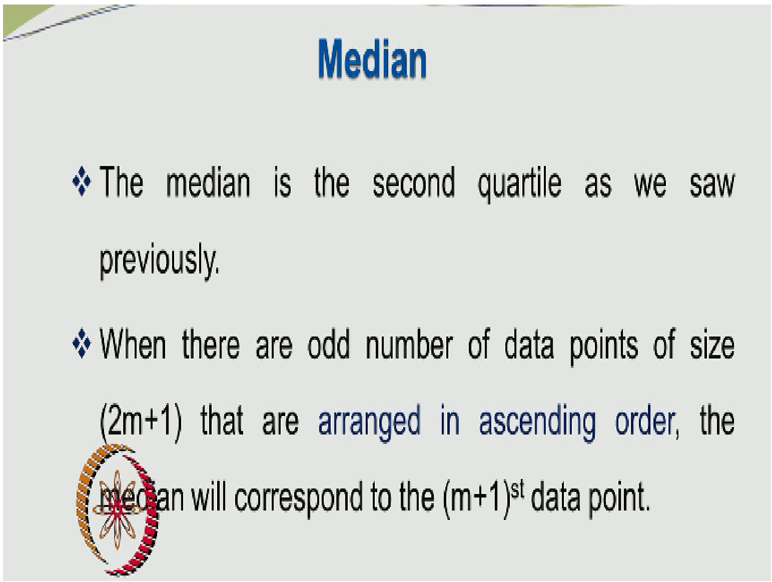
Properties of a Data Set

The mean may be influenced by either an unusually large or small value. Hence if there are outliers in the data set the mean may not correctly represent the central value. This problem gets acute in a small data set.

The problem with this kind of definition is the presence of unusually large value or an unusually a small value may influence the average value. The average is a measure of the overall data set and let us say that the batsman is playing in a 3 test series and if he has score than 4 innings 200 runs okay the average may look to be a healthy 50. On the other hand, if he has scored 200 and then he scored ducks in the remaining 3 innings then the average of 50 is not a good representation of his performance.

He has performed very well in the first innings and then done nothing in the remaining 3 innings. So when you are having a small data set and you are having extreme values the mean value may be influenced by the presence of these extreme numbers.

(Refer Slide Time: 26:35)



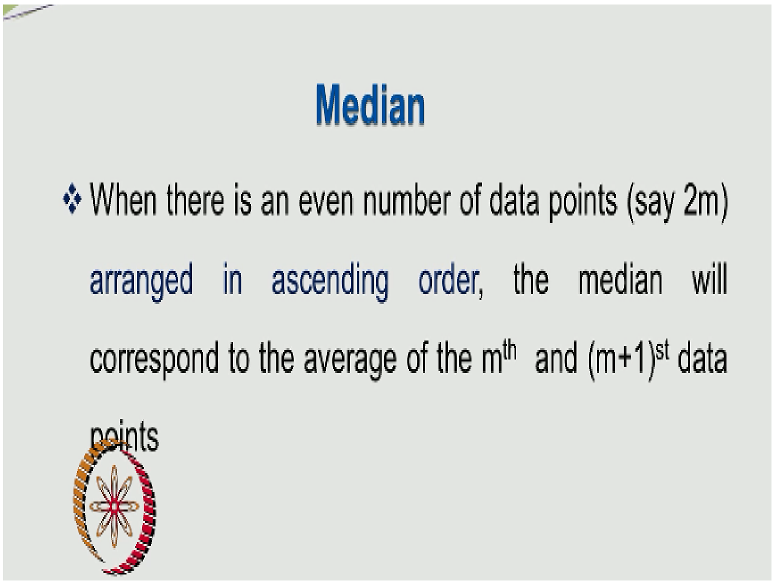
Median

- ❖ The median is the second quartile as we saw previously.
- ❖ When there are odd number of data points of size (2m+1) that are arranged in ascending order, the median will correspond to the $(m+1)^{st}$ data point.

The median is also a measure of the central value in the distribution and we also saw in the box plot discussion that the median is the second quartile how do we find the median. It depends upon whether the data set has odd numbers or even number you arrange the data points in the ascending order you put the smallest number first and the largest number last and then you find out the median.

If the number of data points in the sample is an odd number, then the calculation of median is quite simple. We identify the data points which is in the middle of this spread. Suppose you have 2 m+1 data points which are arranged in the ascending order the median will correspond to the m+1th data point.

