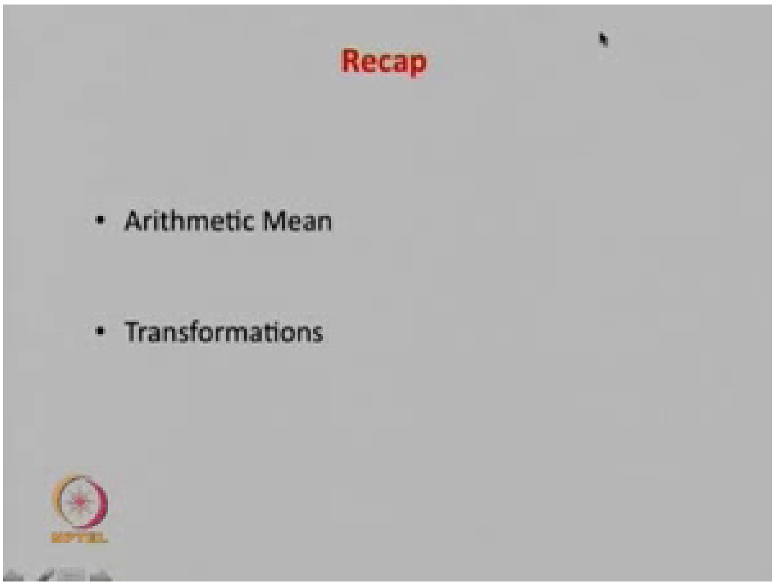
Introduction to Biostatistics
Prof. Shamik Sen
Department of Bioscience and Bioengineering
Indian Institute Technology, Bombay

Lecture – 04
Geometric mean

Dear students, welcome to today's lecture. As always would begin with a brief recap of what we had covered in last lecture and then go on from there.

(Refer Slide Time: 00:35)

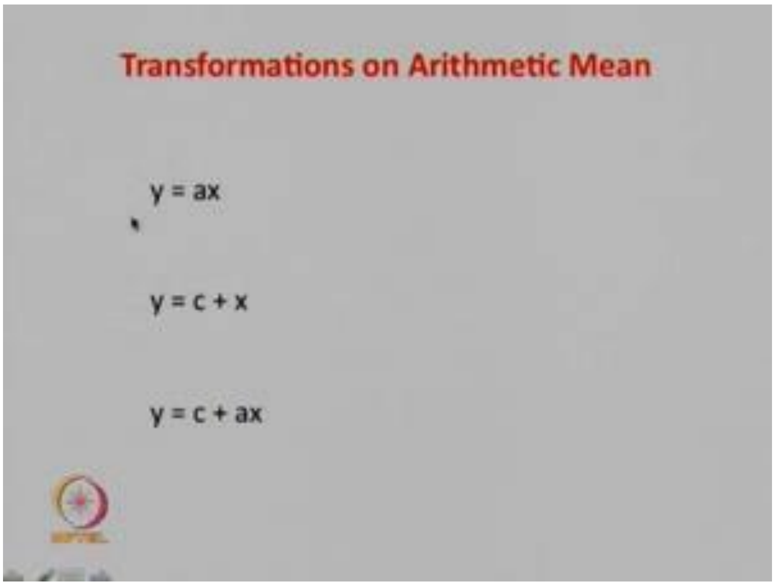


Recap

- Arithmetic Mean
- Transformations

In last lecture, we had discussed two things when. So, how to calculate arithmetic mean given a data set. So, arithmetic mean represents one of the most widely used matrix for representing your data and took and we also briefly discussed about transformations is given a data set how can you make use of transformations to make it easier for you to calculate arithmetic mean particularly if you are doing it by hand. So, there were three main transformations we had discussed in last class y is equal to a x y is equal to c plus x and y is equal to c plus a x.

(Refer Slide Time: 01:01)



Transformations on Arithmetic Mean

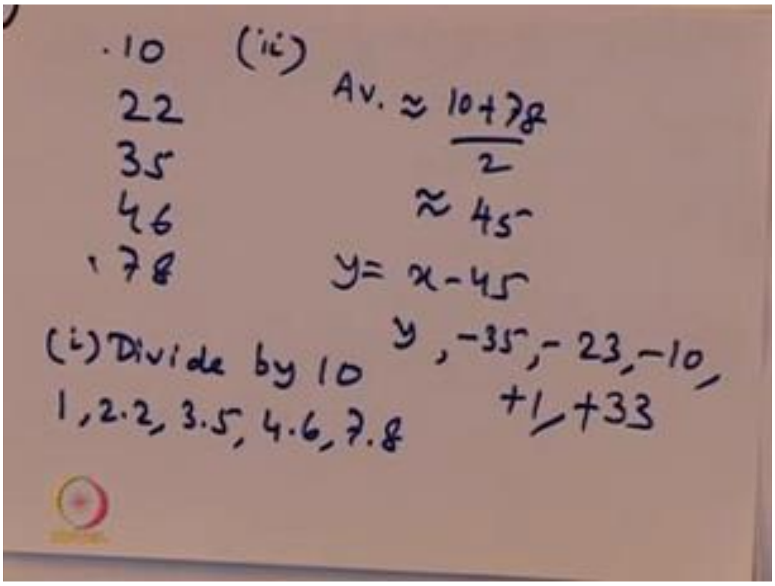
y = ax

y = c + x

y = c + ax

So, just to give you an example, if you; let us say have numbers like 10, 20, 22, 35, 46, 78. So, what should you do? So, what you see there are two things which are going on these are increasing roughly in terms of in jumps of 10 or so and you have a range from 10 to 80.

(Refer Slide Time: 01:14)



10

22

35

46

78

(ii)

$\text{Av.} \approx \frac{10+78}{2}$

≈ 45

$y = x - 45$

y, −35, −23, −10, +1, +33

(i) Divide by 10

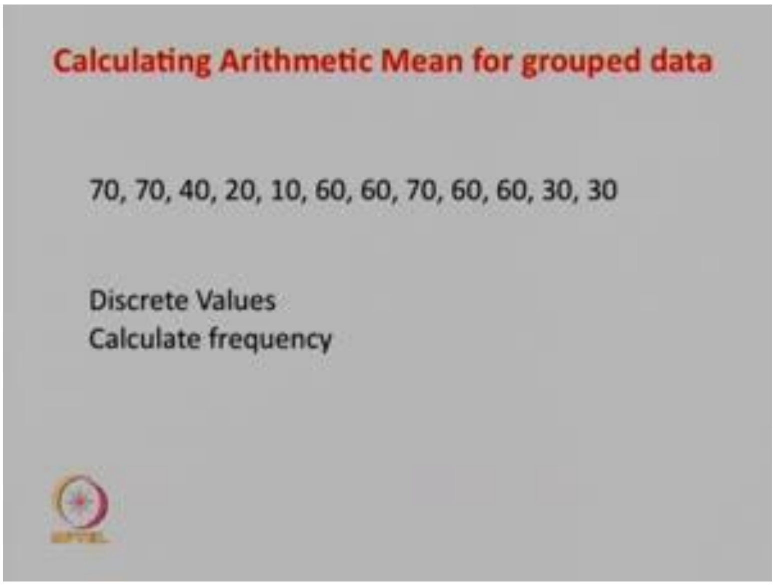
1, 2.2, 3.5, 4.6, 7.8

So, either; what we can do is, particularly if you are doing by hand and in this particular case there are only 5 numbers. So, we can even add up by hand. So, one possibility is divided by 10.

