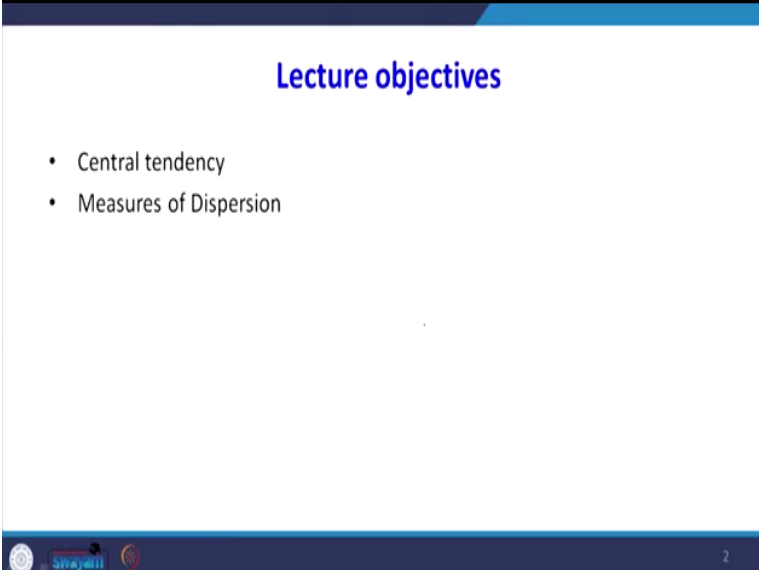**Data Analytics with Python**
Prof. Ramesh Anbanandam
Department of management studies
Indian Institute of Technology, Roorkee

Lecture No 4
Central Tendency and Dispersion

Good morning students, today is we are going to the lecture 4. In this lecture we are going to talk about central tendency and how to measure the dispersions. The lecture objectives we talk about different types of central tendency.
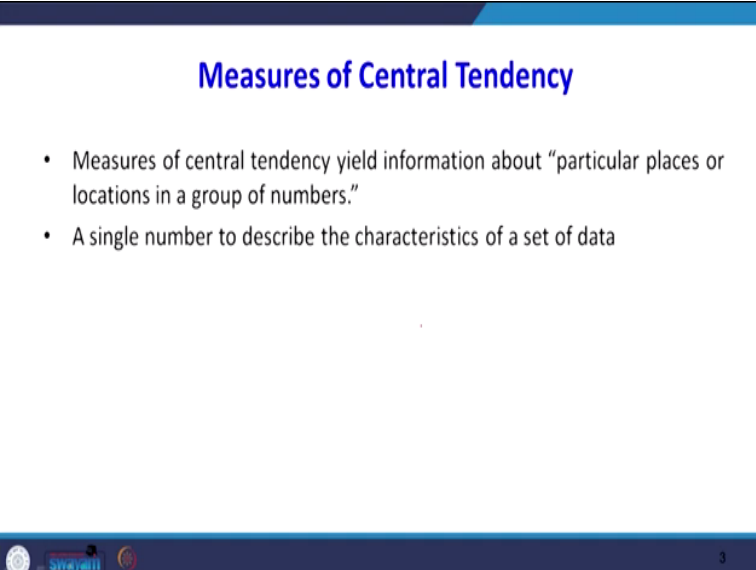
(Refer Slide Time: 00:42)

Then different types of dispersions

(Refer Slide Time: 00:44)

What is measure of central tendency? Measure of central tendency yield information about particular places or locations in a group of numbers. Suppose there are a group of number is there that number group of numbers has to be replaced by a single number that single number we can call it as central tendency. That is a single number to describe the characteristics of a set of data.

(Refer Slide Time: 01:08)

Some of the central tendency which we are going to see in this lecture is arithmetic mean, weighted mean, median, and percentile. In the dispersions we are going to talk about skewness, kurtosis, range, interquartile range, variance, standard score and coefficient of variation.

(Refer Slide Time: 01:25)

First we will see the first central tendency arithmetic mean. Commonly it is called as the mean it is the average of a group of numbers; it is applicable for interval and ratio data. This point is very important it is not applicable for nominal and ordinal data. It is affected by each value in the data set including extreme values, one of the problem of the with the mean is that it is affected by the extreme values computed by summing all values in the data set and dividing the sum by the number of values in the data set.

