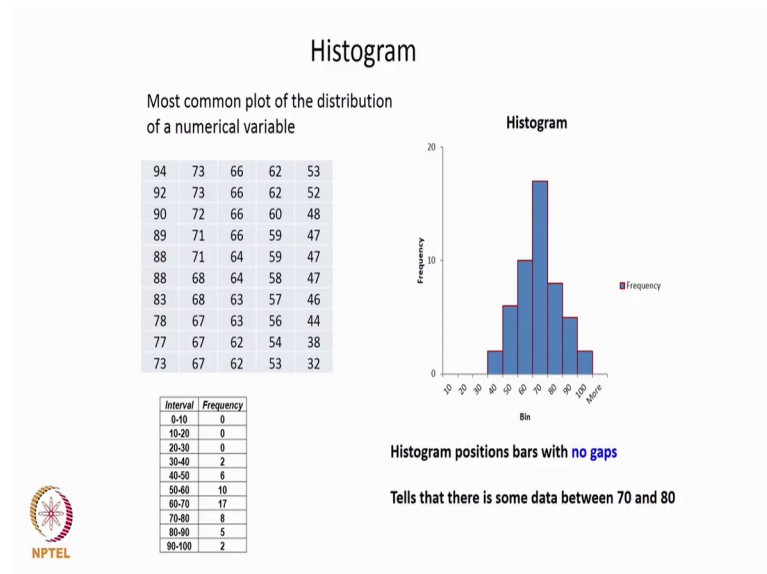


**Introduction to Probability and Statistics**  
**Prof. G. Srinivasan**  
**Department of Management Studies**  
**Indian Institute of Technology, Madras**

**Lecture - 06**  
**Describing Numerical Data**

In this lecture we describe methods to represent, understand and present numerical data.

(Refer Slide Time: 00:27)



So, let us look at this slide, and in this slide we are showing let us say the marks obtained by 50 students in a class, in an examination. When we classified data, we first classified them as categorical and numerical and then we categorized them into four, where we had the nominal and ordinal as categorical data and interval and ratio as numerical data, and we will marks obtained by students can be taken as ratio level data, because not only the differences are measurable and comparable. We can also say that if somebody is called a 50 and somebody else called a 100, we could say that this one got 2 times the mark that the other person cut. So, we have data which represents marks obtained by 50 students in an exam, and one can assume that the maximum is 100 and out of 100 these marks are given.

So, first one way of representing is to, there are 50 pieces of data; therefore, we can have, we can represent this data in form of interval and the corresponding frequency. So, if we assume that this data are marks between say 0 and 100. Now we could divide that range

of 100 to 10 intervals 0 to 10, 10 to 20, 20 to 30 and so on as shown here and then the frequency is the number of instances, where the mark fits into this particular interval or range.

So, there is nobody who has got marks between 0 and 10 and so on. So, between 30 and 40 we have two people getting marks between 30 and 40 and so on. So the frequency would add up to 50 which you can observe. So, one way to represent this type of data is to put it into a table like this where we have intervals and then we have corresponding frequency for each of these intervals.

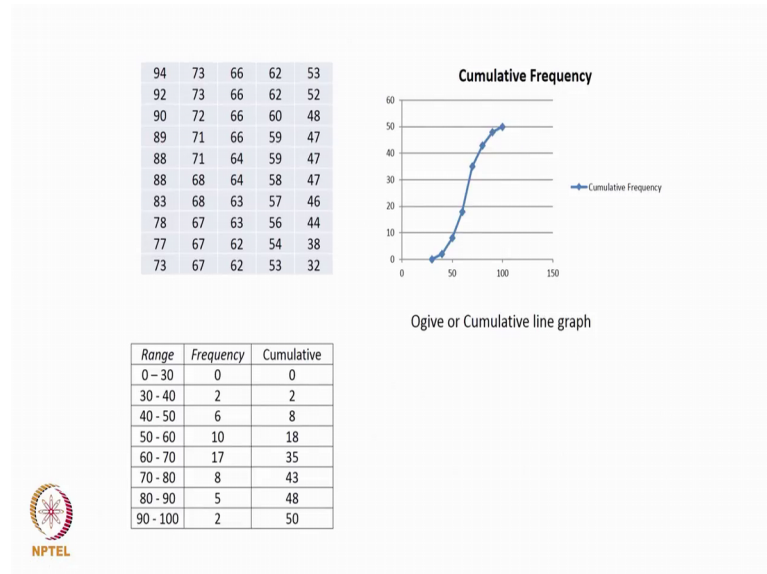
So, the immediate question is can we put this in a picture or can we represent this pictorially; like we did for the earlier types of data. So, this picture which we have shown here is the pictorial representation of this and this picture is called a histogram. Now this histogram is looks like a bar chart, but it is different from a bar chart. The most important difference that we see here is shown here through this, histogram positions bars with no gaps. When we use the term bar chart there would be gaps between the bars, whereas, in a histogram there is no gap.