So then you have the numbers 1, 2.2, 3.5, 4.6 and 7.8; these you can add up by hand very easily the second possibility is the second possibility is like we did we identify the smallest number which is 10 the largest number which is roughly 80. So, which is 90; so, on 1 you know, so the average is roughly you know 45. So, average here is roughly equal to 10 plus 78 by 2 which is roughly equal to 45, sorry, 89-90. So, 45, so, we can apply the transformation y is equal to x minus 45 here and then in that case my numbers will be minus 35. So, y values are minus 35 minus 23 minus 10 plus 1 plus 33 then I can add them up very easily and find out what is my average.

So, what you see that there is no not one rule which will make things work depending on what kind of transformation you use? There can be more than one rule. So, that is one thing. So, today we will discuss the next thing which is let us say you have a discrete data set.

(Refer Slide Time: 02:57)



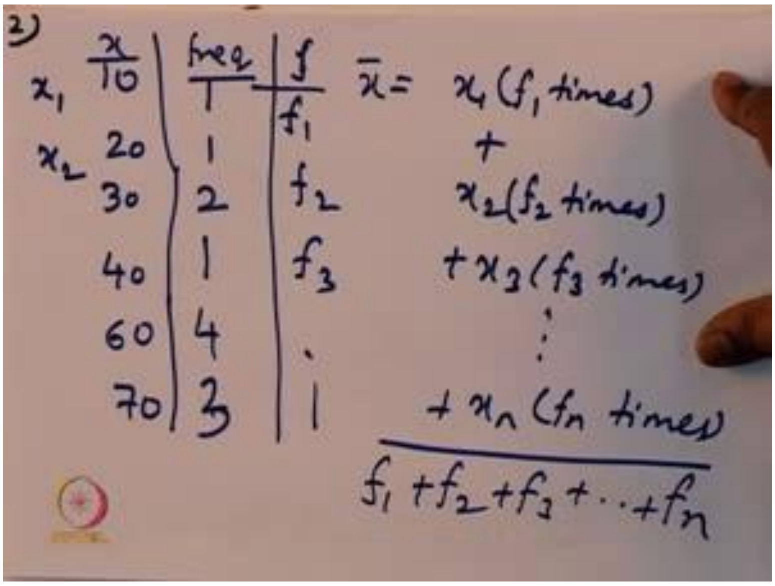
Calculating Arithmetic Mean for grouped data

70, 70, 40, 20, 10, 60, 60, 70, 60, 60, 30, 30

Discrete Values
Calculate frequency

So, if you look at the data set here. So, you have numbers which are all first of all discrete and then they are repeated.

(Refer Slide Time: 03:11)



2)

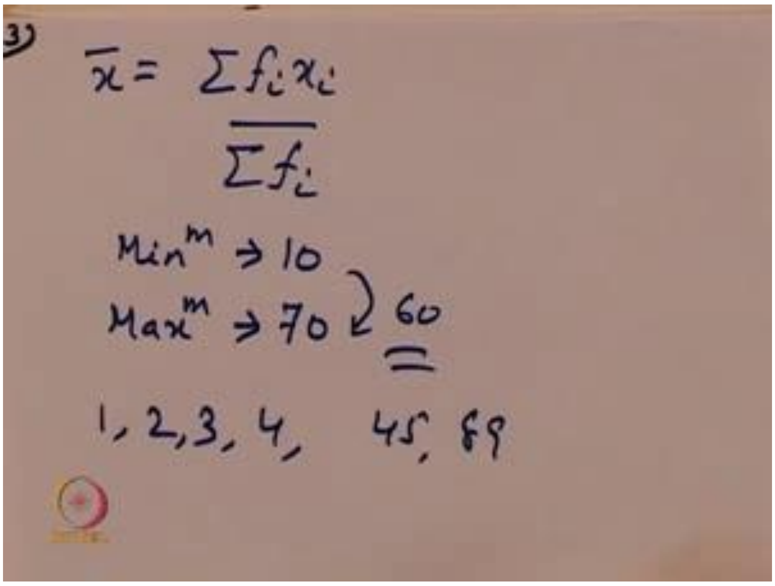
	x	freq	f
x_1	10	1	f_1
x_2	20	1	
	30	2	f_2
	40	1	f_3
	60	4	.
	70	3	|

$$\bar{x} = \frac{x_1(f_1 \text{ times}) + x_2(f_2 \text{ times}) + x_3(f_3 \text{ times}) \vdots + x_n(f_n \text{ times})}{f_1+f_2+f_3+\dots+f_n}$$

So, if this was entire data set what you can clearly see is the number 10 has been repeated. So, by frequency, these are my x values, my frequency 10 has been repeated once, 20 has been repeated once, 30 has been repeated twice, 40 has been repeated once, 60, 60 has been repeated 1, 2, 3, 4, times and 70 has been repeated 2 times; no sorry 3 times. So, what you see? These are discrete values which are repeated.

So, if I were to go back to the way we calculated our average we want to do. So, in general we can write it as frequency is f 1 f 2 f 3, so on and so forth and these guys we call as x 1, x 2. So, we can calculate my x bar is simply x 1, into f 1 times plus x 2 which has been repeated f 2 times plus x 3 which is repeated f 3 times so on and so forth plus x n which has been repeated f n times and we want to divide by the total number of observation which is nothing by f 1 plus f 2 plus f 3 plus dot, dot, dot plus f n.

(Refer Slide Time: 04:42)



3)

$$\bar{x} = \frac{\sum f_i x_i}{\sum f_i}$$

$\text{Min}^m \rightarrow 10$

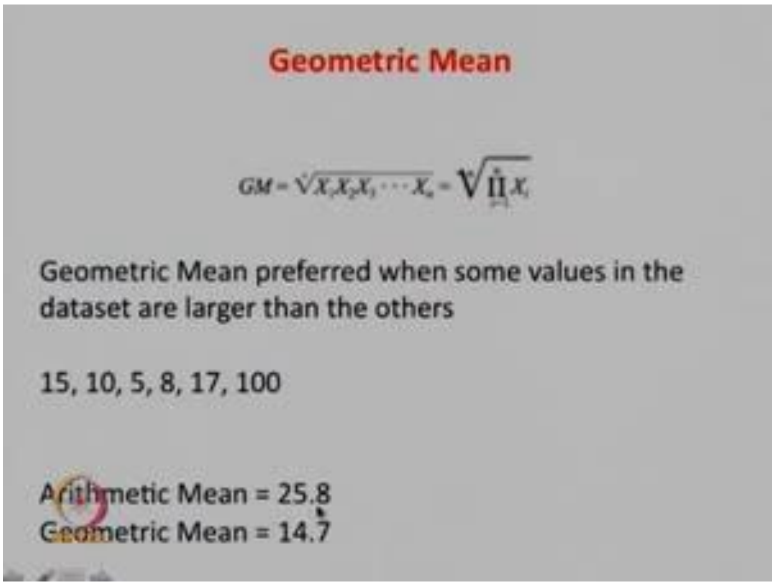
$\text{Max}^m \rightarrow 70 \Big\} \underline{\underline{60}}$

1, 2, 3, 4, 45, 89

So, we can the final expression becomes x bar is nothing, but summation f i into x i by summation f i. So, as you know as before I did not put the start and end values, but this just summation alone means that I am doing a sum from i is equal to 1 to i equal to n. So, you can do this as an exercise. So, you know, what are the values of f 1, f 2? What are the values of x 1, x 2 and please calculate the final average. So, this is what it is, you can have to basically calculate the frequency of each number is discrete variable and their value and then calculate the total sum as per this particular equation.

