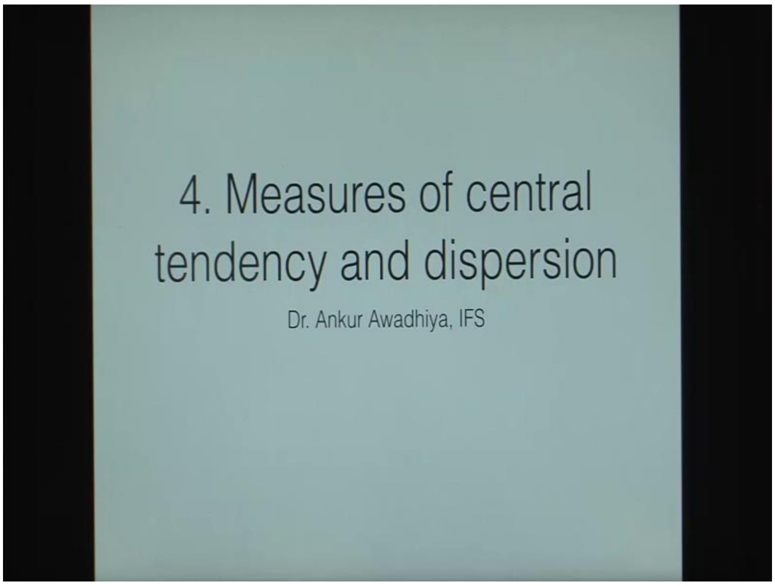
# Forest Biometry

Prof. Mainak Das
Dr. Ankur Awadhiya
Department of Biological Sciences & Bioengineering & Design Programme
Indian Institute of Technology, Kanpur

## Lecture – 04
## Measures of central tendency and dispersion

Today we will look at the measures of central tendency and dispersion. Let us look at a distribution.

(Refer Slide Time: 00:22)

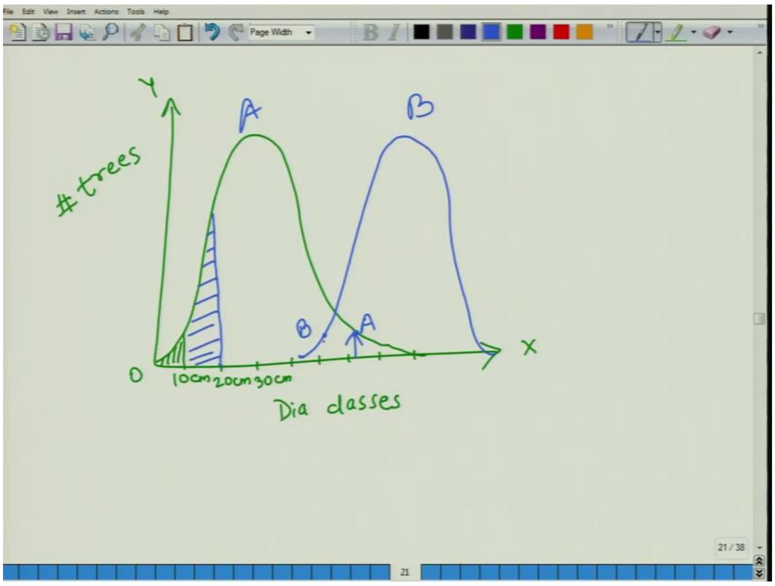


4. Measures of central tendency and dispersion

Dr. Ankur Awadhiya, IFS

So, for instance suppose we want to plot.

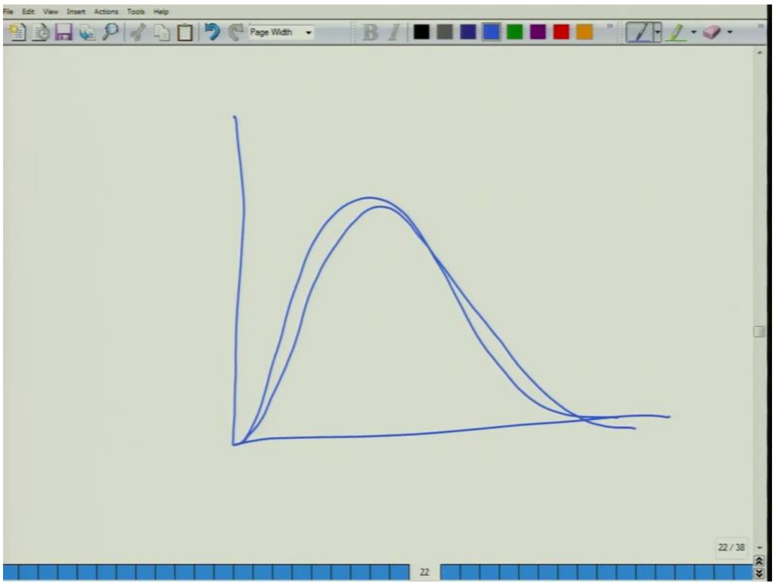
(Refer Slide Time: 00:28)



The number of trees on the y axis and the diameter classes on the x axis. And let us plot a distribution like this. This thing is called a bell shaped distribution. So, basically if we suppose this is 10 centimeters, this is 20 centimeters, this is 30 centimeters and so on, what this distribution tells us is that these many number of trees are have a diameter of 0 to 10 centimetres, whereas, these trees have a diameter between 10 to 20 centimeters and so on. Now if I were to ask you is there a single measure that can describe this distribution. Now why would such a measure be required? Suppose we wanted to compare between 2 distributions, one is this and the second distribution goes like this. Now if I wanted to ask you which of these has trees of a greater diameter class. Of course, we can intuitively see that this distribution let us call it A and let us call this distribution as B.

So, intuitively we can say that because trees in distribution A are on the on the left side whereas, tree is on in the in the distribution B are on the right side. So, probably distribution Be has trees of a greater diameter class, but consider this point, a point here and a point here. Now this point lies in distribution A and this point lies in distribution B. So, here we can see that we have some trees that are there in distribution A that have a greater diameter as compared to some trees that are there in distribution B. Now in this particular example the situation is simple because both these distributions are very far apart, but let us consider some other distributions.

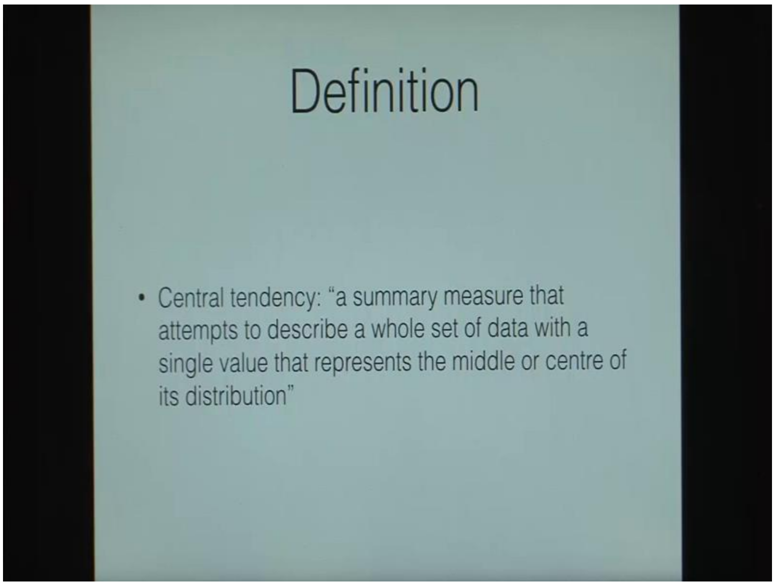
(Refer Slide Time: 02:59)



Suppose We had a distribution that went like this, and then there was another distribution that went like this. What about in this case? Can we see anything with a with a greater certainty it will be difficult right.