So, what is that represent? It represents that for example, there is no gap between 70 and 80, it means there is some data between 70 and 80. Whereas, if we left the gap as we did in a bar chart, then we realize that there was actually no data in the place where there is a gap in a histogram. There is no gap and that is something which we need to understand numerical data are represented in the form of histogram and this is the histogram for this.

Once again as mentioned in an earlier lecture you would observe that all the bars are the, the bars that are present in this histogram we have used the same color and we have not used different colors. Sometimes when we represent data, we feel that if we used different colors, it might be even more pleasing to the eye, so it is not suggested, because all these represent only as the single variable or a single thing under consideration which is the mark.

So, when we are representing multiple things then it is customary to use different colors. So, long as all the data is marks obtained which is a single variable; we use the same color to represent this.

(Refer Slide Time: 04:38)



There other ways of representing it. There is also the thing called accumulative frequency, which means, we look at the previous data, that the original data is given here and then we say that earlier table we started from 0 to 10, 10 to 20, and so on. So, right now we say 0 to 30 the, there is nobody, so frequency cumulative is 0. 30 to 40 there are two people, so the cumulative is 2. 40 to 50 you realize that the cumulative is 2 plus 6 8.


So, what it represents is even though we have used the range here, so we will say less than or equal to 50 there are eight, less than or equal to 60 there are 18 and so on and then we plot the cumulative numbers and you see at the end the cumulative number adds up to 50 which is the total number. So, this is another way to represent which is called the cumulative line graph to represent this kind of data.

(Refer Slide Time: 05:39)

Stem and Leaf

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

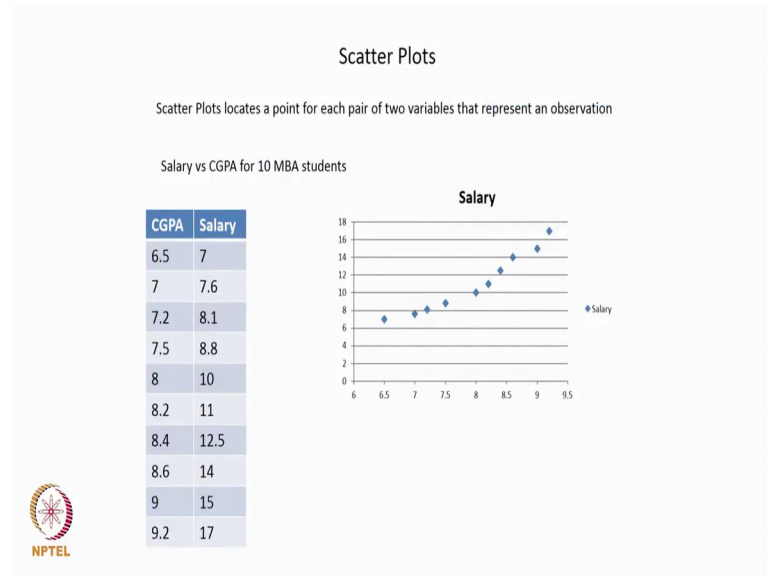
Cumulative Frequency	Stem	Leaf
2	3	2 8
8	4	4 6 7 7 8
17	5	2 3 3 4 6 7 8 9 9
35	6	0 2 2 2 2 3 3 4 4 6 6 6 6 7 7 7 8 8
44	7	1 1 2 3 3 3 7 7 8
48	8	3 8 8 9
50	9	0 2 4



The third representation is also called a stem and leaf representation and then we realized that when we had a cumulative frequency of 2, we had numbers 32 and 38. So, these two numbers 32 and 38, the 3 is the stem in this example, since all of these are two digit numbers, the 10s digit or the left most digit will act as a stem and the right digit acts as the leaf and then this means there are two numbers, so 32 and 38.

Now, in this case the cumulative frequency is eight, but there are 6 numbers which are in the 40s, so the stem is 40 and they are 44 46 47 47 47 48 and so on. So stem and leaf representation as another way to represent, but the most common way to represent is the histogram that we saw in the first slide.