Now, let us I wanted to discuss one more thing. So, even in this case, let us say these examples that we took the variables are 10, 20, 30 so on and so forth. So, you have a range. So, your minimum value is 10 your maximum value is around 70, fine you have a range. So, you can see that the jump is roughly by 60, imagine I have data which is 1, 2, 3, 4 and then 45, 89 so on and so forth. So, what I can clearly see is if I take a mean arithmetic mean I will get a value which is going to be much greater than these 2 and will be much more biased toward these bigger numbers. So, this is one of the principal weaknesses of calculating arithmetic mean if you have large variation in your data set then arithmetic mean might give you a value which really does not mean anything. So, what might be an alternative in our approach to come up with a better number?

(Refer Slide Time: 06:32)



Geometric Mean

$$GM = \sqrt[n]{X_1 X_2 X_3 \cdots X_n} = \sqrt[n]{\prod_{i=1}^{n} X_i}$$

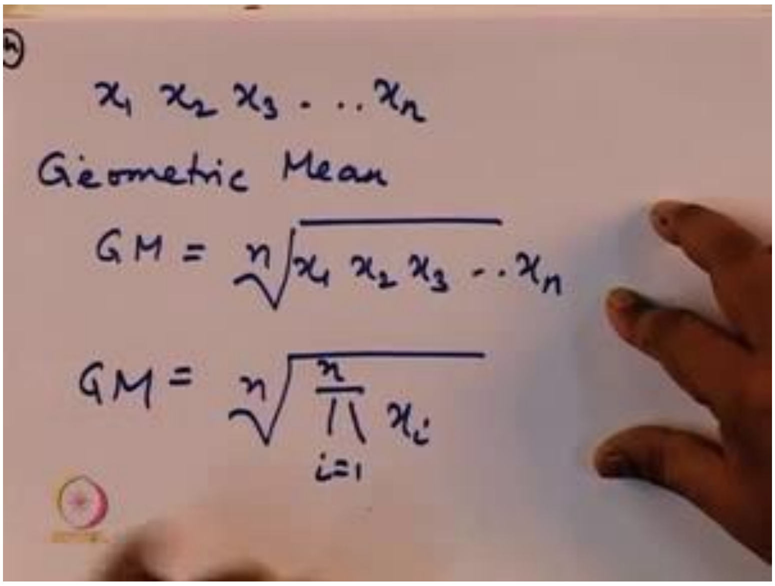
Geometric Mean preferred when some values in the dataset are larger than the others

15, 10, 5, 8, 17, 100

Arithmetic Mean = 25.8
Geometric Mean = 14.7

So, one number which is often used or you know which is used to calculate this is called the geometric mean. So, as opposed to adding up the numbers as opposed to adding up the numbers you take their product and you take their root nth root.

(Refer Slide Time: 06:45)



$x_1 \; x_2 \; x_3 \; . \; . \; . \; x_n$

Geometric Mean

$$GM = \sqrt[n]{x_1 \, x_2 \, x_3 \, . \, . \, x_n}$$

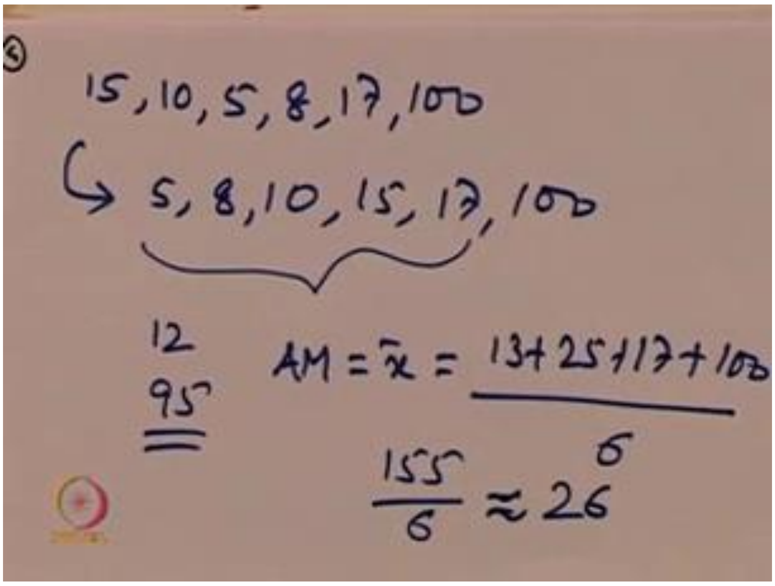
$$GM = \sqrt[n]{\prod_{i=1}^{n} x_i}$$

So, if we have n numbers x 1, x 2, x 3, dot, dot, dot, x n, the way we root geometric mean and in short we write GM is equal to nth root x 1 into x 2 into x 3, dot, dot, x n. So, the way to write it in short mathematically is to write nth root and you put this thing called pi; pi of x i and i is equal to 1 to n here. So, this means product. So, if i is equal to

1 to n equal to 1 then it will simply be 1. If i n is 5 then you have x 1, x 2, x 3, x 4, is 5 so on and so forth. So, this is a simple representation of geometric mean.

Now, let us take an example and see what this you know what this rule does. So, imagine you have a data set as follows which is 15, 10, 5, 8, 17, and 100.

(Refer Slide Time: 07:49)



③

$15, 10, 5, 8, 17, 100$

$\hookrightarrow \underbrace{5, 8, 10, 15, 17}, 100$

$\underline{\underline{\begin{matrix}12\\95\end{matrix}}}$

$AM = \tilde{x} = \dfrac{13+25+17+100}{6}$

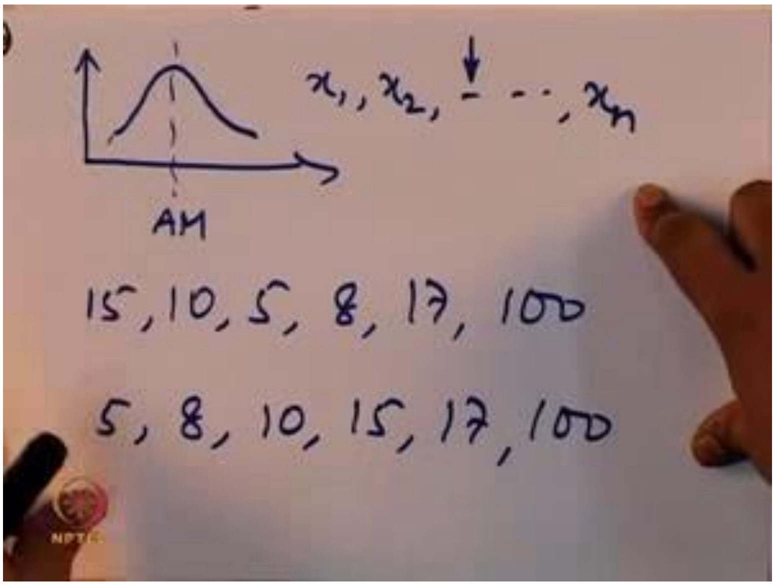
$\dfrac{155}{6} \approx 26$

So, my numbers are 15, 10, 5, 8, 17, and 100. So, if I were to you know sort these numbers and put them next to each other in terms of ascending order. So, I have 5, 8, 10, 15, 17, 100 so, but just by looking at the data I can clearly see that the number 100 is really an outlier in other words outlier means maybe this data was taken when there is some error in the procedure of acquisition of data some experimental error because I can see that my number is you know the other numbers are much closer to each other. So, here my 5 to 17 range is 12 only. So, 17 minus 5 is 12 versus 100 minus 5 is 95 now how can I; so if I account for 100 in calculating the average what number do I get? So, my arithmetic mean x bar becomes 13 plus 25 plus 17 plus 100 c, 4 c of 6 numbers. So, 30, 55, 155; 155 by 6 is equal to 2s square 155. So, if I would to roughly write, it is close to 26, this close to 26, the exact value is 25.8.