So, to solve these problems we need to get one value that can depict a distribution. And a measure of central tendency gives us that value.

(Refer Slide Time: 03:25)

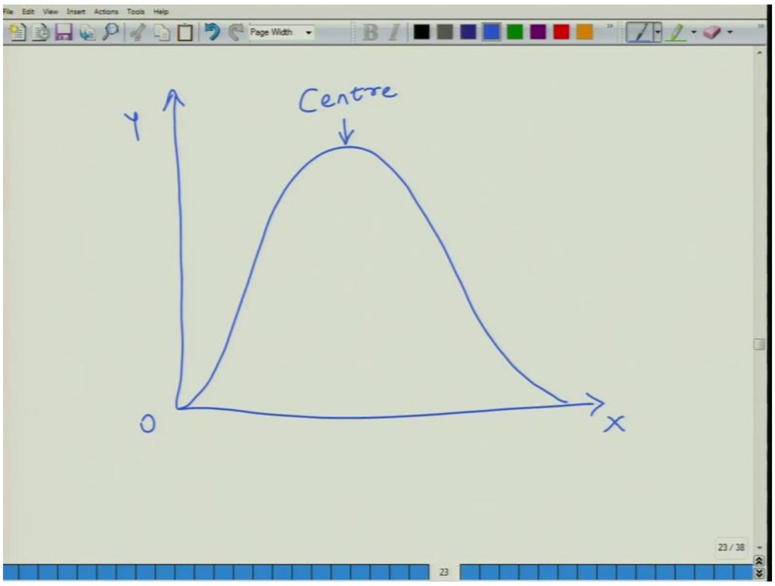


## Definition

- Central tendency: "a summary measure that attempts to describe a whole set of data with a single value that represents the middle or centre of its distribution"

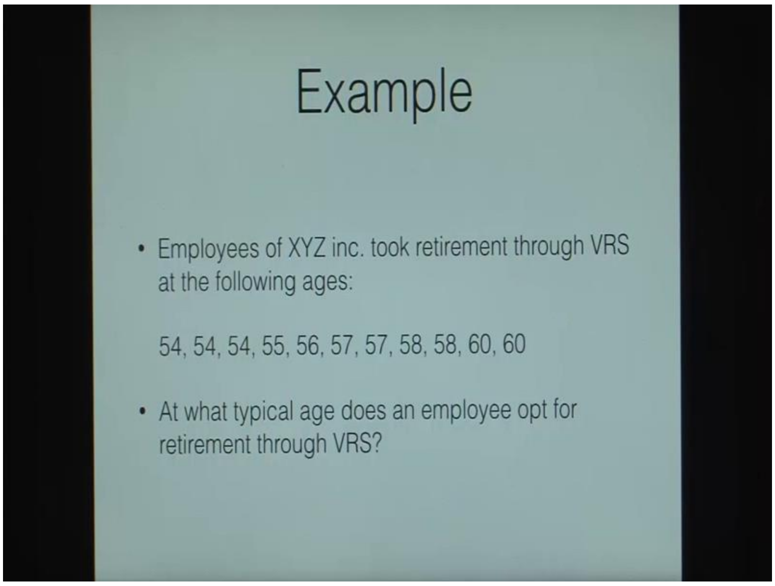
A central tendency is defined as a summary measure that attempts to describe a whole set of data with single value that represents the middle or center of it is distribution. So, essentially if we have a distribution, and if we were to say that this is the center of the distribution. Then this would be a measure of the central tendency.

(Refer Slide Time: 03:48)



Now let us look at this point in more detail with an example. So, as you can see on your screen employees of xyz incorporated took retirement through VRS that is voluntary retirement scheme at the following ages.

(Refer Slide Time: 04:09)



## Example

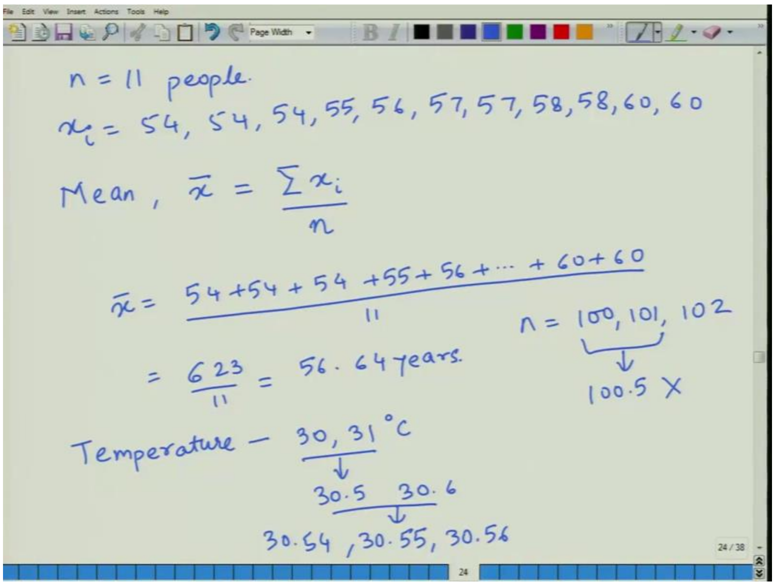
- Employees of XYZ inc. took retirement through VRS at the following ages:

  54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

- At what typical age does an employee opt for retirement through VRS?

The ages are 54 54 54 55 56 57 57 58 58 60 and 60. At what typical age does an employ up for retirement through VRS? So, essentially we have a number of values n equal to 1 2 3 4 5 6 7 8 9 10 11.

(Refer Slide Time: 04:34)



$n = 11$ people.

$x_i = 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60$

Mean, $\bar{x} = \frac{\sum x_i}{n}$

$\bar{x} = \frac{54+54+54+55+56+\cdots+60+60}{11}$

$= \frac{623}{11} = 56.64$ years.