(Refer Slide Time: 06:41)



We can also represent it in the form of scatter plots. So, scatter plot locates a point for each pair of two variables that represent an observation. For example, if we collect data from say 10 MBA students about their salary and their CGPA or their academic performance. And in this we have shown the academic performance as the x axis and the salary as the y axis. Now there are these pairs, there is a point which represents and academic performance of 6.5 out of 10 and say a salary of 7 lakhs and that is represented by this point.

Similarly, the academic performance of 9.2 and a salary of 17 lakhs is represented by this point and this kind of a plot is called a scatter plot, and what is more important to understand here is, the scatter plot locates a point for each pair of two variables that represent an observation. So, in this case the pair of variables are the academic performance and the associated salary.

(Refer Slide Time: 07:51)

Measures of central tendency


Mean, Median, Mode

$$\text{Sample mean } \bar{x} = \frac{\sum_{i=1}^n X_i}{n}$$

Mean = 64.5  
Median = 64  
Mode = 66, 62

94	73	66	62	53
92	73	66	62	52
90	72	66	60	48
89	71	66	59	47
88	71	64	59	47
88	68	64	58	47
83	68	63	57	46
78	67	63	56	44
77	67	62	54	38
73	67	62	53	32

Sorted data



Now, how do we represent this data even better in the sense? If there are 50 observations or 50 marks that we have, are there measures of central tendency. In the previous lectures when we looked at categorical data, we looked at nominal data we said mode was the measure of central tendency. Whereas, in ordinal data where we could order the data, we said both mode and median would be the measures of central tendency.

Now, with numerical data we now observe that all the three exist; the mean the median and the mode are measures of central tendency. Most of us almost all of us know how to calculate these three values, so we would do that one more time. So, in order to calculate all of them, the easier thing to do is to sort them in decreasing order or increasing order as the case may be. Well that is required to calculate the median, it is not absolutely necessary to calculate the mean and the mode.

So first we have already seen how to calculate the median and the mode, so mode as we have seen earlier is that, is the number that has the highest frequency. Mode, we observe that we have four places we have 62 1, 2, 3 and 4. We also have 66 in four places and therefore, we have two modes for this data which is 66 and 62. We also have seen earlier how to calculate or find out the median, sort these values in decreasing or increasing order.

So, which has been shown, already it is shown in the sorted order and since there are 50 observations which is an even number, 50 divided by 2 is 25, the median is actually the

average of the 25th and the 26th observation. So, in this case, this is the 10th observation, twentieth observation, 21, 22, 23, 24. Both the 25th and the 26th observation are 64 and therefore, the median is 64.

Arithmetic mean calculation all of us know, we have done it so many times. Add all the values and divided by the total number of values, so there are 50. So, sum all these 50 numbers and that is shown as sigma which represents summation  $i = 1$  to  $n$   $X_i$   $X_i$  is the individual observation, so this means sum  $X_1$  to  $X_{50}$ . So, this is  $X_1 + X_2 + X_3$  and so on divided by  $n$  which is 50.

So, sum all of them and divide it by 50 to get 64.5 as the mean or the arithmetic average or arithmetic mean. So these three measures of central tendency are used to represent numerical data which are the mean, the median and the mode. And for this data that be now have with us, the mean is calculated to be 64.5, the median is 64 and the mode is 66 and 62. Other measures which we will also see as we move along.


(Refer Slide Time: 11:23)

Exercise

Pay package in lakhs for 50 students is given below:

18	11	10.2	8.5	7.7
17.4	11	10.2	8.5	7.7
16.5	10.6	9.9	8.4	7.7
15.9	10.6	9.9	8.4	7.7
11.2	10.6	9.9	8.4	7.7
11.2	10.6	9.6	8.4	7.7
11.2	10.6	9.6	8.4	7.7
11.2	10.5	9.6	8.3	6.9
11.2	10.5	9.3	8.2	6.9
11	10.5	9.3	7.7	6