(Refer Slide Time: 02:01)

Population Mean

$$\mu = \frac{\sum X}{N} = \frac{X_1 + X_2 + X_3 + \ldots + X_N}{N}$$

$$= \frac{24 + 13 + 19 + 26 + 11}{5}$$

$$= \frac{93}{5}$$

$$= 18.6$$

See here I have used a notation mue, mue means capital letters mue represents, mean for the population. The formula is Sigma X / N that is, X 1 + X 2 + X 3 and so on +XN / N here N is the number of elements. For example; the values are 24, 13, 19, 26, 11 add these numbers and divided by 5 because there are 5 elements. So 93 / 5, 18.6 is the mean of these 5 numbers. So now the 18.6 can be replaced by these set 5 numbers. Okay?

Suppose in your class if you see the average mark is 60. So the whole marks of all the students can be represented to be a single number that is 60, 60 will give an idea about the performance of the whole class.

(Refer Slide Time: 02:55)

$$\bar{X} = \frac{\sum X}{n} = \frac{X_1 + X_2 + X_3 + \ldots + X_n}{n}$$

$$= \frac{57 + 86 + 42 + 38 + 90 + 66}{6}$$

$$= \frac{379}{6}$$

$$= 63.167$$

Next what is the sample mean? Make sure that the difference here the X bar. Previously for the population mean we have used mue for the sample mean we are using X bar. X bar is Sigma X /n X 1 + X 2 + X3 / n. For example; 6 element is there 57, 86, 42, 38, 19, 66 so divided by 6, the mean is 63.167

(Refer Slide Time: 03:19)

- Weighted average of class midpoints
- Class frequencies are the weights

$$\mu = \frac{\sum fM}{\sum f}$$

$$= \frac{\sum fM}{N}$$

$$= \frac{f_1 M_1 + f_2 M_2 + f_3 M_3 + \cdots + f_i M_i}{f_1 + f_2 + f_3 + \cdots + f_i}$$

Now how to find out the mean of a grouped data? The mean of your grouped data is nothing but weighted average of class midpoints, class frequencies are the weight. For the formula is mue is Sigma f into m divided by Sigma of f. Here f represents the frequency, m represents the midpoint. So Sigma f is nothing but f1 m1 + f2 m2 + f3 m3 and so on + fi mi / sum of all f. That is nothing but your N. We will see you an example;

(Refer Slide Time: 03:58)

Calculation of Grouped Mean

Class Interval	Frequency(f)	Class Midpoint(M)	fM
20-under 30	6	25	150
30-under 40	18	35	630
40-under 50	11	45	495
50-under 60	11	55	605
60-under 70	3	65	195
70-under 80	1	75	75
	50		2150

$$\mu = \frac{\sum fM}{\sum f} = \frac{2150}{50} = 43.0$$

See this is the grouped data. What is given class interval is given frequency is given class midpoint is given and multiplied value of frequency and midpoint also we can find out. For example; see here 20 to 30 there are 6 numbers is their frequency 6. Suppose if you say the marks of here if you say this this is an example of here marks obtained by in your class. So between 20 and 30 there are 6 students is there between 30 under 40 there are 18 students is there.

Suppose for this the data is in this format that is in grouped format how to find out the mean? Okay? First what do you have to do first you have to find out the class midpoint. See 20 to 30 that is a class interval the midpoint is 25, for 30 and 40. The class midpoint is the middle value 35 like this 45, 55, 65, and 75. Next one you have two multiplied by frequency and class midpoint so 6 into 25 is 150, 18 into 35 is 630, 11 into 45 is 495 and so on.

What the formula says it is a sigma of fm / sigma f last column the sum value is 2150, 2150/50 Sigma f is some of the frequency so for this kind of grouped data the mean is 43.

(Refer Slide Time: 05:29)

Weighted Average

- Sometimes we wish to average numbers, but we want to assign more importance, or weight, to some of the numbers.

- The average you need is the *weighted average.*

Now we will go to the next central tendency that is the weighted average. Some time if you look at the previous values the each value is given equal weightage. Suppose it is not always the case there may be some marks there some values where there may be higher weightage. So for that case we have to go for weighted average. Some time you see this we will list two average numbers but we want to assign more importance or weight to some of the numbers. The average you need is the weighted average.

(Refer Slide Time: 05:58)

Formula for Weighted Average

$$Weighted\ Average = \frac{\sum xw}{\sum w}$$

where x is a data value and w is the weight assigned to that data value. The sum is taken over all data values.

So the weighted average is sum of the product of weightage and that value/sum of values. Where x is the data value and w is the weight assigned to that data value. The sum is taken over all data values.

(Refer Slide Time: 06:20)

Example

Suppose your midterm test score is 83 and your final exam score is 95. Using weights of 40% for the midterm and 60% for the final exam, compute the weighted average of your scores. If the minimum average for an A is 90, will you earn an A?

$$Weighted\ Average = \frac{(83)(0.40)+(95)(0.60)}{0.40+0.60}$$

$$= \frac{32+57}{1} = 90.2$$

You will earn an A!