$n = \underbrace{100, 101}_{\downarrow}, 102$

100.5 X

Temperature – $\underline{30, 31}$ °C

$\downarrow$

$\underline{30.5 \quad 30.6}$

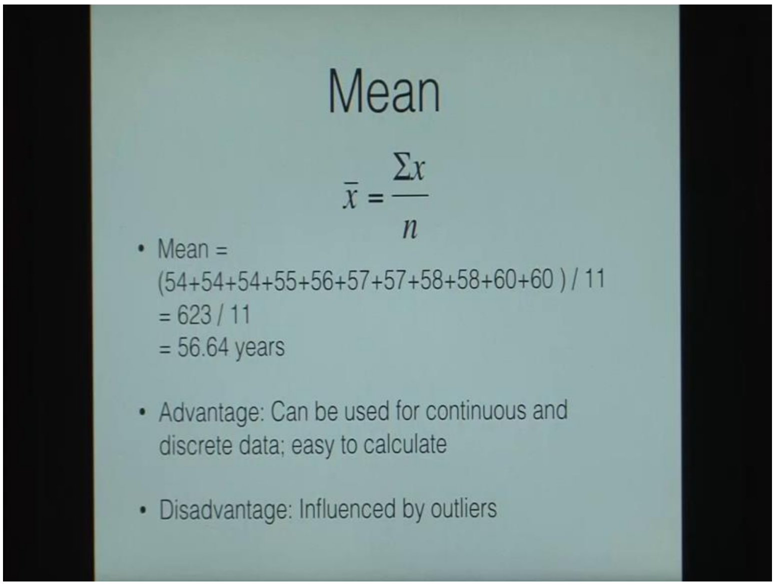
$\downarrow$

30.54, 30.55, 30.56

So, we have 11 people. And the values of x are 54 54 54 55 56 57 57 58 58 60 and 60. So, 1 2 3 4 5 6 7 8 9 then 11 values. Now we want to find out what is the typical age at which an employee opts for retirement through the voluntary retirement scheme. So, this would give us an idea of the measure of the central tendency. So, let us look at the first measure. The first measure is called the mean. Mean that is shown by the symbol x bar is given by the sum of the values sum of over xi whole divided by the total number of values. So, in this case if you wanted to calculate the mean x bar would be 54 plus 54 plus 54 plus 55 plus 56 plus So on, till you have the last 2 values whole divided by 11.

Which comes out to be 623 by 11 equal to 56.64 years. So, this is the average age or the mean age at which an employee opts for retirement through the VRS. Now calculation of a mean is an easy task, besides it also has some other advantages.

(Refer Slide Time: 06:23)



Mean

$$\bar{x} = \frac{\Sigma x}{n}$$

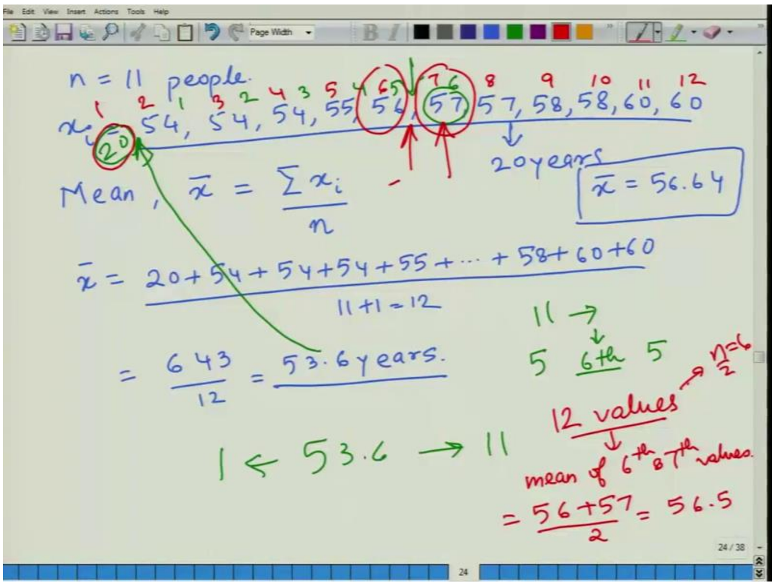
- Mean =
  (54+54+54+55+56+57+57+58+58+60+60 ) / 11
  = 623 / 11
  = 56.64 years
- Advantage: Can be used for continuous and discrete data; easy to calculate
- Disadvantage: Influenced by outliers

The advantage of using a mean is that it can be used for continuous and discrete data, and it is easy to calculate and interpret. So, when we say continuous and discrete data what do we mean, consider a variable called temperature. Now temperature if we consider 2 values 30 and 31 degree celsius, we can always have another value in between that is 30.5, now if we considered 30.5 and 30.6 we can always have another value that is in between 30.55 for instance or 30.56 for instance. Or maybe 30.54 for instance. Now such variables that vary continuously are called continuous variables.

On the other hand we have some discrete variables, discrete variables like number of trees in a forest. And so, the number of trees can be 100 it can be 101 or 102, but you cannot have any value in between 101 or 100 and 101. So, you can never have 100.5 freeze that is not possible. So, those are discrete data. Now mean is a value that can be calculated both for continuous data and for discrete data.

It is easy to calculate and interpret, because we know that it is the sum of the values some of the values divided by the total number of values. So, this is the average. An average is something that we are very used to using intuitively; however, the problem with mean is that it is greatly influenced by outliers. So, suppose amongst these values you added a single extra value and that was 20 years. So, how would the calculation vary in this case? Remember that currently we had the average of 56.64 years. Now if we added one extra value of 20 years what would the mean be? So, let us calculate the mean again. So now, x bar would be equal to 20 plus 54 plus 54 plus 54 plus 55 plus so on, till 58 60 60 whole divided by now the total number of values becomes 11 plus 1 equal to 12.

(Refer Slide Time: 08:40)



n = 11 people.

$x_i$ = 20, 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60

20 years

Mean, $\bar{x} = \frac{\sum x_i}{n}$

$\boxed{\bar{x} = 56.64}$

$\bar{x} = \frac{20+54+54+54+55+\cdots+58+60+60}{11+1=12}$

$= \frac{643}{12} = \underline{53.6 \text{ years.}}$

11 →

5 $\underline{6\text{th}}$ 5

1 ← 53.6 → 11

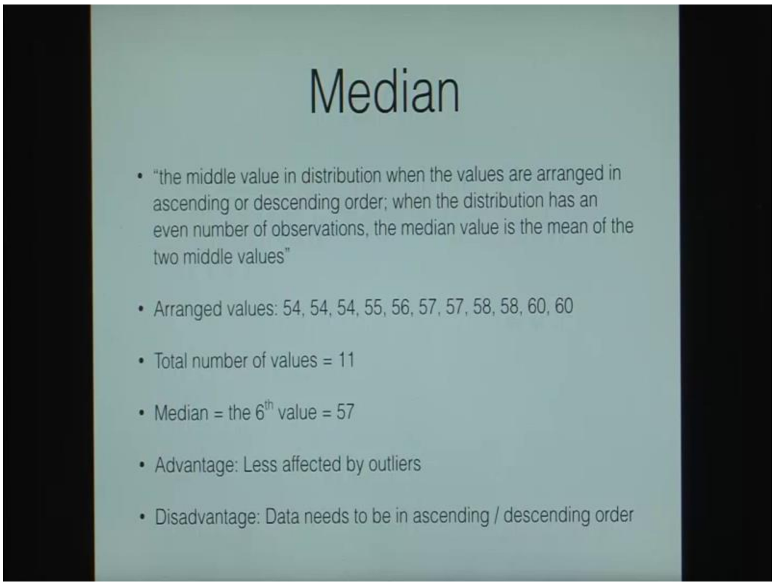
$\underline{12 \text{ values}}$ → $\frac{n}{2} = 6$

mean of 6th & 7th values

$= \frac{56+57}{2} = 56.5$

So now this value comes to be 643 divided by 12, which is 53.6 years. Now consider the original value 56.64 years. If you look at the distribution, we will find that 56.64 lies somewhere in between. But now that we have added a single value of 20 in front of it, this mean has become 53.6 years which is right next to 20. So, essentially one value lies to the left side of 53.6 years and we have 11 values on the right side of it. So, can we still call it a measure of the central tendency? And how good would it be to be used for our distributions. So, these are the limitations of mean it is very much influenced by the outliers. Which is why we need another measure of central tendency which is called the median.