Now, so, clearly I can see that 25.8 would be a value which is right here which is way further than most of the other numbers. So, really this arithmetic mean gives me a number which does which is not representative of the whole population how can I come up with an alternative number. So, geometric mean is one such approach to get a number which is much more representative of your data. So, if I do the same exercise with the geometric mean. So, what I get is a value of 14 point 7 which is still much. So, we have a value which is you know which is between 10 and 15, which is still more representative of the population. So, you have changed from 25.8 to 14.7.

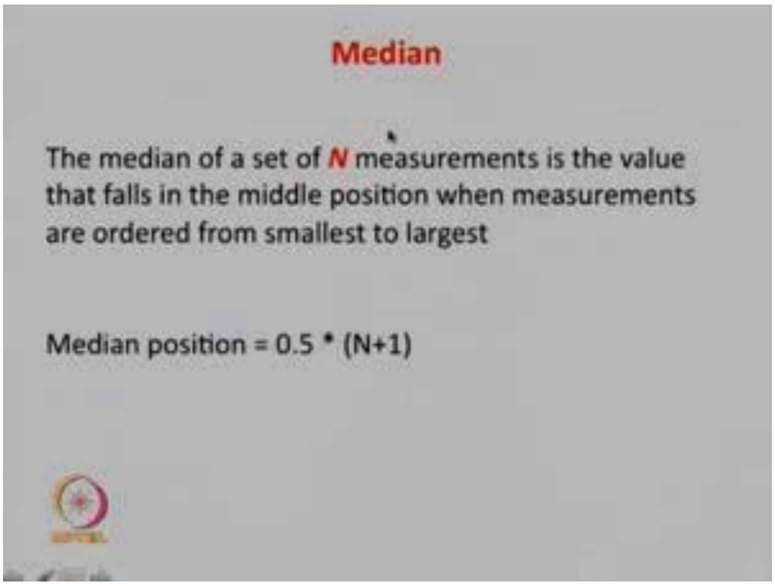
So, this is one of the big advantages of geometric mean over arithmetic mean particularly when your data set in your data set there is huge heterogeneity and there are 1 or 2 values which are way larger then what is the most of the other values.

(Refer Slide Time: 10:22)



So is one more important measure. So, let us say that you draw a histogram and you plot the data and your plot data looks like this. So, logic would dictate that if you do the arithmetic mean of all these population where arithmetic mean will somewhere lie here this is where your arithmetic mean lies now there is one more way of doing it. So, what you can do is you can sort these numbers entire numbers x 1, x 2, up to x n in ascending or descending order and try to find; which is the number; which is directly in the middle. So, it is some. It is a; it would be a very nice alternative to the arithmetic mean and this particular metric is an has a name it is called the median.

(Refer Slide Time: 11:15)



Median

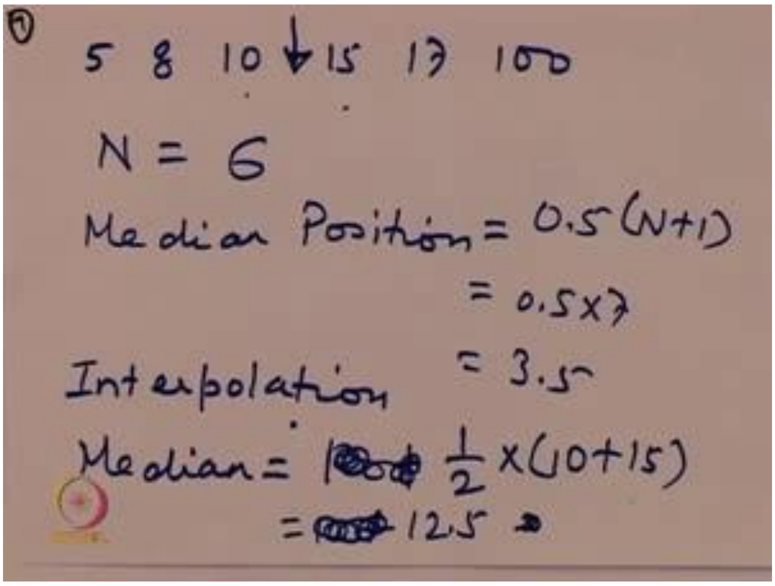
The median of a set of *N* measurements is the value that falls in the middle position when measurements are ordered from smallest to largest

Median position = 0.5 * (N+1)

So, this is how you define the median the median of a set of n measurements is the value that falls in the middle position when measurements are ordered from smallest to largest. So, of course, whether you order from smallest to largest or largest or smallest that does not matter, but. So, you basically collect find the number which is at this position which is 0.5 times n plus 1.

So let us do the previous numbers you had if I go back to the previous slide and my numbers were 15, 10, 5, 8, 17, 100, 15, 10, 5, 8, 17 and 100, so if I sort them and I write it together 5, 8, 10, 15, 17, 100.

(Refer Slide Time: 12:08)



5 8 10 ↓ 15 17 100

$N = 6$

Median Position $= 0.5(N+1)$

$= 0.5 \times 7$

Interpolation $= 3.5$

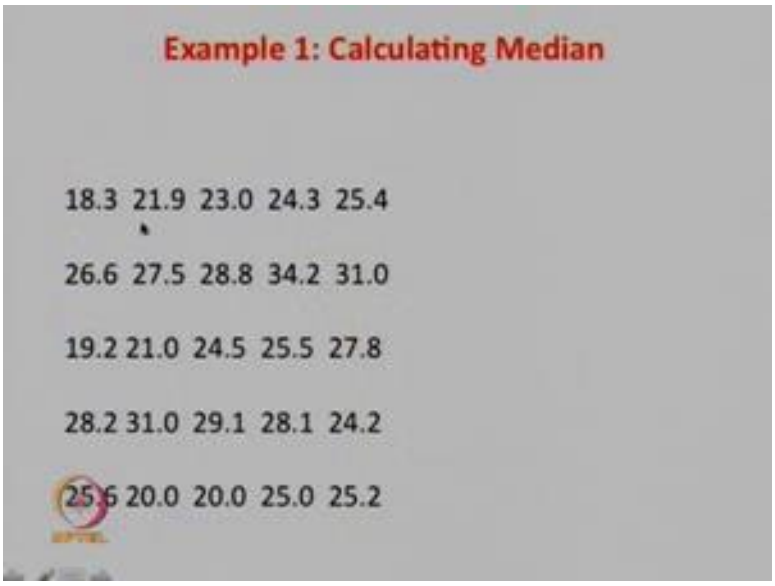
Median $= \frac{1}{2} \times (10+15)$

$= 12.5$

So, these are this is the sorted list. So, I put them together. So, I have these numbers 5, 8, 10, 15, 17, 100, what is my n? N is the total number of numbers I have which is 1, 2, 3, 4, 5, and 6. So, my median position is what is equal to half into n plus 1. So, 0.5 times n plus 1 is equal to 0.5 into 7 is equal to 3.5. So, 3.5 means what the position is somewhere right in the middle of 10 and 15, this is the third number, this is the fourth number, but my median is the number which is right at position 3.5.