(Refer Slide Time: 27:44)

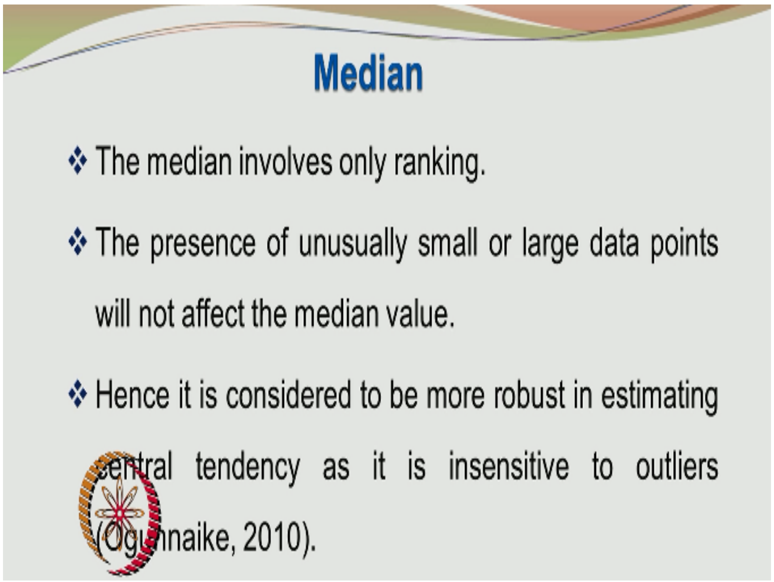


Median

- When there is an even number of data points (say 2m) arranged in ascending order, the median will correspond to the average of the m^{th} and $(m+1)^{st}$ data points

On the other hand, you have even number of data points say 2 m which are arranged in ascending order again. The median will correspond to the average of the mth and m+1th data point. Suppose you have let us say 4 numbers which are arranged in ascending order then you find out the second number which is m and then you find out the third number which is m+1 take the average of those 2 numbers to get the median.

(Refer Slide Time: 28:25)



Median

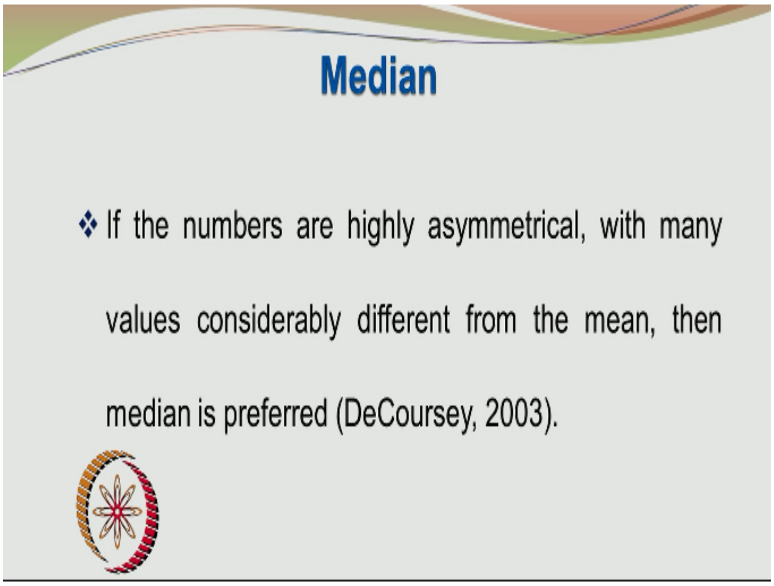
- ❖ The median involves only ranking.
- ❖ The presence of unusually small or large data points will not affect the median value.
- ❖ Hence it is considered to be more robust in estimating central tendency as it is insensitive to outliers (Ogunnaike, 2010).

The median involves only a ranking and the presence of unusually small or large data points will not affect the median value and hence it is considered to be more robust in estimating the central tendency of the distributions as it is not affected that much by the outliers. The extreme points they may take any value, but here you are not actually doing the adding and then dividing by the total number.

So the value of these numbers are really not affecting the calculation it is just ranking them and then seeing what is the number which is going to be there in the middle. And the outliers or obviously the extreme data points they are very, very low data values or very, very high ones. So you would not get outliers in the middle of a distribution it does not make any sense you will have outliers only in the extremes of the distribution.

So in that sense the median is the more robust way to find the central tendency of the distribution.