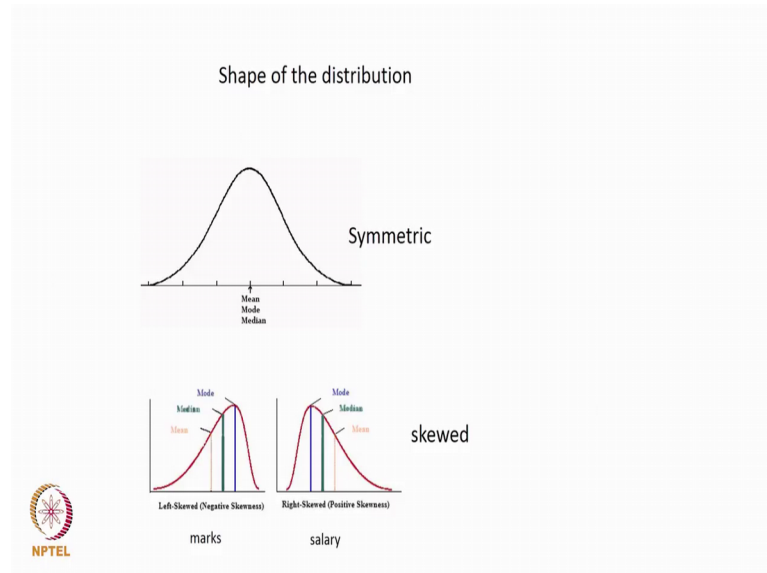
Compute the mean, median and mode



Now, we have a small exercise which you can do. So, this could be pay package in lakhs for 50 students of a management school, this could be given here and then you can calculate the mean, the median and the mode, in exactly the way we did in the earlier case. So, let us do the mode and one could, perhaps see that 7.7 has 1 2 3 4 5 6 7 8 9 observations and would become the mode, the median is the, again the, the average between the 25th and the 26th, so we have 9.9 and 9.6 as the 25th and 26th values.

Therefore, the median is the average of these two which will become 9.75. The arithmetic mean can be calculated by adding all these and dividing it by 50.

(Refer Slide Time: 12:18)



Sometimes we represent these data also in the form of a curve or a distribution and the from the shape of these distributions we also draw some conclusions, we will see about this as we move along. Now, this is called a symmetric distribution, and in a symmetric distribution the mean, median and the mode are all in the middle. Now these are examples of skewed distributions or skewed distributions, so this is called a right skewed.

So, the skew is only the longer part comes to the right, so it is called a right skewed and here this is left skewed and it is customary that this is how the mode, median and mean are when the distribution is skewed. So, when the distribution is positive skewed or right skewed, you realize the mean is higher and then the median and then the mode, whereas if it is left skewed, the mode is higher and then the median and then the mean. We will see some of these as we move along




(Refer Slide Time: 13:20)

Measures of variation (dispersion)

Range, Inter quartile range	<table border="1" style="border-collapse: collapse; font-size: small;"> <tr><td>94</td><td>73</td><td>66</td><td>62</td><td>53</td></tr> <tr><td>92</td><td>73</td><td>66</td><td>62</td><td>52</td></tr> <tr><td>90</td><td>72</td><td>66</td><td>60</td><td>48</td></tr> </table>	94	73	66	62	53	92	73	66	62	52	90	72	66	60	48					
94	73	66	62	53																	
92	73	66	62	52																	
90	72	66	60	48																	
Variance and Standard deviation	<table border="1" style="border-collapse: collapse; font-size: small;"> <tr><td>89</td><td>71</td><td>66</td><td>59</td><td>47</td></tr> <tr><td>88</td><td>71</td><td>64</td><td>59</td><td>47</td></tr> <tr><td>88</td><td>68</td><td>64</td><td>58</td><td>47</td></tr> </table>	89	71	66	59	47	88	71	64	59	47	88	68	64	58	47					
89	71	66	59	47																	
88	71	64	59	47																	
88	68	64	58	47																	
Coefficient of Variation	<table border="1" style="border-collapse: collapse; font-size: small;"> <tr><td>83</td><td>68</td><td>63</td><td>57</td><td>46</td></tr> <tr><td>78</td><td>67</td><td>63</td><td>56</td><td>44</td></tr> <tr><td>77</td><td>67</td><td>62</td><td>54</td><td>38</td></tr> <tr><td>73</td><td>67</td><td>62</td><td>53</td><td>32</td></tr> </table>	83	68	63	57	46	78	67	63	56	44	77	67	62	54	38	73	67	62	53	32
83	68	63	57	46																	
78	67	63	56	44																	
77	67	62	54	38																	
73	67	62	53	32																	

Sorted data

Five number summary of data



We now study measures of variation or dispersion of data. We will study range, inter quartile range, variance standard deviation and coefficient of variation. We use the same example which has marks obtained by 50 students as the data, using which we will learn these concepts. We will also learn what is called a five number summary of data. Now we have shown the marks data sorted in descending order or decreasing order in this table. We would also be using the same data sorted in the increasing or ascending order which we will show in subsequent slides.