(Refer Slide Time: 10:16)



# Median

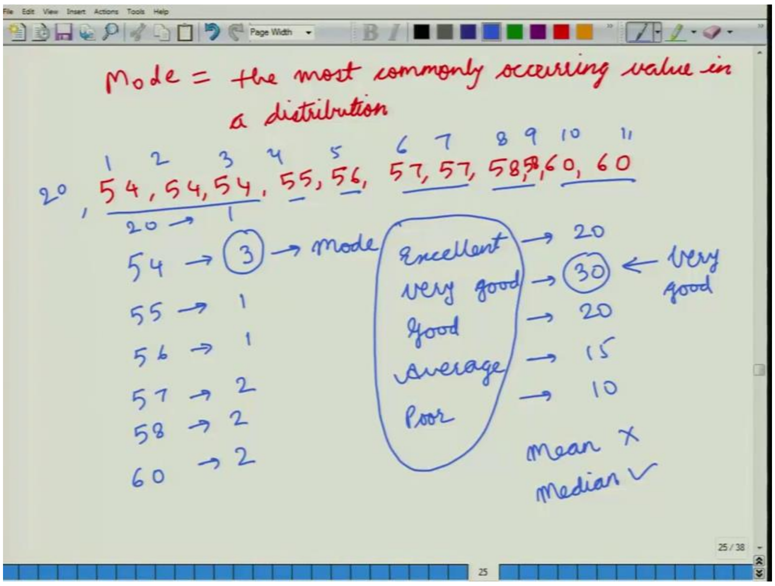
- "the middle value in distribution when the values are arranged in ascending or descending order; when the distribution has an even number of observations, the median value is the mean of the two middle values"
- Arranged values: 54, 54, 54, 55, 56, 57, 57, 58, 58, 60, 60
- Total number of values = 11
- Median = the $6^{th}$ value = 57
- Advantage: Less affected by outliers
- Disadvantage: Data needs to be in ascending / descending order

So, if we consider the original distribution 54, 54, 54 and so on, and if we wanted to calculate the median the medium the median would be defined as the middle value in the distribution, when the values are arranged in an ascending or descending order. When the distribution has an even number of observations the median value is the middle of the 2 middle values, it is the mean of the 2 middle values.

So, in the case of 11 people. So, 11 is an odd number. So, we can have 5 to the left of it 5 to the right of it and the 6th value would be the mean value or would be the median value. So, let us find out the 6th value this is 1 2 3 4 5 6 57. Now 57 is the median value in this particular case. Now consider our original example. Suppose we added one extra value that was 20. So, after we have added 20, 20 becomes the first value this becomes the second third 4th 5th 6th 7th 8th 9 10 11 and the twelfth value. Now in the case of 12 values because this is an even number the median would be the mean of 6th that is n by 2 is 6 and 7th values. So, essentially here the 6th value is 56. So, it becomes 56 plus the 7th value is 57 by 2 which is 56.5. Now even though we have added a very extreme value of 20 years, the median has shifted from 57 to 56.5. So, median is very less affected by outliers. So, that is an advantage of using the median.

However the one disadvantage of using a median is that your data needs to be arranged in ascending or descending order in order to get the median. So, that is easy if you have a small data set, but when you have a larger data set then it becomes more of involved. The third measure of central tendency is called the mode.

(Refer Slide Time: 12:46)



Mode = the most commonly occurring value in a distribution

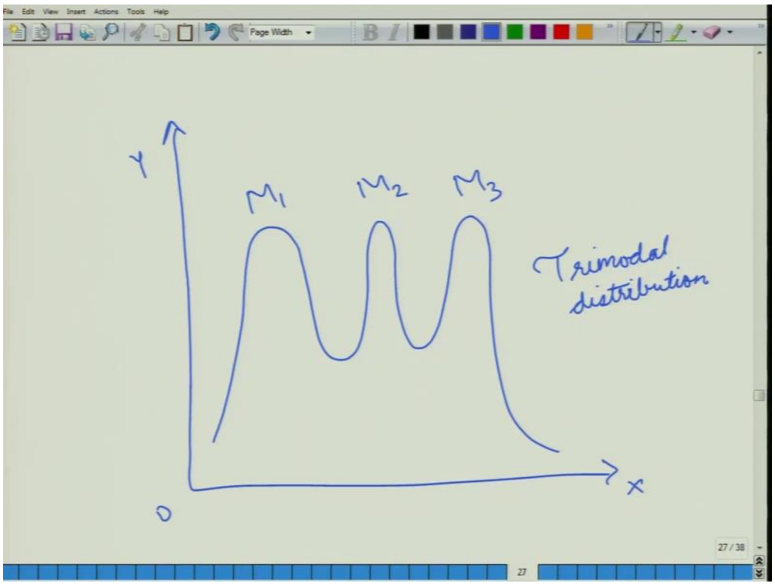
And a mode is defined as the most commonly occurring value in a distribution. So, what does that mean? Again if we look at our original values we had 54 54 54 55 56 57 57 58 60 and 60. This is before we added odd number 20. So, this is the first second third 4th 5th 6th 7th 8th 9th 10th maybe you have missing one 58 58 60 and 60. So, this is the 9th 10th and the 11th values. Now which of these values occurs most frequently most commonly. So, if we wrote down these values 54 55 56 57 58 and 60. How many times do we have 54? We have it 3 times, 55 occurs only once, 56 also occurs only once.

57 occurs twice, 58 occurs twice and 60 occurs twice. So, in this particular example 54 is occurring the maximum number of times that is 3 times. So, this is the mode. So, the mode is the most commonly occurring value in a distribution. Now what is the advantage of using a mode? One it is also a very less affected by outliers. So, suppose we added this number 20 here. So, 20 add it here and it occurs only once. And the mode remains unchanged it remains the same as before 54. Another example oh another advantage of using a mode is that it is useful even for a non numerical data. So, for instance if you wanted to ask people to rate something, on a scale that said average good very good excellent and maybe poor. And suppose the number of people who opted for these values was say 20 said it was excellent, 30 said it was very good, 20 again said it was good

average was said by 15 people and 10 people said that it was poor. Here also we can calculate the mode it would be very good. Because it is opted for by the maximum number of people it is the most commonly occurring value in this distribution.