We will see one application of this weighted average. Suppose your midterm test score is 83 and your final exam is score is 95 using weights of 40% is for the midterm and 60% is for the final exam compute the weighted average of your scores if the minimum average for an A grade is 90 will you earn an A grade. So first we find out the weighted average so the mark is 83 weights age is 40% for midterm for interim your mark is 95 weightage is 60 %.

So multiply that then divided by some of the weight that is 0.4 + 0. 6 is 1 so 90.2. So if you are above 90 you will get A, because you are crossing 90 obviously you will get the A grade.
(Refer Slide Time: 07:12)

Median

- Middle value in an ordered array of numbers

- Applicable for ordinal, interval, and ratio data

- Not applicable for nominal data

- Unaffected by extremely large and extremely small values

Now we will go to the next central tendency Median, the middle value in ordered array of number is called Median. It is applicable for ordinal interval and ratio data. You see previously the mean is applicable only for interval and ratio data but the median is applicable

for ordinal data. There is a point has to be remembered and it is not applicable for nominal data and one advantage of median is it is unaffected by extremely large and extremely small values.

(Refer Slide Time: 07:46)

Next we will see how to compute median there are 2 procedures, first procedure is arrange the observations in an ordered array. If there is an odd number of a term the median is the middle term of the ordered array. If there is the even number of terms the median is the average of middle two terms. Another procedure is the medians position of an ordered array is given by $n + 1 / 2$, n is the number of data set.

(Refer Slide Time: 08:15)

We will see this example; I have taken some exam some numbers that is arranged in an order that is an ascending order 3, 4, 5, 7 up to 22. There are 17 terms in the ordered array the

position of the median is, with respect to previous let n + 1 / 2. So n + 1 / 2, 17 + 1, 18 / 2 is 9. So the median is the 9th term, 9th term here is 15. If see the 22 which is the highest number is replaced by 100 still the median is 15.

See if the 3 is replaced by -103 still the median is 15. So there is the advantage of this median over mean is median is not disturbed by extreme values.

(Refer Slide Time: 08:59)

Previously the number of items are odd now let us see the another situation; there are 16 terms in the ordered array there is an even number, the position of the median is n + 1 / 2 that is 16 + 1 / 2 is 8.5. So we have to look at the term where it is the position of 8.5. That is the median is between 8th and the 9th term here the 8th term is 14, 9th term is 15 so average of that one is 14.5. Again, if the 21 is replaced by 100 the median is same 14.5, if the 3 is replaced by - 88 still the median is 14.5.

(Refer Slide Time: 09:42)

Median of Grouped Data

$$Median = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

Where :

L = the lower limit of the median class

cf_p = cumulative frequency of class preceding the median class

f_{med} = frequency of the median class

W = width of the median class

N = total of frequencies

Now let us see how to find out the median of your grouped data but it will be grouped data, here if the data is given in the form of a frequency table. This case the formula to find out the median of a group data is median equal to L + N / 2_cfp / f median multiplied by W. Where, L is the lower limit of the median class before using this formula from the given table you have to find out what is the median class.

Then cfp cumulative frequency of the class preceding the median class f median, the frequency of the median class, W is the width of the median class; N is the total number of frequencies.

(Refer Slide Time: 10:26)

Median of Grouped Data -- Example

Class Interval	Frequency	Cumulative Frequency
20-under 30	6	6
30-under 40	18	24
40-under 50	11	35
50-under 60	11	46
60-under 70	3	49
70-under 80	1	50
N = 50		

$$Md = L + \frac{\frac{N}{2} - cf_p}{f_{med}}(W)$$

$$= 40 + \frac{\frac{50}{2} - 24}{11}(10)$$

$$= 40.909$$

18

See this is an example; as I told you before using this formula first order to find out the median class. What is the median class is when you add the frequency it is a 50. 6 + 18 + 11

+ 11 + 3 + 1 is 50. So divide this 50 / 2 it is a 25. In the community frequency column in the last column look at where that 25 is lying it is not between 30 and 40; it is going to lie on between 40 and 50 because 24 for the next term is 35.

So the median class for this given group data is 40 and 50. So as usual L, is the lower limit of the median class that is a 40 + N is 50, 50 / 2 - 24. You see the cumulative frequency of the preceding interval is 24. So 50 / 2 - 24 / the frequency of the median class is 11, 11 W is 10, because the width interval is 10. When you simplify you would get 40.909, so this is the way to find out the median of your grouped data.