(Refer Slide Time: 29:45)

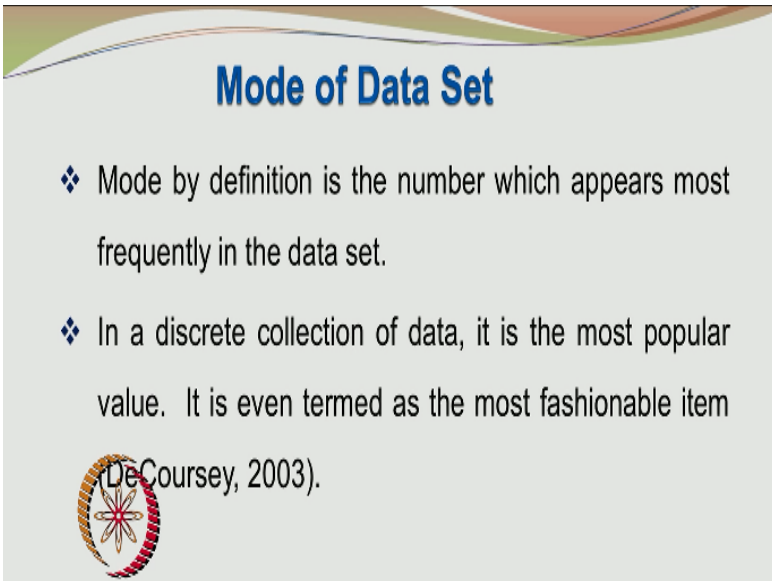


Median

- If the numbers are highly asymmetrical, with many values considerably different from the mean, then median is preferred (DeCoursey, 2003).

If the numbers are highly asymmetrical with many values considerably different from the mean, then the median is preferred.

(Refer Slide Time: 30:00)

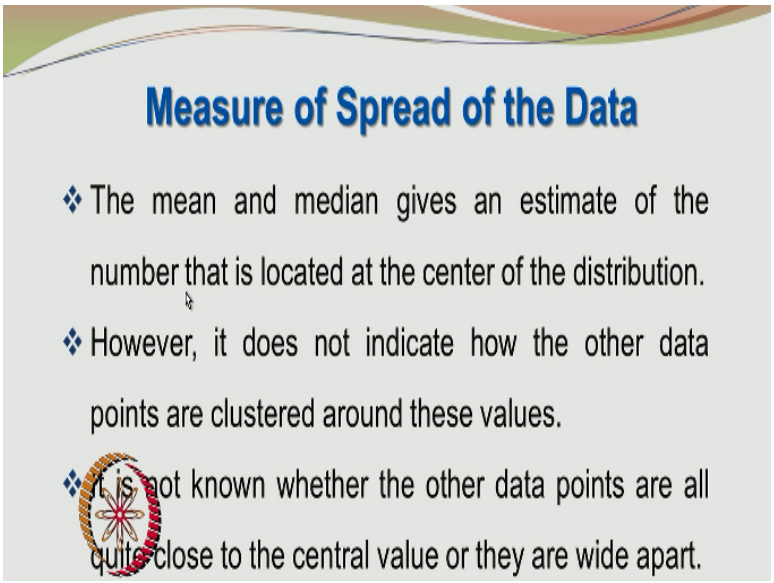


Mode of Data Set

- Mode by definition is the number which appears most frequently in the data set.
- In a discrete collection of data, it is the most popular value. It is even termed as the most fashionable item (DeCoursey, 2003).

We also have the mode. The mode by definition is the number which appears most frequently in the data set. This you might have studied in your high school itself. In the discrete collection of data, it is the most popular value. DeCoursey terms it as the even the most fashionable item in the data set. Who said numbers are dull they have very interesting properties.

(Refer Slide Time: 30:42)