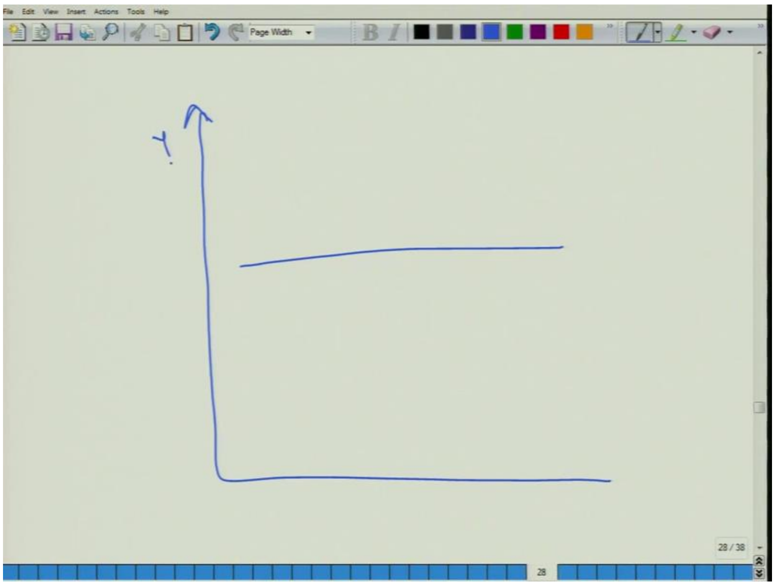
However In the case of these non numerical data excellent very good average and poor we cannot calculate a mean. And also we cannot calculate a median or maybe median would be possible in some cases, but mean definitely cannot be calculated in this case; however, there is a disadvantage in using a mode.

(Refer Slide Time: 16:30)



Some distributions Suppose if you have a distribution like this. So, this value occurs the most number of times. So, this is the mode. So, we can very easily calculate a mode for a unimodal distribution. And this is something that we can very easily use as well; however, consider this distribution. Now this distribution would have 3 modes, mode 1 2 and three. So, this is a trimodal distribution. Now remember that we started our discussion on the measure of central tendency by asking whether we could have a single value that could describe this distribution. So now, if we have a trimodal distribution what do we make of these 3 modes? Similarly you can have a distribution that is flat.

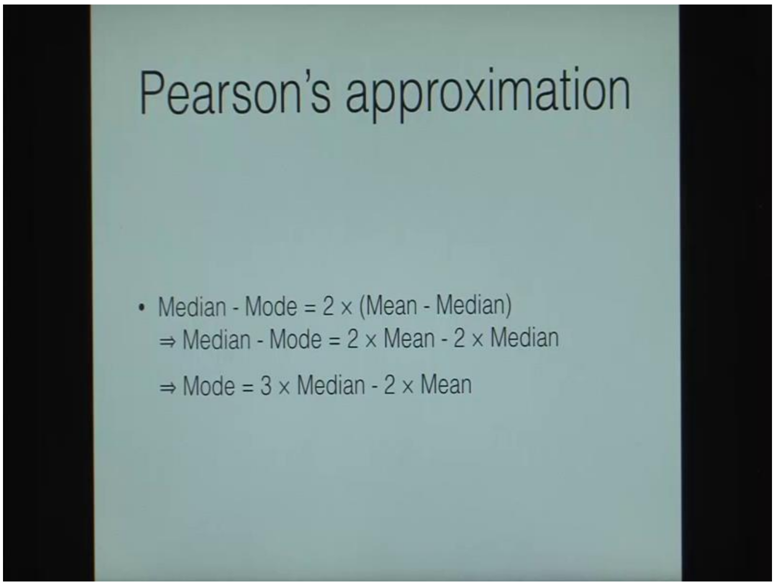
(Refer Slide Time: 17:42).



And in this case you will not have any single mode. So, you can have a non modal distribution or you can have a bimodal trimodal or a multimodal distribution.

Now, mode can be used intuitively only in the case of a unimodal distribution. So, that is a disadvantage in using the mode. Now pearson has given us an approximation for most of the distributions.

(Refer Slide Time: 18:14)

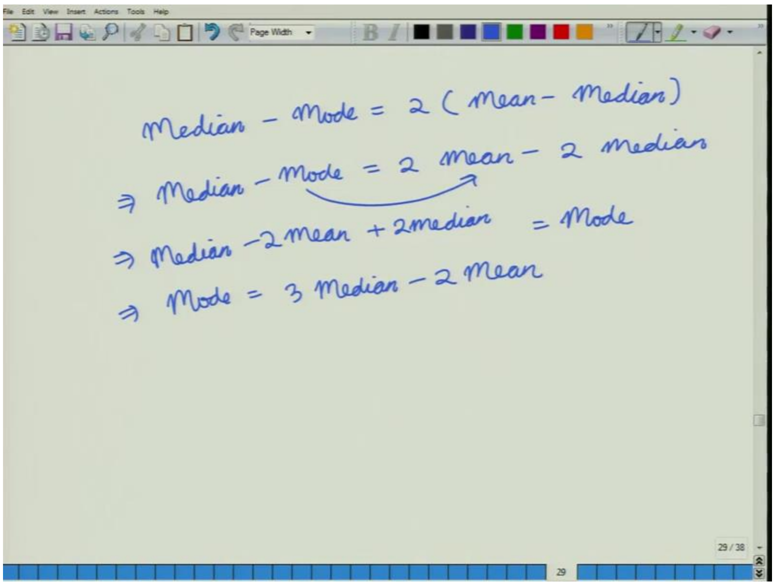


Pearson's approximation

- Median - Mode = 2 x (Mean - Median)
  ⇒ Median - Mode = 2 x Mean - 2 x Median
  ⇒ Mode = 3 x Median - 2 x Mean

This holds valid, it goes like median minus mode equals twice mean minus median.

(Refer Slide Time: 18:18).



Median − Mode = 2 ( Mean − Median)

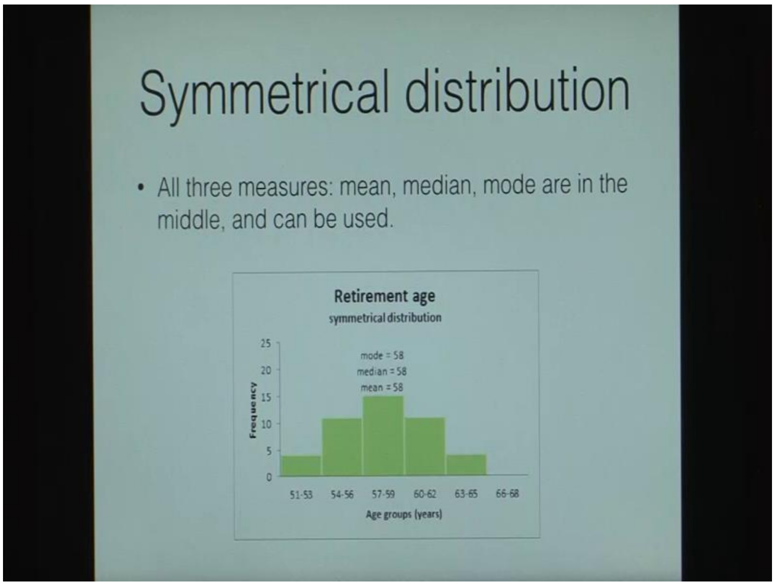
⇒ Median − Mode = 2 mean − 2 median

⇒ Median −2 mean + 2median = Mode

⇒ Mode = 3 Median − 2 Mean

Implying that median minus mode equals twice mean minus twice median. So, bringing mode to this side and all the other entries to this to the other side we would have mode equals median minus 2 mean plus 2 median. Or we could also write it as mode equals 3 median minus 2 mean. Now this distribution is in this formula is an approximation. So, it may not always be correct; however, it is useful for some cases.