(Refer Slide Time: 11:45)

Now mode the most frequently occurring value in a data set is mode applicable to all level of data, measurement nominal, ordinal, interval and ratio. Sometimes there is a possibility the data set may be bimodal. Bimodal means data sets that have two modes. That means two numbers are repeated same number of time multimodal data sets that contain more than two modes.

(Refer Slide Time: 12:12)

Mode -- Example

- The mode is 44
- There are more 44s than any other value

35	41	44	45
37	41	44	46
37	43	44	46
39	43	44	46
40	43	44	46
40	43	45	48

See this one sample data as it is given for this data set the mode is 44, because the 44 is appearing more number of time. How many number of time 1, 2, 3, 4, 5. Okay? So the mode is 44. That is there are more 44s than any other values.

(Refer Slide Time: 12:37)

Mode of Grouped Data

- Midpoint of the modal class
- Modal class has the greatest frequency

Class Interval	Frequency
20-under 30	6
30-under 40	18
40-under 50	11
50-under 60	11
60-under 70	3
70-under 80	1

$$Mode = L_{Mo} + \left(\frac{d_1}{d_1 + d_2} \right) w =$$

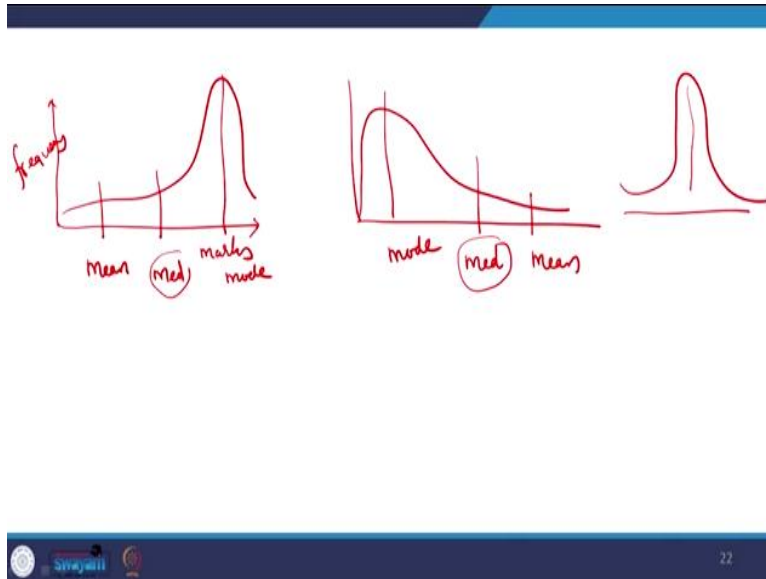$$30 + \left(\frac{12}{12 + 7} \right) 10 = 36.31$$

That is the formula for finding mode of a grouped data. Here first we have to find out the mode class. For that look at the frequency column there 18 is the highest frequency. So corresponding the n class interval is called mode interval. Okay? The mode interval L Mo is the lower limit of that mode interval is 30 + the d1 is difference between 18 and 6. And d2 is difference between 18 and 11.

See 30 + see d1 is nothing but 18 is the mode interval and then the previous frequency is 6, so 18 - 6 is 12 / d1 is 12 + d2 is the difference between your 18 and 11 that is 7. So 12 + 7

multiplied by width is 10, so 36.31 is the mode of your grouped data. Yes? We have studied mean, median, mode for group data and ungrouped data. Now the question is when to use mean? When to use median? When to use mode? Okay?

Many time even though we study mean, median, mode we are not exactly told how to use or when to use mean or when to use median or when to use mode?

(Refer Slide Time: 14:00)

For example look at this data set this is left skewed data because the tail is on the left hand side. The example for this is supposing see the exam is very easy question paper and the x axis is the marks and y axis is frequency. So there is more number of students who got higher marks .Where the question paper is easy situation this is an example of left skewed data. So what will happen here, here will be mean here will be median this will be mode.
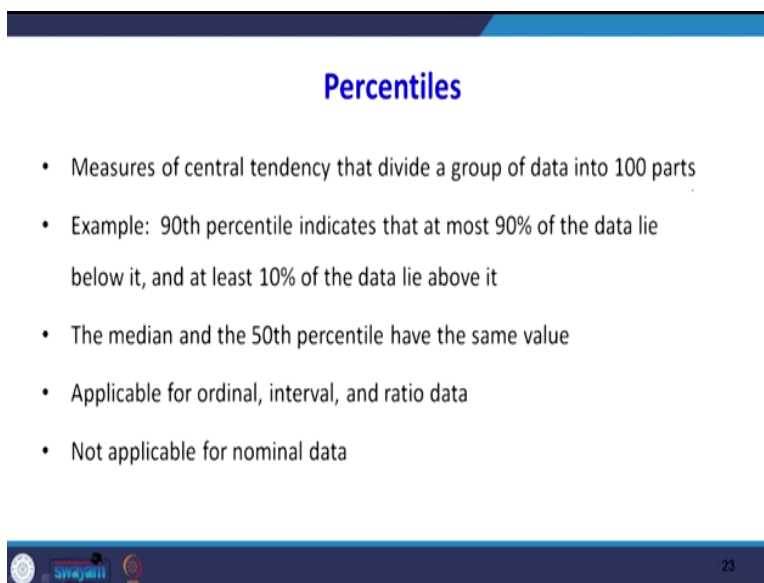
You see another example; where the question paper is very tough. So this is called right skewed data. You know here what is happening how we are saying that since the question paper is very tough. There are more number of students who got the lesser marks that is why the skewness on this side. So here there will be mean here will be median this will be mode Okay? There will be another situation it is symmetric it is a bell looking at bill shaped curve in this situation.