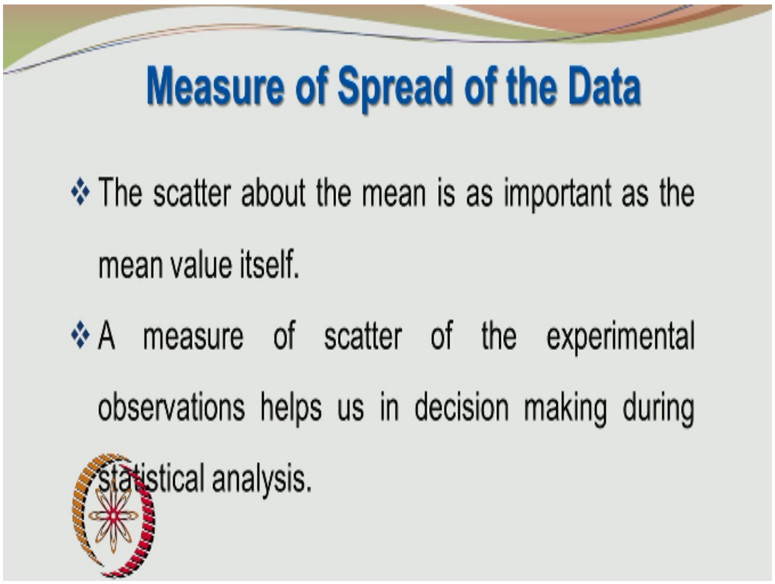


Measure of Spread of the Data

- The mean and median gives an estimate of the number that is located at the center of the distribution.
- However, it does not indicate how the other data points are clustered around these values.
- It is not known whether the other data points are all quite close to the central value or they are wide apart.

So now we want to look at the spread of the data. We have looked at the central tendency now we will look at the variability in the data. The mean and median gives an estimate of the number that is located at the center of the distribution. However, it does not indicate how the other data points are clustered around the central point whether the points are very close to the mean values or they are wide apart from the mean value.

(Refer Slide Time: 31:20)



Measure of Spread of the Data

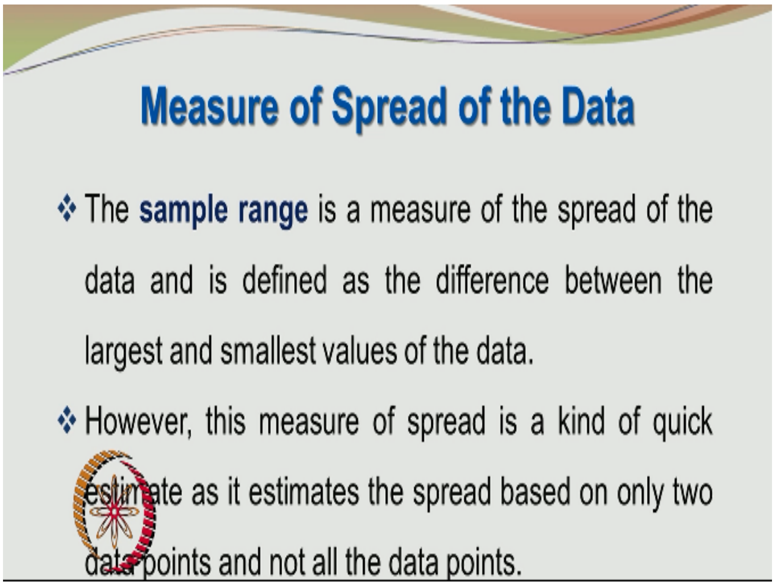
- The scatter about the mean is as important as the mean value itself.
- A measure of scatter of the experimental observations helps us in decision making during statistical analysis.

It is very important for us to know the scatter about the mean value. It is as important as knowing the mean value itself and the variability in the data is what influences us when we make decisions during experiments. The variance is the parameter which influences our decision making in statistical data analysis. What is variance? Variance is based on the deviation from the mean, but we know that the deviation from the mean add up to 0.

So our aim is not to get the actual values we want to get our overall idea about the spread. So whether is negative deviation or positive deviation we want to give them equal importance and so we square these deviations and once they are squared then there is no difficulty because the sum will not be=0 in most cases. So we will have the square of the deviations from the mean and we add up those deviation square.

And then divide it by a suitable number that suitable number we will discuss very soon.

(Refer Slide Time: 32:55)



Measure of Spread of the Data

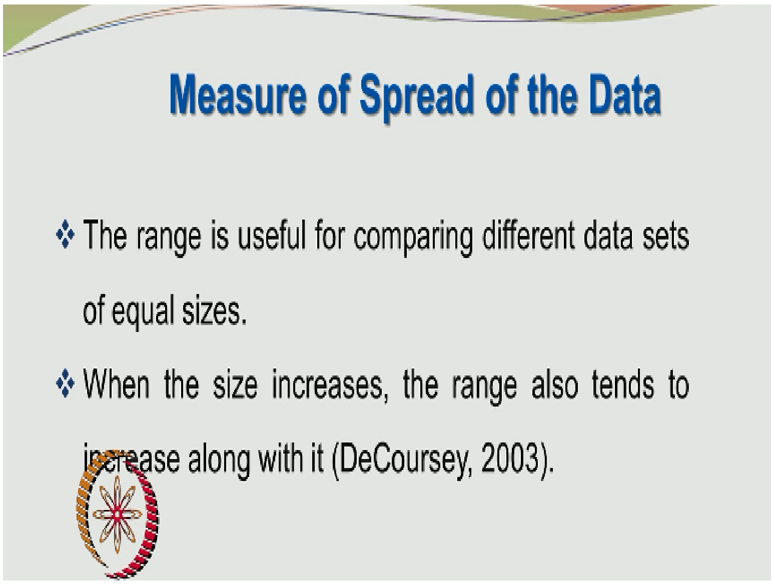
- The **sample range** is a measure of the spread of the data and is defined as the difference between the largest and smallest values of the data.
- However, this measure of spread is a kind of quick estimate as it estimates the spread based on only two data points and not all the data points.