(Refer Slide Time: 19:27)

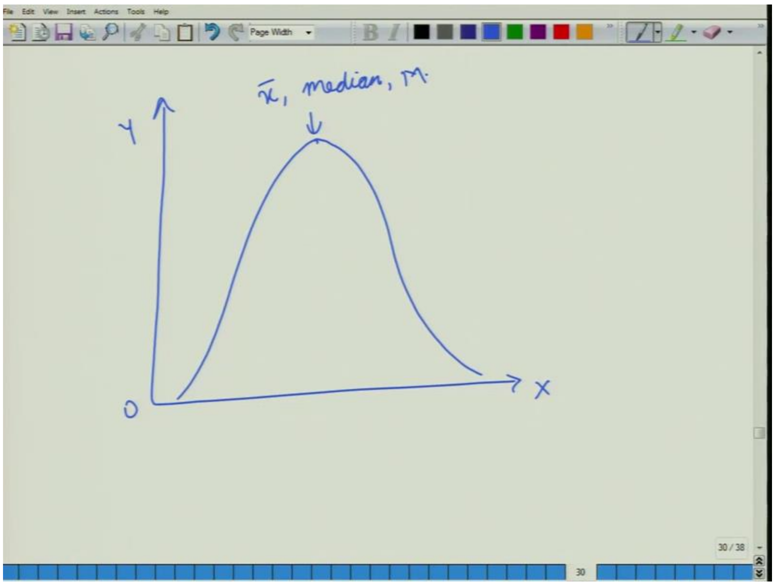


## Symmetrical distribution

- All three measures: mean, median, mode are in the middle, and can be used.

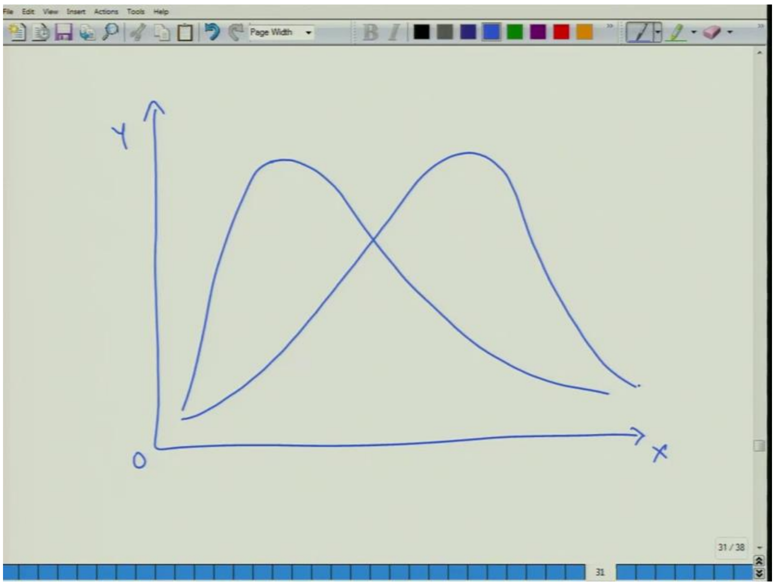
Now, when do we use which cen which measure of central tendency. If you look at a symmetrical distribution.

(Refer Slide Time: 19:35)



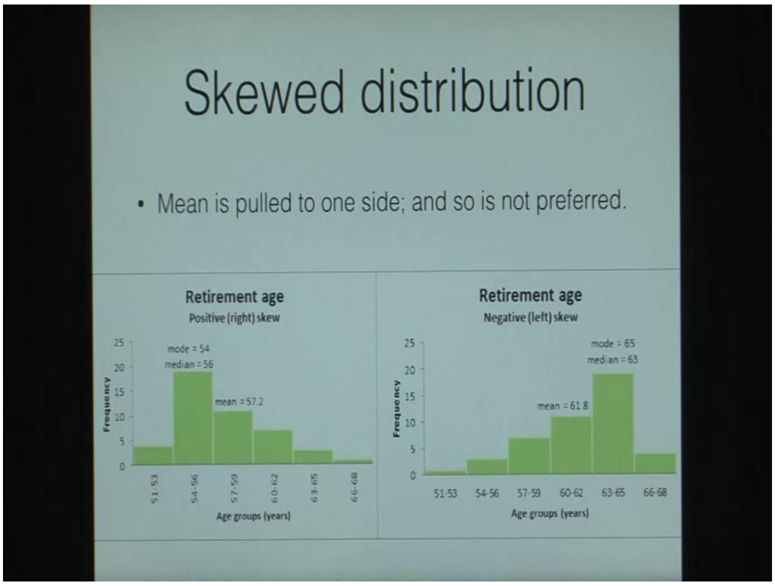
In this case mean median and mode will all be in the center. So, so we can use any one of these 3 values; however, if a distribution is skewed to one side.

(Refer Slide Time: 20:01)



Something like this, then we will have the mean pull to one side, but median and mode will remain the same. So, in this case mean is not used. Similarly we could have a distribution with a skew on the other side.

(Refer Slide Time: 20:36)

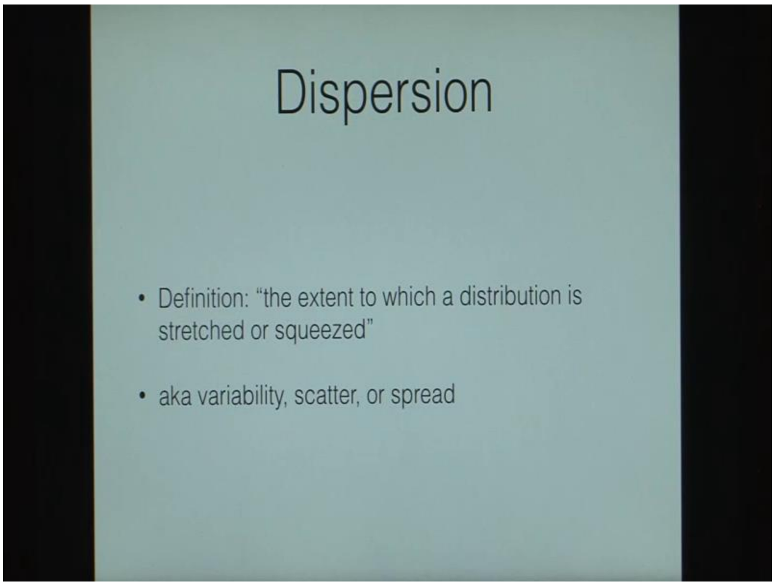


# Skewed distribution

- Mean is pulled to one side; and so is not preferred.

So, in a skewed distribution it is always preferable to use a median or a mode, in case of a mean; now a related concept to that of the measure of central tendency lot of dispersion.