So, how do I find the number which is 3.5? So, I use what is called as interpolation. So, my median as per this will be the average. So, will be 10 plus half into 10 plus 15, this is the position it is middle of this. So, we will be get 10 plus half into 10 plus 15 is 12.5 is equal to so and then its position is half, sorry, sorry, it is simply at this position, it will is going to be half into 10 plus 5; 15 is 12.5. So, your median is 12.5. So, you get a number which is even better it is; so, it is much more you know representative of this entire data set the effect of 100 has completely been eliminated. So, and the other advantage is median is of course, much more easier to calculate than geometric mean. So, let us solve another problem of calculating the medium.

(Refer Slide Time: 14:08)



Example 1: Calculating Median

18.3 21.9 23.0 24.3 25.4

26.6 27.5 28.8 34.2 31.0

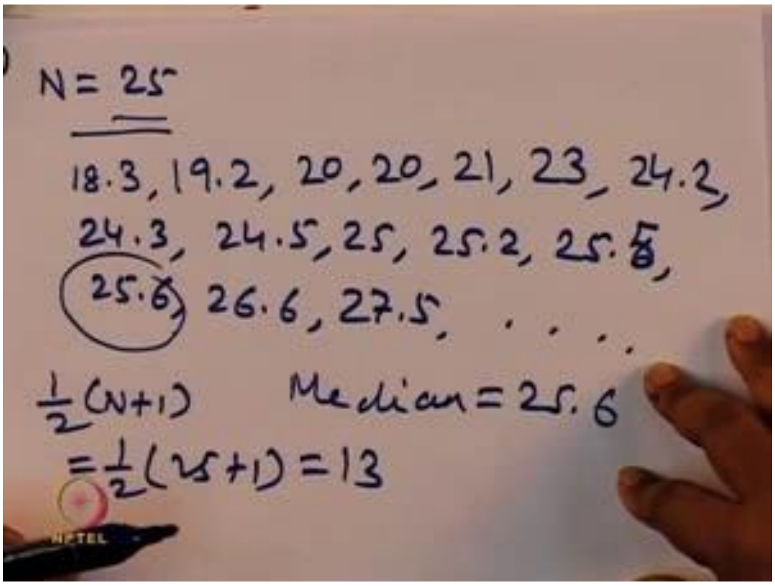
19.2 21.0 24.5 25.5 27.8

28.2 31.0 29.1 28.1 24.2

25.6 20.0 20.0 25.0 25.2

So let us take this following data set again. So, what I can have? So, let me first count. So, I want to calculate the median, I want to number them and I want to see what is the pose. So, first of all how many numbers way a 1, 2, 3, 4, 5, 5 into 5 is 25.

(Refer Slide Time: 14:22)



$N = 25$

18.3, 19.2, 20, 20, 21, 23, 24.2,
24.3, 24.5, 25, 25.2, 25.5,
25.6, 26.6, 27.5, ...

$\frac{1}{2}(N+1)$ Median = 25.6

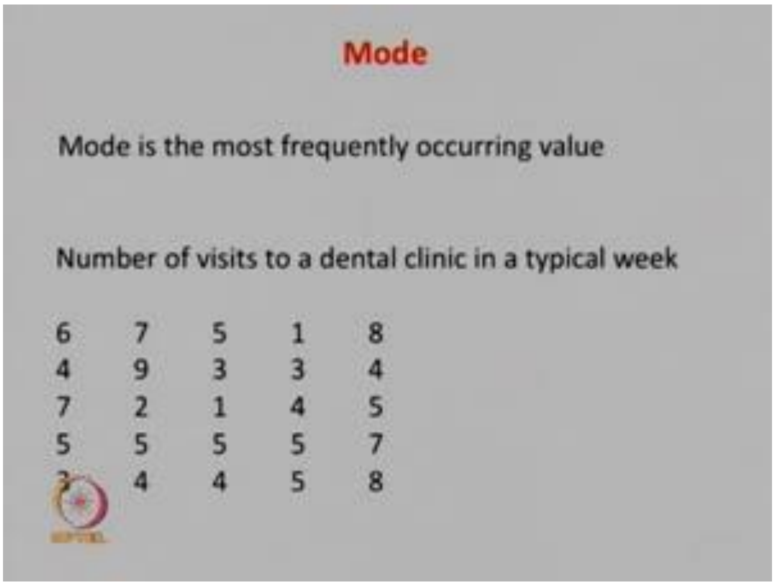
$= \frac{1}{2}(25+1) = 13$

So, n is equal to 25, I want to order their numbers together. So, I have to sort them in terms of you know lowest to highest number. So, my lowest number is 18.3, the next highest number is 19.2, the next highest number is 19.2, you have 20, 20, the next highest number is 20, I have one 21 here, after 21, I have 23 here. So, 23, after 23 is 24.3

is there here and 24.2 is here. So, 24.2; 24.3, 24.2 to 24.5 then after 24.5, you have 25, you have 25; 25.2, 25.6. So, it is 25.5 also then 25.6, 25.5 is there, 25.6 after that you have 26.6 then after 26.6, you have 27.5. So, I can write the remaining numbers now. I know that n is 25. So, my median position is basically, half into n plus 1 position is equal to half into 25 plus 1 is equal to 13. So, I did not complete it, but whatever number 1, 1, 2, 3, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13. So, your median is 25.6. So, median is 25.6.

So sometimes you will get an exact value. So, if your number of data is odd. So, when n is odd then half of n plus 1 will give you an exact value? So, you directly choose this value, but when n is even half of n plus 1 will give you a fraction. So, you have to interpolate between 2 numbers as was the case previously. So, in this case I had 6 observations. So, my median position was 3.5 in which case I had to interpolate between the 2 numbers, but when n is equal to odd. So, I can get the exact value because this is an exact position. So, I can just collect the number at that position.

(Refer Slide Time: 17:31)



Mode

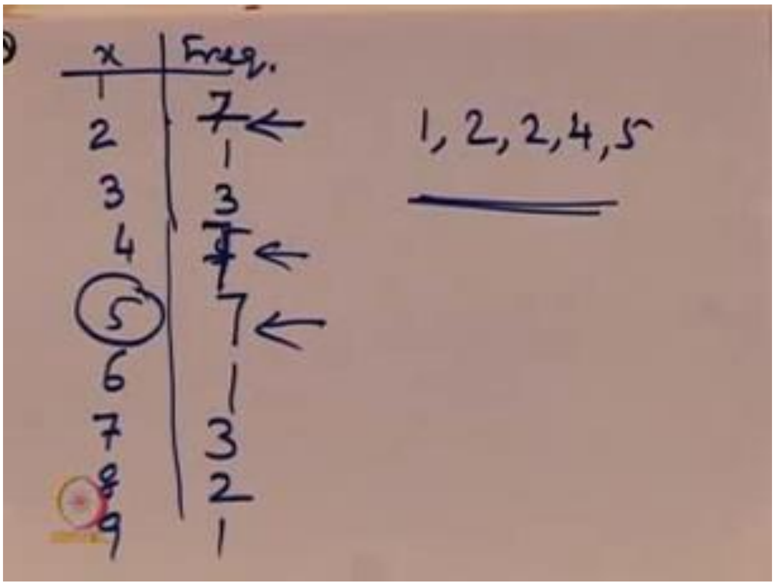
Mode is the most frequently occurring value

Number of visits to a dental clinic in a typical week

6	7	5	1	8
4	9	3	3	4
7	2	1	4	5
5	5	5	5	7
3	4	4	5	8

So, there is one more metric, there is one more metric which is often used this is called the more mode the most frequently occurring value. So, in our case, let us say for example, this is a data set of number of visits to you know n dental clinic in a typical week and we can see that which is the most frequently observing value.