The next measure of the spread is the sample range and it is defined as the difference between the largest and smallest values in the data set. This is a kind of a shortcut to estimate the spread of the data. We are using only 2 data points you are not considering the entire data set. Well if you want to look at the spread it will be better if all the data points are participating in the exercise.

You take only the smallest number and the largest number and find a difference that is usually not rigorous. It is a useful and a quick estimate, but it is not very rigorous. For example, the largest and the smallest data maybe outliers and so the other remaining data points may be very close to be mean value if you go by only the largest and the smallest value you may be over estimating the spread.

Whereas in the actual case the spread may have been quite smaller the overall spread would have been quite smaller then what was reported by the sample range.

(Refer Slide Time: 34:25)



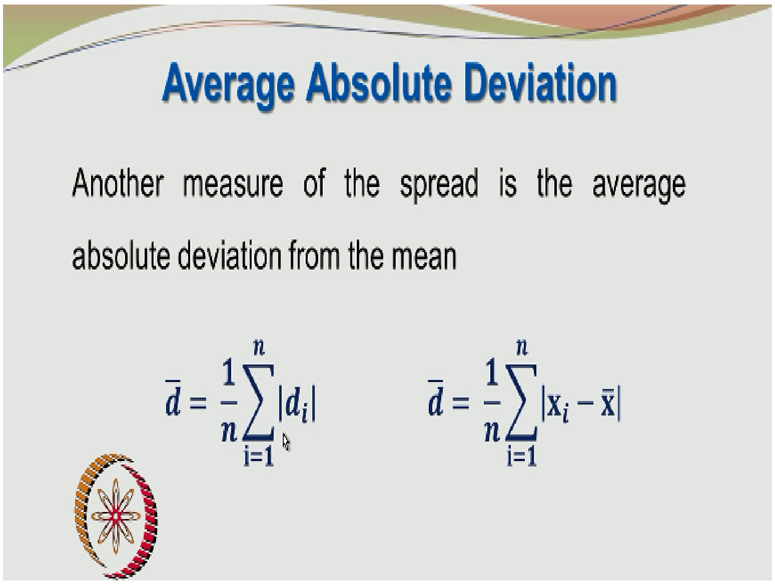
Measure of Spread of the Data

- The range is useful for comparing different data sets of equal sizes.
- When the size increases, the range also tends to increase along with it (DeCoursey, 2003).

The range is useful when you want to compare different data sets of equal sizes. DeCoursey observes that when the size of the data set increases the range also tends to increase along with it. In some research paper you might have come across the average absolute deviation as the name says there are different ways to handle the case where you have positive deviations and negative deviations. One way is to square them, but when you square them you are sort of changing the order of magnitude of the number.

If it is > 1 when you square it, you are having a higher order of magnitude If the number is < 1 you are going to have a lower order of magnitude after the squaring is done. This is slight manipulation of the data and of course after you take the variance you take the square root and get the standard deviation. Another way to handle this issue of positive or negative deviations from the mean is to ignore the sign.

(Refer Slide Time: 35:50)