(Refer Slide Time: 20:44)

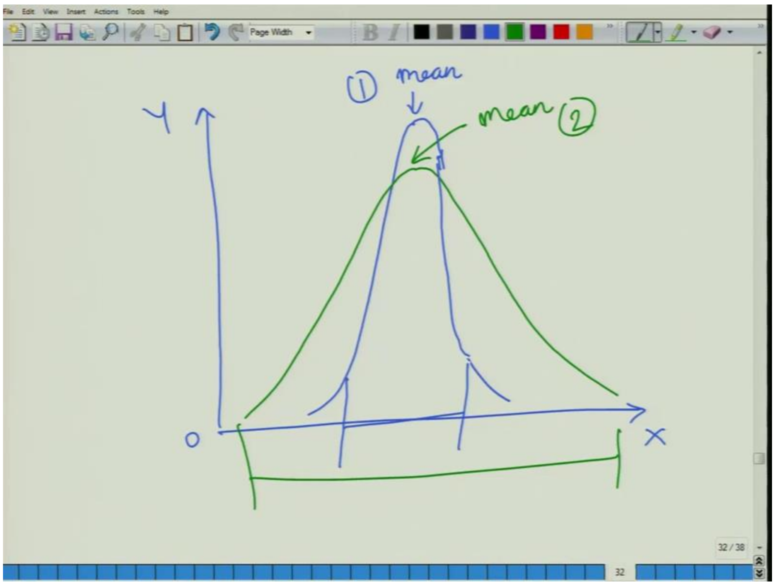


# Dispersion

- Definition: "the extent to which a distribution is stretched or squeezed"
- aka variability, scatter, or spread

A dispersion is defined as the extent to which a distribution is stretched or squeezed, it is also known as variability scatter or spread.

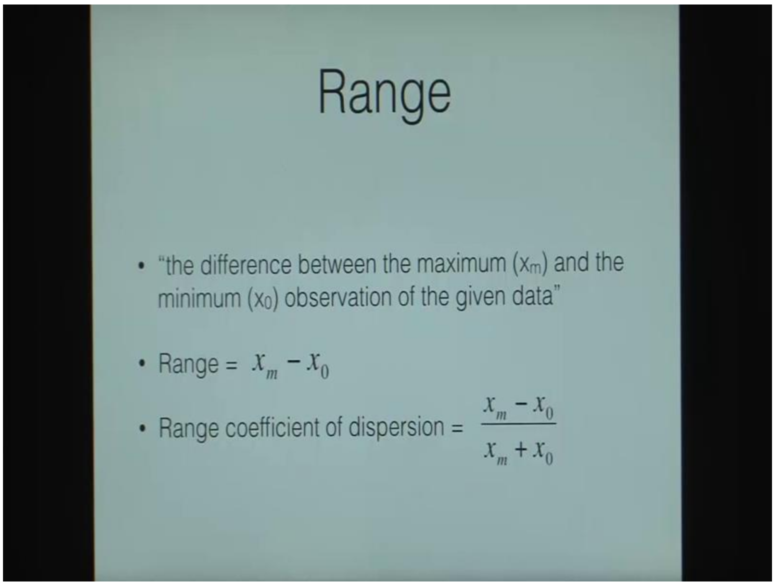
(Refer Slide Time: 20:56)



So, again let us look at this distribution, now this distribution has a mean here. Now consider another distribution, that goes like this. Now both these distributions have the same mean; however, we can observe that this distribution say let us call this distribution one and let us called this is distribution 2. Now distribution one has a smaller spread.

Whereas distribution 2 has a greater spread. So, while both these distributions have the same measure of central tendency, they have very different dispersions. So, how do we get a value for dispersion. Well dispersion is defined by several different Measures of dispersion.

(Refer Slide Time: 22:12)

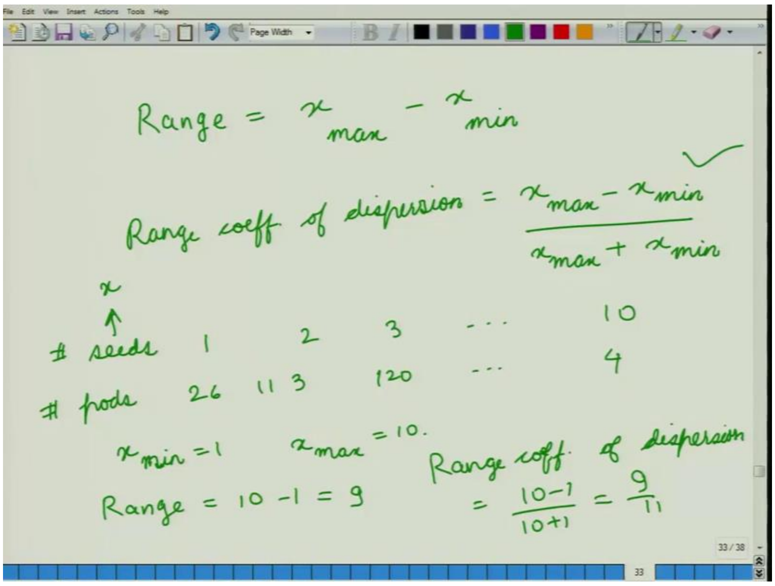


## Range

- "the difference between the maximum ($x_m$) and the minimum ($x_0$) observation of the given data"
- Range = $x_m - x_0$
- Range coefficient of dispersion = $\dfrac{x_m - x_0}{x_m + x_0}$

One of which is called the range. A range is defined as the difference between the maximum and the minimum value observed in a given data. We can also calculate the range coefficient of dispersion which is given by dividing the range by the sum of the maximum and the minimum values, or to put it mathematically range Equals x max minus x min.

(Refer Slide Time: 22:38)



Range = $x_{max} - x_{min}$

✓

Range coeff of dispersion = $\dfrac{x_{max} - x_{min}}{x_{max} + x_{min}}$

$x$

$\uparrow$

| | | | | | |
|---|---|---|---|---|---|
| # seeds | 1 | 2 | 3 | ... | 10 |
| # pods | 26 | 113 | 120 | ... | 4 |

$x_{min} = 1$ $\quad x_{max} = 10.$

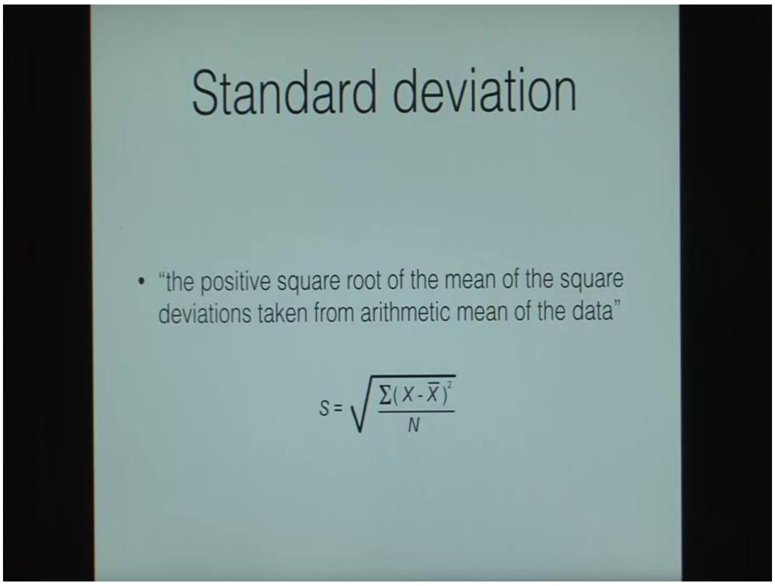
Range = $10 - 1 = 9$

Range coff. of dispersion $= \dfrac{10-1}{10+1} = \dfrac{9}{11}$

So, these are the maximum and the minimum values. On the other hand the range coefficient of dispersion equals x max minus x min divided by x max plus x min. So, to take an example, here we have the number of seeds per pod in a tree. So, we have number of seeds and the number of pods. And we are required to find out the range and the range coefficient of dispersion.