(Refer Slide Time: 17:50)



x	Freq.
1	7
2	1
3	3
4	7
5	7
6	1
7	3
8	2
9	1

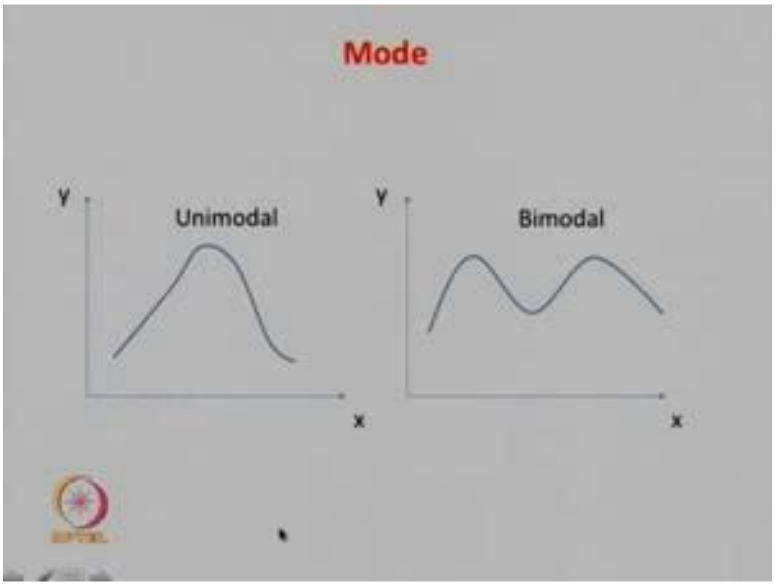
1, 2, 2, 4, 5

So, let us tabulate again, I have my x and I have my frequency. So, number of visits 1, I have 2, I have 3, I have 4, I have 5, I have 6, I have 7, I have 8, I have 9.

So, frequency for 1 is 2, frequency for 2 is 1, frequency for 3 is 1, 2, 3, frequency for 4 is 1, 2, 3, 4, 5. So, frequency for 5 is 1, 2, 3, 4, 5, 1, 2, 3, 4, 5, 6, 7 frequency for 6 is 1, frequency for 7 is 1, 2, 3, frequency for 8 is 1, 2, frequency for 9 is 1. So, clearly you look at the frequency column identify the maximum number which is 7; that means, your mode is 5. So, mode is the most frequently occurring value which is 5, so on an average 5 number of visits are there in a given week.

But let us say in this; you did this; you went through this exercise and let us say hypothetically; let us say you have another case where even 4 gives you the number 7 or you know one also gives you the number 7 then what do you do?

(Refer Slide Time: 19:27)

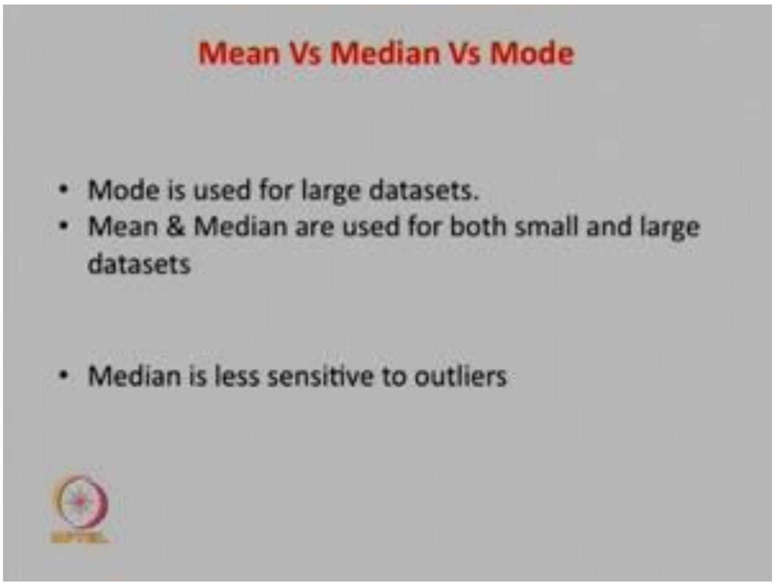


Mode

So, you do not have one unique answer and that is exactly what brings us to this plot where you have you know your distribution might be Unimodal; that means, there is a heap which is with a unique peak or you can have bimodal and there are 2 peaks in this curve. So, this is a bimodal distribution in which case you have to report both the mode values what is the Unimodal distribution.

So, now that we have 3 different matrix of quantifying the data one is mean one is median one is mode. So, which one do we choose? Which one do we choose?

(Refer Slide Time: 20:00)



Mean Vs Median Vs Mode

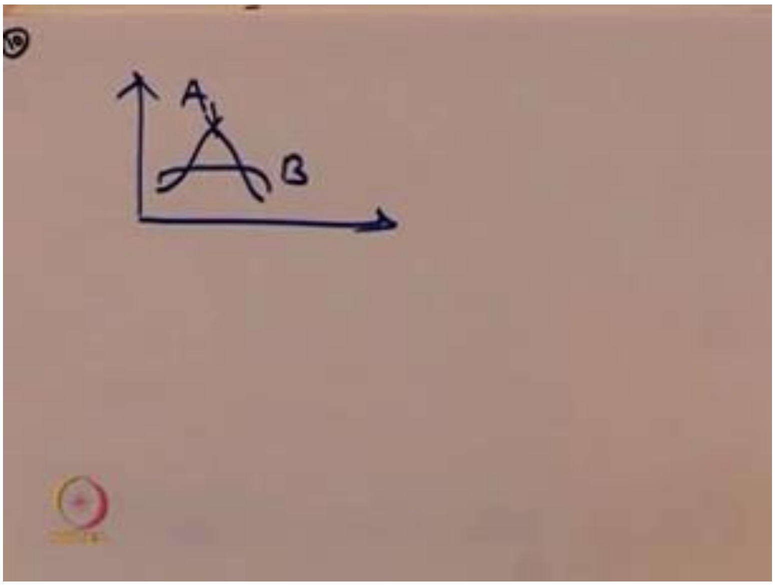
- Mode is used for large datasets.
- Mean & Median are used for both small and large datasets
- Median is less sensitive to outliers

So, a rule of thumb is mode is used for large data sets let us just say for hypothetical case you had only 5 observations and these operation 1, 2, 3, 4, 5. So, in this case reporting a mode does not make sense because your data set is too small. So, in this case it is better to do either a mean or a median. So, as the rule of thumb if you look back at the rule of thumb mode is used for large data sets what is large if you have 50 observations I would say that is a reasonable number to calculate or to report mode as one of the representative matrix.