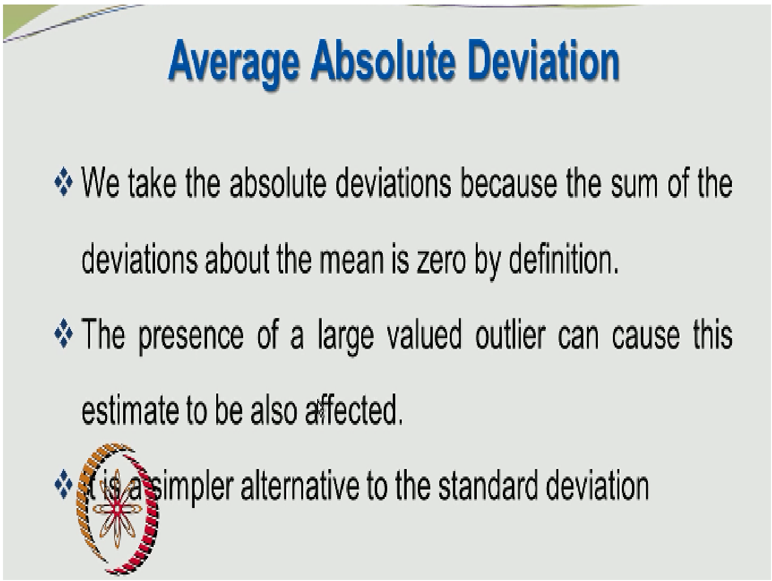
Average Absolute Deviation

Another measure of the spread is the average absolute deviation from the mean

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n}|d_i| \qquad \bar{d} = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

So what we do here is we take the absolute value of the deviation from the mean. We write it as d bar which is=1/n sigma= 1 to n di. Where di is the deviation xi from the arithmetic mean.

(Refer Slide Time: 36:10)



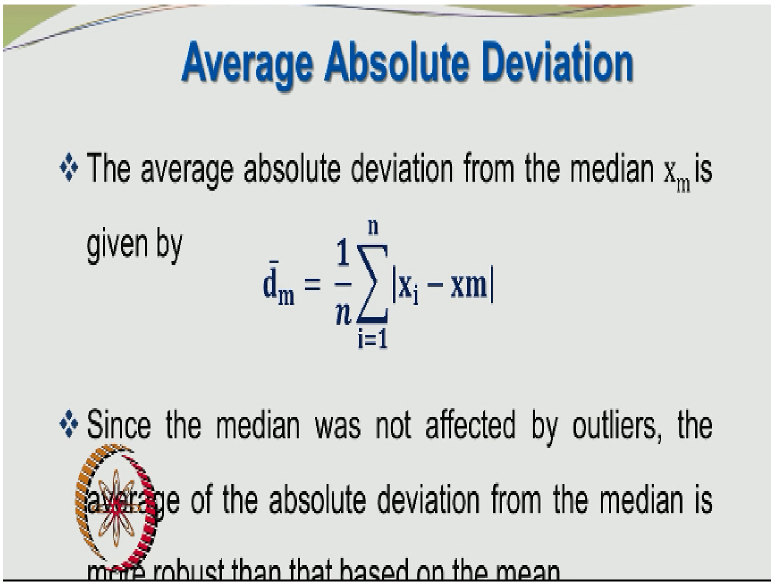
Average Absolute Deviation

- We take the absolute deviations because the sum of the deviations about the mean is zero by definition.
- The presence of a large valued outlier can cause this estimate to be also affected.
- It is a simpler alternative to the standard deviation

So as far as the average absolute deviation is concerned the presence of a large valued outlier can cause this estimate to be also affected. When you want to present your deviations between your model predictions and exploratory data you may want to do so by using the average absolute deviation. It may so happen that except in one case in all other cases the data is matching rather well with the model predictions. However, because of one outliers the model may be showing a prediction much different from that of the experimentally observed value.

And this may increase your average absolute deviation and make it appear as if the comparison between the model and experimental data is not that good. So you may have to check for this outlier of course you have to go into the route of the matter rather than simply removing the outlier. The average absolute deviation is simpler alternative to the standard deviation.

(Refer Slide Time: 37:24)