Now after looking at this hypothetical problem now the question arises when to use mean, when to use median, mode look at the location of the median. The median is always in the middle. Whether the data is left skewed or right skewed the median is always the middle. So

whenever the data is skewed you should go for median as a central tendency. If your data is following a bell-shaped curve then you can use mean, median, mode.
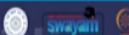
There is no problem at all the clue for that choosing the correct central is first you have to plot that curve go to plot the data outer plotting the data you have to get an idea of the skewness of the data set. How it is distributed? Whether it is right skewed or left skewed or it is bell shaped curve. If it is skewed data you go for median as the center tendency. If it is following a bell-shaped curve you go for mean or median or mode as a central tendency.

(Refer Slide Time: 16:39)

Percentiles

- Measures of central tendency that divide a group of data into 100 parts

- Example: 90th percentile indicates that at most 90% of the data lie below it, and at least 10% of the data lie above it

- The median and the 50th percentile have the same value

- Applicable for ordinal, interval, and ratio data

- Not applicable for nominal data

Now you go to next one is a Percentile, mainly this you might have seen some of the cat examination scores or gate examination scores their performance is expressed in terms of percentile not the percentage because percentile is having some advantage over percentage because percentage is absolute term but the percentile is the relative term the measure of central tendency that divide a group of data into 100 parts it is called percentile.

For example; somebody say 90th percentile my score is 90th percentile indicates that at most 90% of the data lie below it and at least 10% the data lie above it. Okay? The median and the 50th percentile have the same value. It is applicable for ordinal, interval and ratio data it is not applicable for nominal data.

(Refer Slide Time: 17:44)

Percentiles: Computational Procedure

- Organize the data into an ascending ordered array
- Calculate the p th percentile location:

$$i = \frac{P}{100}(n)$$

- Determine the percentile's location and its value.

- If i is a **whole number**, the percentile is the average of the values at the i and (i+1) positions

- If i is **not a whole number**, the percentile is at the (i+1) position in the ordered array

Okay we will see an example how to compute a percentile the first step is organize the data into an ascending ordered array calculate the pth percentile location. Suppose if I want to know 30th percent location for that you have to find out the value i, i is nothing but P / 100 multiplied by n, n is the number of data set, the i is nothing but the percentiles location we got to find out the i value.

If i is a whole number the percentage is the average of the values at the i and i + 1 positions.
If i is not a whole number the percentile is the i + 1 position in the ordered array.

(Refer Slide Time: 18:35)

Percentiles: Example

- Raw Data: 14, 12, 19, 23, 5, 13, 28, 17
- Ordered Array: 5, 12, 13, 14, 17, 19, 23, 28
- Location of 30th percentile:

$$i = \frac{30}{100}(8) = 2.4$$

- The location index, i, is not a whole number; i+1 = 2.4+1=3.4; the whole number portion is 3; the 30th percentile is at the 3rd location of the array; the 30th percentile is 13.
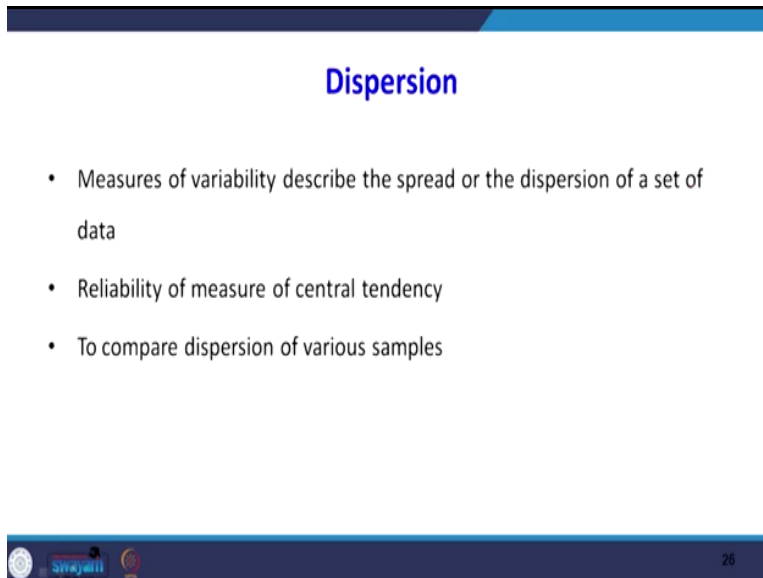
Look at this example the raw data is given 14, 12, and 19 up to 17. I have arranged in the ascending order the lowest value is 5, the highest value is 28. Suppose I want to know 30th percentile for knowing the 30th percentile, first I have to find out i that is a 30 / 100

multiplied by 8, so 2.4. The i is nothing but location index as I explained the previous slide, i is not the whole number. So you have to add i + 1, so 2. 4 + 1 is 3. 4.

In the 3.4 the whole number portion is 3 rights? So the 30th percentile is at the 3rd location of an array. When you look at the 3rd location is 13, that means a person who scored 13 marks his corresponding percentile is 30.
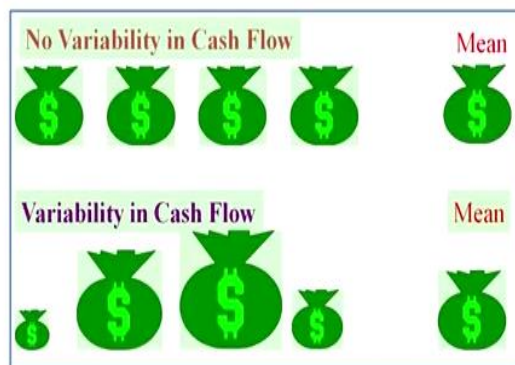
(Refer Slide Time: 19:26)

So far we talked about these different central tendencies will go for differing. Now we are going for measuring dispersion measures of variability describes the spread or the dispersion of the set of the data. The reliability of measure of central tendency is the dispersion because many times, the central tendency will mislead the people. So the reliability of that central tendency is calculated by or identified by its corresponding dispersion.

It is used to compare dispersion of various samples that is why whenever you plot the data you not only show the mean you have to show the central tendency also because the reliability of mean is explained by dispersion.

(Refer Slide Time: 20:14)

You look at this data, when you see the first two rows is no variability in cash flow mean is same. The second one is variability in cash flow see there is a lot of variability in the second one but the mean is same. If you look at only the mean it look like same when you look at only the mean the mean value same but when you look at see the left hand side the second dataset is having more variability. The quality of the mean is explained by its variability there is nothing but dispersion.

(Refer Slide Time: 20:51)

There are different measures to measure the variability one is the range, inter-quartile range, mean, absolute deviation, variance, standard deviation, z scores and coefficient of variations. We will see one by one.

(Refer Slide Time: 21:07)

Suppose there is ungroup of data is there see this one you have to find out the range. The range is nothing but the difference between the largest and the smallest value in a set of data. It is very simple to compute the problem here is it ignores all data points except the two extremes. So the range is the largest value is 48 in this data set the smallest value is 35. It is only 13 you see that only the two values are taken care in between the values is not taken into consideration for finding the range.
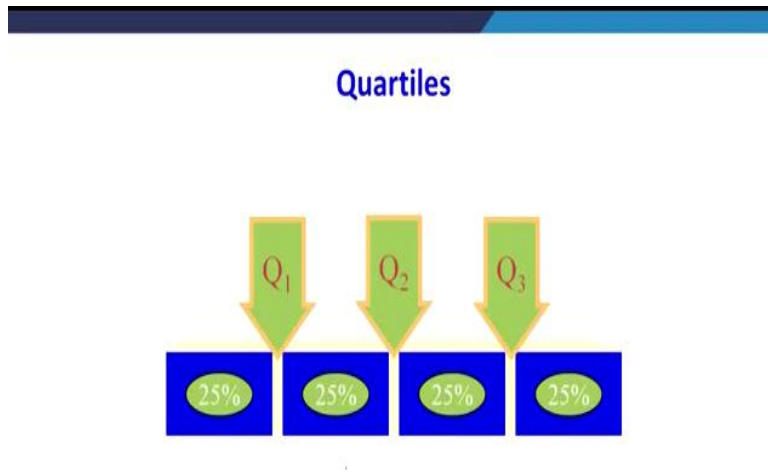
(Refer Slide Time: 21:46)

It is a quick estimate to measure the dispersion of a set of data. I will go for a quartile; quartile measures the central tendency that divided group of data into 4 subgroups. We say Q 1, Q 2, Q 3, Q 1 is nothing but 25 % of the data set is below the first quartile. Q2, 50 % of the data set is below the second quartile. Q3, 75 % the data set is below the third quartile. So we can say Q 1 is the 25th percentile Q 2 is the 50th the percentile nothing but the median.