The other 2 matrix mean and median are used for both small and large data sets. So, as we said that you know one of the caveats of mode is depending on your distribution you may not have one unique value, but there might be more than one, but both mean and median gives you one particular value the other beauty about median is median is insensitive to outliers advert seen in the previous case where we could clearly see that you know if you when we are doing the arithmetic mean this guy 100 will directly influence your observations, but when you are doing the median then essentially you are finding out the position which is here and then that eliminates the possibility or effect of outliers.

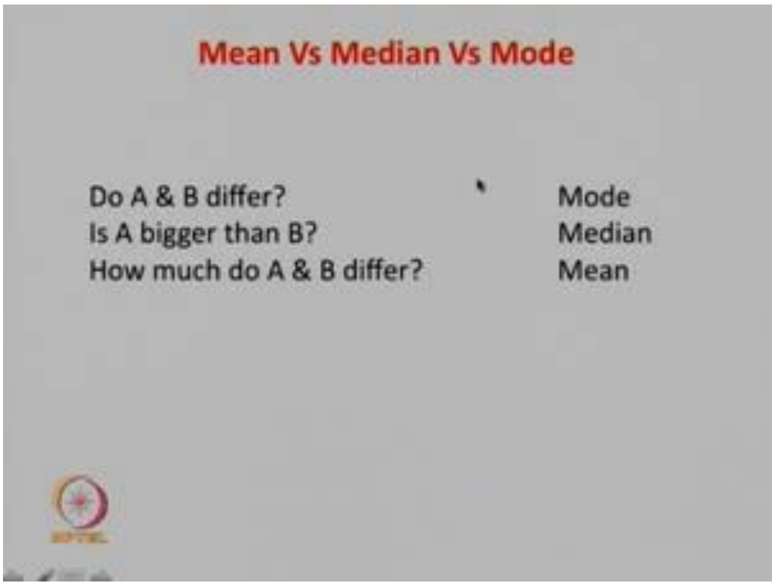
So, this is one of the you know beauty of this median metric that median is not is less sensitive to outliers, but you have multiple values which are all big and they are spread across then median might have a outlier there is another way of thinking about when to use mean when to use median and mean to use mode.

(Refer Slide Time: 22:08)



So let us say you did experiments where there were 2 conditions are 2, you know 2 experimental conditions A and B and that you have obtained your data set and you want to ask do A and B differ in other words if I; so, let us say you have done your frequency histogram this is for condition A and for condition B, it is this is for condition B. So, from a peaked distribution which has A distinct maxima, B is reasonably flat in other words all values are equally reasonable. So, clearly this would tend to say that there is a big difference. So, and you can report mode as this value, but in the example that I took there is no mode to report.

(Refer Slide Time: 22:35)

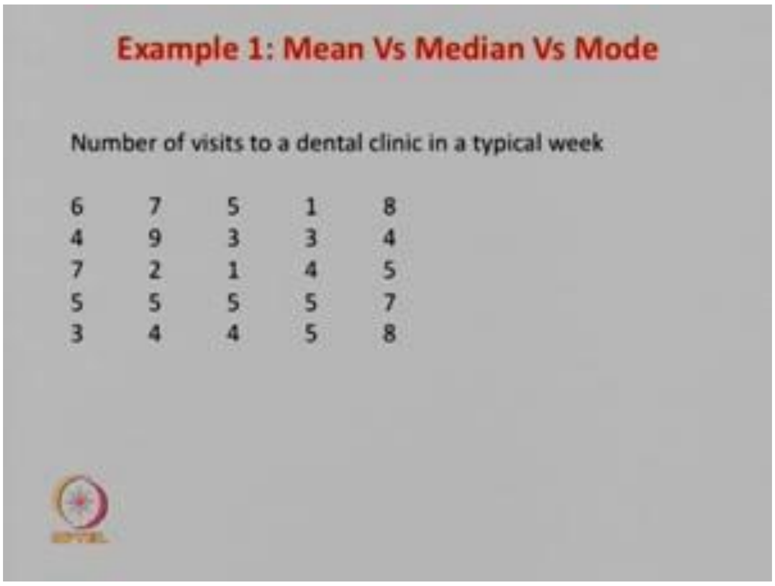


Mean Vs Median Vs Mode

Do A & B differ?	Mode
Is A bigger than B?	Median
How much do A & B differ?	Mean

Second thing is A bigger than B, in that case median might be a good way to look at it and last how much you want to quantify the extent to which A and B differ and in that case we can report the mean. So, there are different you know context in which you might use a mean median or a mode.

(Refer Slide Time: 23:06)



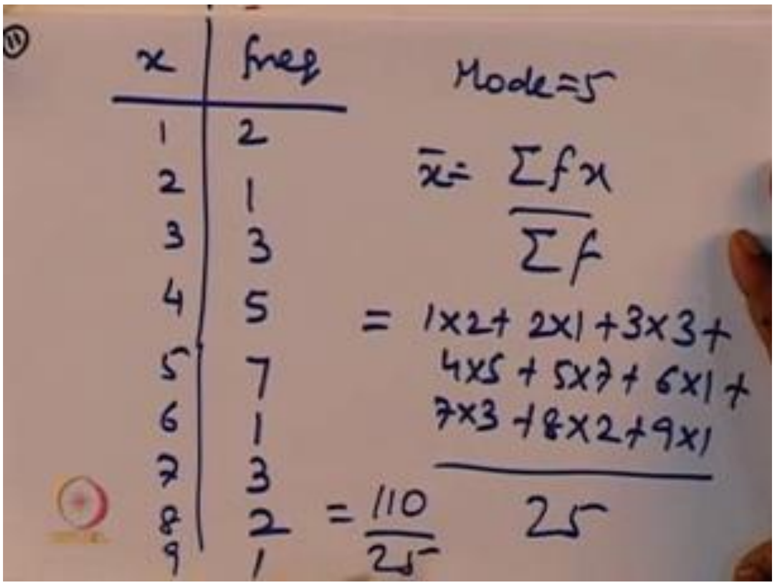
Example 1: Mean Vs Median Vs Mode

Number of visits to a dental clinic in a typical week

6	7	5	1	8
4	9	3	3	4
7	2	1	4	5
5	5	5	5	7
3	4	4	5	8

Let us go through few examples where we compute all these 3 quantities. So, let us again come to this particular example the number of visits to a dental clinic in a typical week. So, what we wanted to; so, I had calculated that frequency distribution. So, this was the frequency distribution earlier. So, my mode sorry; I let me rewrite it let me rewrite it. So, I have x and I have frequency. So, you have where is a 1, 2, 3, 4, 5, 6, 7, 8, and 9.