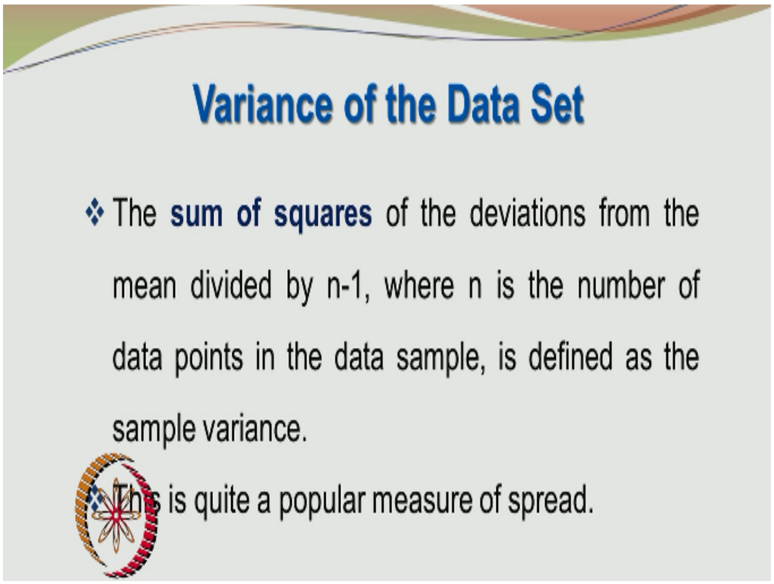
Average Absolute Deviation

- The average absolute deviation from the median x_m is given by $$\bar{d}_m = \frac{1}{n}\sum_{i=1}^{n}|x_i - xm|$$
- Since the median was not affected by outliers, the average of the absolute deviation from the median is more robust than that based on the mean

So now we will be again talking about average absolute deviation, but this time we will talking about it with respect to the median value. Earlier we were talking with respect to the arithmetic mean now we are going to find the deviation with respect to the median. So a small typo is there I will just correct it the subscript has again not been implemented so I will just make the subscript.

So the average of the absolute deviation from the median is more robust than that based on the mean. It is pretty useful and we denote the absolute deviation from the median in terms of d bar m.

(Refer Slide Time: 38:31)

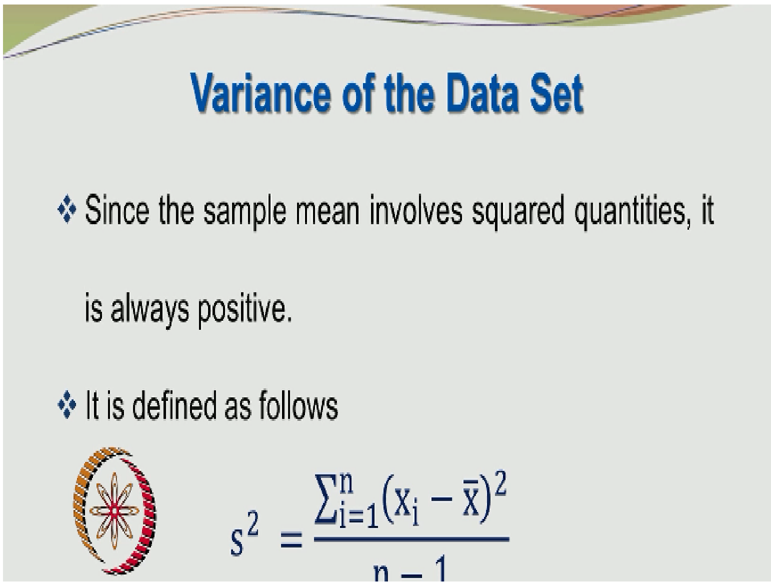


Variance of the Data Set

- The **sum of squares** of the deviations from the mean divided by n-1, where n is the number of data points in the data sample, is defined as the sample variance.
- This is quite a popular measure of spread.

Now let us come to the most rigorous form of identifying the spread in the data. Here we find the sum of squares of the deviations from the mean and divided by n-1 where n is the number of data points in the data set or sample. This is called as sample variance and this is very popular.

(Refer Slide Time: 38:57)



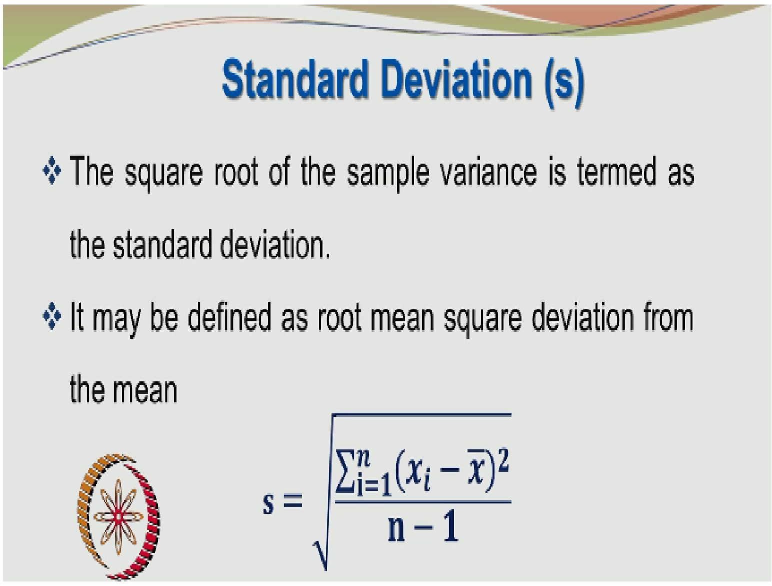
Variance of the Data Set

- Since the sample mean involves squared quantities, it is always positive.
- It is defined as follows

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Please note that the variance is always a positive quantity because all the deviations have been squared and all of them have now become positive. Of course we have to deal with only a real number we do not deal with imaginary quantities. So after squaring we always have positive values with us. The mathematical formula or equation for variance is given by S square= sigma i= 1 to n Xi-X bar whole square/ n-1.