This is a very important point; Q2 is nothing but the median. Q3 is the 75th percentile and another point is the quartile values are not necessarily members of the data set.

(Refer Slide Time: 22:34)

You see this lets say Q1, Q2, Q3. So Q1 see first 25 % the data set, Q2 first 50 % of the data set Q3, 75 % of the data set Okay? It is nothing but the quartile is used to divide the whole data set into 4 groups first 25, second 25, third 25 and last 25.

(Refer Slide Time: 22:59)

Quartiles: Example

- Ordered array: 106, 109, 114, 116, 121, 122, 125, 129
- Q_1 $\qquad i = \frac{25}{100}(8) = 2 \qquad Q_1 = \frac{109+114}{2} = 111.5$

- Q_2: $\qquad i = \frac{50}{100}(8) = 4 \qquad Q_2 = \frac{116+121}{2} = 118.5$

- Q_3: $\qquad i = \frac{75}{100}(8) = 6 \qquad Q_3 = \frac{122+125}{2} = 123.5$

Suppose an example for finding the quartile, suppose the data is given 106, 109 and so on. Okay? We have arranged it in the ascending order. First we got to find out the Q 1, Q 1 as I told you but the 25th percentile so the location of the 25th percentile. First you have to find out the location index i for the 25/ 100 into 8 it is a 2. Since the 2 is the even number. As I

explained previously if it is the location takes it 2 you have to find out that position plus the next position and its average.

So in the second positions data set is 109 the next one is 114 / 2 to 111.5. So the Q1 is nothing but here 111.5, Q 2 is 50th percentile 50 / 100 multiplied by 8 it is a 4, again the 4 is the even number. So the 4th location is 1, 2, 3, 4th location is 116 and 5th the location is 121 / 2, 118.5. So the Q2 our median is 118.5. Then Q3, 75 /100 multiplied by 8 again 6, 6 is the even number the average of 6th and 7th values are 122 + 125 / 2. So the value is 123.5.

(Refer Slide Time: 24:28)

Interquartile Range

- Range of values between the first and third quartiles
- Range of the "middle half"
- Less influenced by extremes
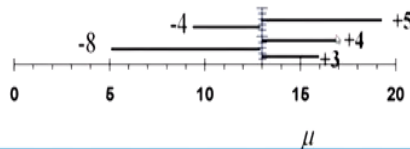
$$Interquartile\ Range = Q_3 - Q_1$$

This is the way to calculate Q 1, Q 2, and Q 3. Now the next term is interquartile range .So the dispersions in the data set is measured with help of interquartile range by using this formula Q3 - Q1. As we know Q 3 is 75th percentile Q1 is the 25th percentile so range of values between the first and third quartile is called interquartile range. It is a range of middle of why we are using quartile range because it is the less influenced by extreme values.

Because when we collect the data set we are not going to consider at very low values at the same time very high values. So the middle values which is not affected by extremes that is taken for further calculation .For that purpose we are using interquartile range.

(Refer Slide Time: 25:15)

Deviation from the Mean

- Data set: 5, 9, 16, 17, 18
- Mean: $\mu = \dfrac{\sum X}{N} = \dfrac{65}{5} = 13$
- Deviations from the mean: -8, -4, 3, 4, 5

There is a Q3, now we will go for deviation from the mean so dataset is the given 5, 9, 16, 17, and 18. To find the deviation from the mean first to find the mean, mean is 13. Suppose there is a graph is there so this is the 13 Okay? See the first value 5 the difference is 5 - 13 that is - 8. So this distance is your first deviation the second data is 9. So 9 - 13 is - 4 this is - 4. So this deviation is expressed by these lines.

Look at that there is a negative deviation there is a positive deviation. Suppose if we want to add the deviation general it will become 0. That is why we should go for mean absolute deviation.