(Refer Slide Time: 23:31)



11

x	freq
1	2
2	1
3	3
4	5
5	7
6	1
7	3
8	2
9	1

Mode = 5

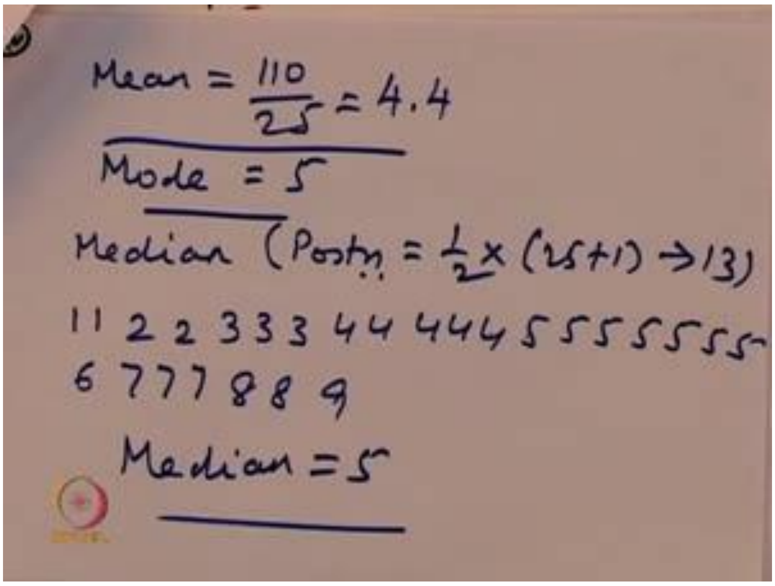
$$\bar{x} = \frac{\Sigma fx}{\Sigma f}$$

$$= \frac{1\times2 + 2\times1 + 3\times3 + 4\times5 + 5\times7 + 6\times1 + 7\times3 + 8\times2 + 9\times1}{25}$$

$$= \frac{110}{25}$$

So 1 is 2 times; 1 is 2 times, 2 is 1 time, 3 is 3 times, 4 is 5 times, 5 is 7 times, 6 is 1 times, 7 is 3 times, 8 is 2 times, 9 is 1 time. So, for this particular distribution we want to identify we want to identify the mode which we already have. So, my mode is equal to 5 let us calculate the mean then. So, mean; I can use the formula x bar is equal to summation f x by summation f this would give the value 1 into 2 plus 2 into 1 plus 3 into 3 plus 4 into 5 plus 5 into 7 plus 6 into 1 plus 7 into 3 plus 8 into 2 plus 9 into 1. My total number of observations is 2 plus 1; 3, 3 plus 3; 6, 6 plus 5; 11 11 plus 7; 18 plus 1; 19, 19 plus 3; 22, 22, 24, 25. So, let us calculate the sum here. So, you here we have 2 plus 2; 4, 4 plus 9; 13, 13 plus 20; 33, 33 plus 35; 68, 68 plus 6; 74, 74 plus 21 is 95, 95 plus 6 is 101; 101 plus 9 is 110. So, it is 110 by 25. So, we just go to the next page.

(Refer Slide Time: 25:27)



Mean = $\frac{110}{25} = 4.4$

Mode = 5

Median (Postn = $\frac{1}{2} \times (25+1) \rightarrow 13$)

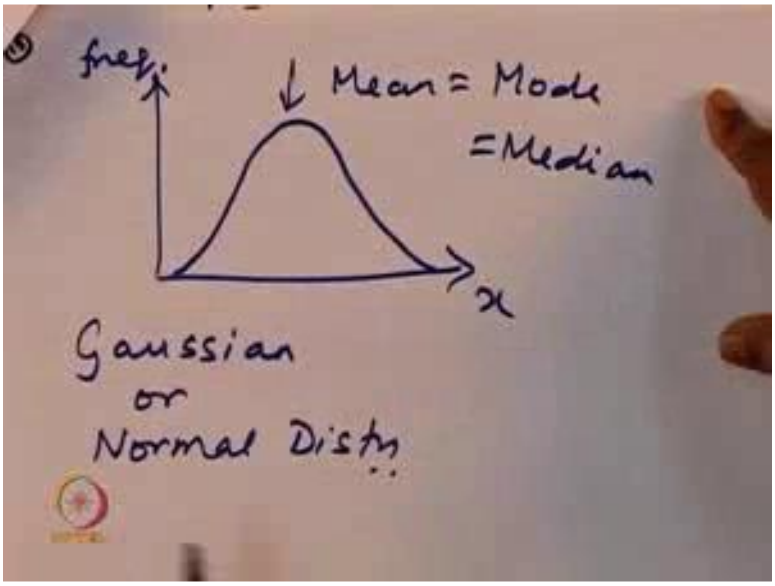
1 1 2 2 3 3 3 4 4 4 4 4 5 5 5 5 5 5 5

6 7 7 7 8 8 9

Median = 5

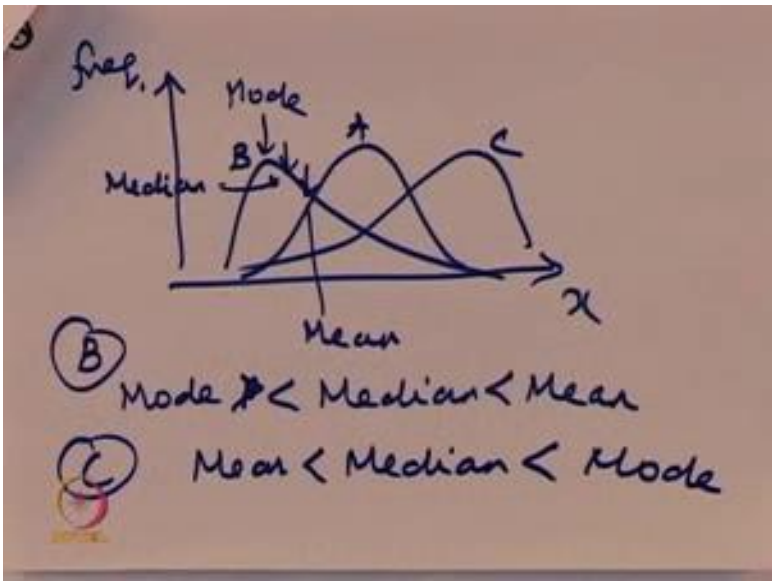
So, we had found mean is equal to 110 by 25; 4 plus 4.4 mode, we have found out as 5 and median. So, if we arrange the numbers. So, for calculating median we have to arrange the numbers we have 1, 1, 2, 2, 3, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 1, 2, 3, 4, 5, 6, 7, 6, 7, 7, 7, 8, 8, 9. So, I want to find out. So, median is at position half into 25 plus 1 which is position 13 and what is position 13? 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, we get a median is equal to 5. So, these are your 3 observations you see in this particular case mean more median they are all very close to each other, but still there is a difference.

(Refer Slide Time: 26:50)



So, in the general context if I were to draw the curves, the best case situation we will discuss this particular curve which is a Gaussian or you know a Gaussian distribution a Gaussian or a normal distribution for this particular case mean equal to mode equal to median, but for all other cases for all other cases in general. So, this is one particular case, but your data might look like this or data might look like this. So, this is x and this is frequency.

(Refer Slide Time: 27:24)



B Mode ~~>~~< Median < Mean

C Mean < Median < Mode

So, in this case for example, let us say this is my mode my median is right in the middle median will probably be somewhere here. So, this is my median an average will somewhere like here this is my mean. So, in this case I have mode greater sorry less than median less than mean and in the other case. So, this is case B, this is case A where all the 3 is equal. So, this is case B and for case C, you will have mean less than median less than mode.

So with that you get an idea that you can have 3 matrix mean median mode and then depending on your data sets. So, you have you know how to calculate mean median mode and you see that depending on the data set either there will be a left shift of the data as in case B here or there will be a right sheet of the data as in case C here or in the best case situation. If you collect enough statistics most processes in nature follow the normal distribution. So, you will have this distribution in which case whether you do the mean median mode you will get the last give the same value.

With that I thank you for today’s lecture. We will meet again in next week.

Thank you.