(Refer Slide Time: 39:42)



Standard Deviation (s)

- The square root of the sample variance is termed as the standard deviation.
- It may be defined as root mean square deviation from the mean

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

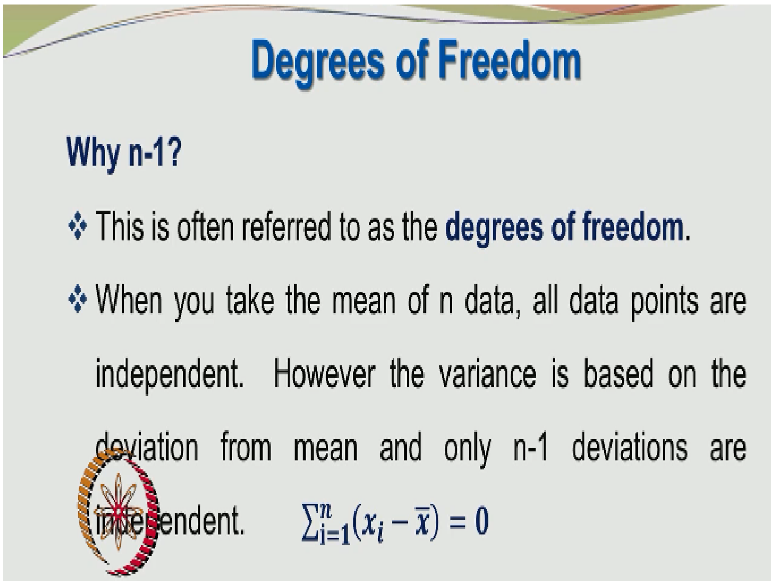
To find the standard deviation s we simply take the square root of the variance. This is the squared deviation Xi- X bar whole bar is referred to as the squared deviation and since we are adding all those squared deviation and dividing it by n-1 we have a mean square deviation and then we take the square root. So the standard deviation is also referred to as the root mean square deviation from the mean. This concept of mean square is very important and we will encounter this frequently in our design and analysis of experiments.

We call it as the mean square error. You may want to refer to the first lecture the introduction where we talked about mean square error for the fertilizer example. Please note that the mean can have negative value. It depends upon the range of number I have already told that in one of the earlier lectures. The standard deviation also has the same units as that of the data set. If you are having the data set in terms of particle diameter the standard deviation will also be expressed in terms of particle diameters.

They may be referring to particle diameters so the dimension would be more appropriately micrometers then the standard deviation will also be in micrometers. The mean of this distribution will also be in micrometers. So the standard deviation and the mean will have the same units. The mean however can take negative values whereas the standard deviation is the positive square root of the variance.

So the standard deviation will have only positive values. In the formula, we used to find the variance or the standard deviation we used n-1 why did we use n-1 why not n. We used n in the calculation of the mean whereas in calculation for the variance we are using n-1.

(Refer Slide Time: 42:40)



Degrees of Freedom

Why n-1?

- This is often referred to as the **degrees of freedom**.
- When you take the mean of n data, all data points are independent. However the variance is based on the deviation from mean and only n-1 deviations are independent. $\sum_{i=1}^{n}(x_i - \overline{x}) = 0$

The term n or n-1 is referring to the degrees of freedom. N of course stands for the size of the data set. We are looking at the number of independent entities in the data set. If you collect the data set the entities in those have been chosen in such a way that they are independent. So when you are finding the mean value you are dealing with n independent entities. However, when you are calculating the standard deviation or the variance you are basing those calculations on the deviation from the mean value not all the deviations are independent.

The mean has been defined in such a way that the sum of the deviations from the mean will be=0. So this acting like a constraint. So you have only n-1 deviations from the mean that are independent. So this is an interesting situation how do we deal with this okay. So the number of independent entities is only n-1 and so we use n-1 in the calculation of the sample variance. There is also another reason why we use n-1 I will come to it shortly.