(Refer Slide Time: 26:12)

Mean Absolute Deviation

- Average of the absolute deviations from the mean

X	$X - \mu$	$|X - \mu|$
5	-8	+8
9	-4	+4
16	+3	+3
17	+4	+4
18	+5	+5
	0	24

$$M.A.D. = \dfrac{\sum |X - \mu|}{N}$$

$$= \dfrac{24}{5}$$

$$= 4.8$$

You see this X is given here there are 2 values are negative deviation 3 are three values are positive deviation. When you add this it is becoming 0 so it seems we are getting 0 we cannot

measure the dispersion. One way is we have to remove this negatives you take only positive value. When you take positive values 24 so 24 / 5 there are 5 data set, so 4.8 is called mean absolute deviation. It is the average of absolute deviations from the mean.

(Refer Slide Time: 26:46)

Population Variance

- Average of the squared deviations from the arithmetic mean

X	$X - \mu$	$(X-\mu)^2$
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$

There was a problem in the mean absolute deviation I will tell you what is the problem there, see the next we will see population variance it is not the average of the squared deviation from the arithmetic mean. Okay? So the X is there, mean is there so when you add the absolute even digit is 0, one way the previously the mean absolute deviation you take an only positive value. Now we are going to square it, the squaring of the deviation having some advantage.

One advantage is we can remove the negative sign second one is but the deviation is less when you square it. For example; - 4 square is 16 see - 8 squared is 64. So what is happening more the deviation more this squared value. Okay? So here the Sigma square is Sigma of X - mue / X - mue whole square / N. So 130/5 here we are squaring the purpose why we are squaring for there are two reason one is to remove the negative sign, the second reason is giving for higher penalty for higher deviation values

The next one is the population standard deviation because already there is a variance but variance is a squared number that we cannot compare. Suppose the two numbers are given say 12 and 13 that is easy intuitively we can say which is higher which is smaller. Suppose 124, 169 is given notice squared number. We cannot compare intuitively and not only that it is in the square root of squared term.

We want to have it in the actual term so for comparison purpose for that purpose we are taking square root of that.

(Refer Slide Time: 28:25)

Population Standard Deviation

* Square root of the variance

X	X − μ	(X − μ)²
5	-8	64
9	-4	16
16	+3	9
17	+4	16
18	+5	25
	0	130

$$\sigma^2 = \frac{\Sigma (X - \mu)^2}{N}$$

$$= \frac{130}{5}$$

$$= 26.0$$

$$\sigma = \sqrt{\sigma^2}$$

$$= \sqrt{26.0}$$

$$= 5.1$$

So 5.1 is the standard deviation next we will go to the sample variance the formula is same but only thing is it is divided by n - 1.

(Refer Slide Time: 28:37)

Sample Variance

* Average of the squared deviations from the arithmetic mean

X	X − X̄	(X − X̄)²
2,398	625	390,625
1,844	71	5,041
1,539	-234	54,756
1,311	-462	213,444
7,092	0	663,866

$$S^2 = \frac{\Sigma (X - \bar{X})^2}{n-1}$$

$$= \frac{663,866}{3}$$

$$= 221,288.67$$

Why we are dividing by n -1, the reason is that to make the variance as the unbiased estimator. This is due to degrees of freedom since we already know the value of the mean will last one degrees of freedom. That we are dividing by n – 1 so it is very important whenever you find the sample variance so the in the denominator there should be a n - 1. So here the variance is 221,288.67.

This is another sample standard deviation; just to take the square root of that it is a 470.41 so a square root of the variance is nothing but standard deviation.

Now the purpose is why we have to study the standard deviation because the standard deviation is giving an indicator of financial risk. Higher the standard deviation is more risk lesser the standard deviation less at risk. In quality control context generally when we manufacture something suppose here plant A and plant B are shift A and shift B whenever the variances in lesser then the quality of the product is high.

The process capability also they should have the lesser variance means in the process capabilities high. Then I suppose therefore comparing the populations household income of 2

cities, employee absenteeism in 2 plants for these purpose it is for comparing the population that means wherever there is a lesser standard deviation so that is having higher homogeneous data set.

(Refer Slide Time: 30:12)

Standard Deviation as an Indicator of Financial Risk

Financial Security	Annualized Rate of Return	
	μ	σ
A	15%	3%
B	15%	7%

You see look at this one mue 1 Sigma, see this is a financial security A and B. See the return rate is 15, 15 it is both are giving equal return but look at this the Sigma standard deviation because in financial context it is, it is measured as the risk. So the first one is 3% second with 7% so the security B having a higher risk, so always we will go for where there is a lesser standard deviation because mean is same.

We are the same time the risk all should be same so far we have seen different central tendencies, different dispersions. In the coming class will use Python will take some sample data set. I will explain you how to find out central tendency and the dispersion of the given data set. Thank you very much.