So, the number of seeds goes from 1 2 3 all the way up till 10, and the number of pods is also given as 26 113 120 So on up till 4. Now we are required to find out the range and the range coefficient of dispersion. So, here we need to keep in mind that the number of seeds is our variable x. So, we have x min equals 1 x max equals 10. So, the range would be 10 minus 1 equal to 9. So, that is the range of this distribution the range coefficient of dispersion would be 10 minus 1 divided by 10 plus 1, which is 9 by 11 using this formula.

(Refer Slide Time: 24:44)



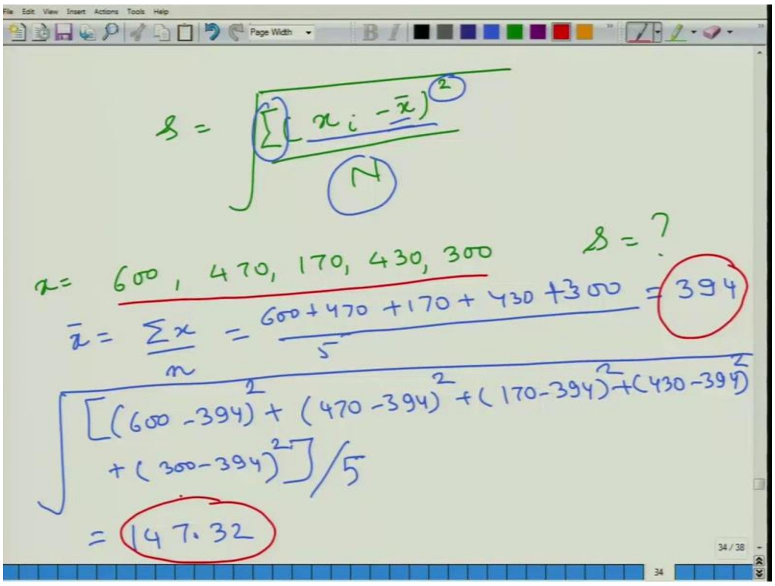
Standard deviation

- "the positive square root of the mean of the square deviations taken from arithmetic mean of the data"

$$S = \sqrt{\frac{\Sigma(X - \overline{X})^2}{N}}$$

Another measure of dispersion that is very widely used is the standard deviation, it is defined as the positive square root of the mean of the squared deviations taken from the arithmetic mean of the data.

(Refer Slide Time: 24:58)



$$s = \sqrt{\frac{\boxed{\sum}(x_i - \bar{x})^{\boxed{2}}}{\boxed{N}}}$$

$x = 600, 470, 170, 430, 300$ $\quad s = ?$

$$\bar{x} = \frac{\sum x}{n} = \frac{600 + 470 + 170 + 430 + 300}{5} = \boxed{394}$$

$$\sqrt{\left[(600-394)^2 + (470-394)^2 + (170-394)^2 + (430-394)^2 + (300-394)^2\right]/5}$$

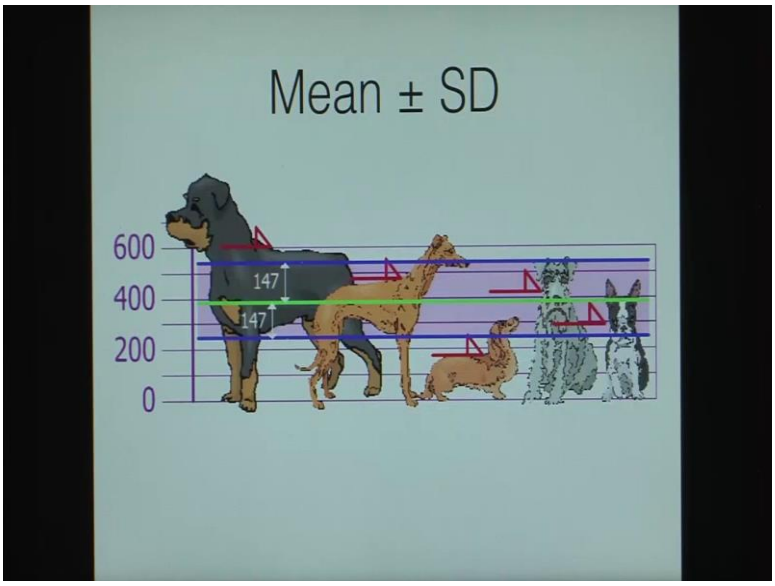
$$= \boxed{147.32}$$

That is s equals the positive square root of xi minus x bar the square of the values divided by n. So, to understand it more thoroughly let us take another example. The heights of 5 dogs in millimeter are as under. So, you have the heights they are given as 600 mm 470 mm 170 mm 430 mm and 300 millimeters.

And we are required to find the standard deviation. So, how do we go about doing it? The first step would be to find out x bar. Now x bar that is the mean is given by sum of all x divided by n. Here we have 5 values and we take 600 plus 470 plus 170 plus 430 plus 300 whole divided by 5. And this value comes to be 394. So, that is the main value that is the measure of the central tendency. Now how do we calculate this standard deviation? To calculate this standard deviation we need to subtract this value of mean from every value. So, you will have 600 minus 394 the second value would be 470 minus 394, then we have 170 minus 394, then we have 430 394. And we have 300 minus 394. So, we take xi minus x bar then we square all of these. So, that is the square. Then we add all of these. And once you have this whole value you divide it by n, n here is 5. And then you take a big square root over all these values.

So, when you calculate this you get it as 147.32. So now, we can say that this distribution has a measure of central tendency has 394 and it has a dispersion of 147.32 . Now what do we make out of these 2 values? Let us plot those. So, if you plot Mean plus minus standard deviation, we see that in this case the green line depicts the mean and we have 2 other lines that that are depicting this standard deviation of 147.32.

(Refer Slide Time: 27:48)



Mean ± SD

So, as you can see there are some dogs that have a height like this third dog has a height that is very much less than the mean, but if you consider mean minus standard deviation it is height is covered. So, mean plus minus standard deviation covers nearly all the values, which is why it becomes such an important measure to use. So, essentially if you want to depict any a distribution, you write the measure of central tendency plus minus the dispersion. So, in this case the measure of the central tendency would tell you the central value and the dispersion would tell you the spread.

Thank you for your attention.