(Refer Slide Time: 14:07)

Median and Inter quartile range

<table border="1" style="border-collapse: collapse; font-size: small;"> <tr><td>53</td><td>52</td><td>63</td><td>59</td><td>62</td></tr> <tr><td>48</td><td>47</td><td>66</td><td>54</td><td>67</td></tr> <tr><td>62</td><td>72</td><td>46</td><td>53</td><td>68</td></tr> <tr><td>58</td><td>77</td><td>38</td><td>66</td><td>83</td></tr> <tr><td>66</td><td>60</td><td>78</td><td>90</td><td>88</td></tr> <tr><td>73</td><td>88</td><td>62</td><td>32</td><td>73</td></tr> <tr><td>89</td><td>94</td><td>68</td><td>47</td><td>62</td></tr> <tr><td>92</td><td>73</td><td>67</td><td>64</td><td>59</td></tr> <tr><td>66</td><td>71</td><td>67</td><td>56</td><td>44</td></tr> <tr><td>57</td><td>64</td><td>71</td><td>63</td><td>47</td></tr> </table>	53	52	63	59	62	48	47	66	54	67	62	72	46	53	68	58	77	38	66	83	66	60	78	90	88	73	88	62	32	73	89	94	68	47	62	92	73	67	64	59	66	71	67	56	44	57	64	71	63	47	<p style="font-size: x-small;">Sorted in ascending order</p> <table border="1" style="border-collapse: collapse; font-size: small;"> <tr><td>32</td><td>53</td><td style="color: blue;">62</td><td>67</td><td>73</td></tr> <tr><td>38</td><td>54</td><td style="color: blue;">62</td><td>67</td><td>77</td></tr> <tr><td>44</td><td>56</td><td>63</td><td>67</td><td>78</td></tr> <tr><td>46</td><td>57</td><td>63</td><td>68</td><td>83</td></tr> <tr><td>47</td><td>58</td><td>64</td><td>68</td><td>88</td></tr> <tr><td>47</td><td>59</td><td>64</td><td>71</td><td>88</td></tr> <tr><td>47</td><td>59</td><td style="color: blue;">66</td><td>71</td><td>89</td></tr> <tr><td>48</td><td>60</td><td style="color: blue;">66</td><td>72</td><td>90</td></tr> <tr><td>52</td><td style="color: blue;">62</td><td style="color: blue;">66</td><td>73</td><td>92</td></tr> <tr><td>53</td><td style="color: blue;">62</td><td style="color: blue;">66</td><td>73</td><td>94</td></tr> </table>	32	53	62	67	73	38	54	62	67	77	44	56	63	67	78	46	57	63	68	83	47	58	64	68	88	47	59	64	71	88	47	59	66	71	89	48	60	66	72	90	52	62	66	73	92	53	62	66	73	94
53	52	63	59	62																																																																																																	
48	47	66	54	67																																																																																																	
62	72	46	53	68																																																																																																	
58	77	38	66	83																																																																																																	
66	60	78	90	88																																																																																																	
73	88	62	32	73																																																																																																	
89	94	68	47	62																																																																																																	
92	73	67	64	59																																																																																																	
66	71	67	56	44																																																																																																	
57	64	71	63	47																																																																																																	
32	53	62	67	73																																																																																																	
38	54	62	67	77																																																																																																	
44	56	63	67	78																																																																																																	
46	57	63	68	83																																																																																																	
47	58	64	68	88																																																																																																	
47	59	64	71	88																																																																																																	
47	59	66	71	89																																																																																																	
48	60	66	72	90																																																																																																	
52	62	66	73	92																																																																																																	
53	62	66	73	94																																																																																																	

Mode = 62, 66

Total marks of 50 students in a course



**Median** is the middle value  
 Median is the 50<sup>th</sup> percentile of the data  
 It is the second quartile of the data

So, we first look at median and inter quartile range again. We have already seen that the median is the middle value. Now to explain the basic data is shown in the left hand side here, total marks of 50 students in a course. Now to do this analysis to find the median or to do the simple computation to find the median, we first sort the data in ascending order or increasing order or non decreasing order and then the sorted ascending order is shown here.

We have already seen that the mode is that value which repeats maximum number of times, and from this data we observed that both 62 marks and 66 marks appear 4 times and therefore, both qualify to be the mode. So, this is called bimodal data where there are two modes. We have seen that the median is the middle value after the data is sorted in ascending order.

Now, we have even number  $n$  equal to 50, so the middle value is 25, but since we have 50 data points which is an even number, the middle value being 25, the median is the average of the 25th and the 26th value and we observe in the sorted order that both the 25th and the 26th value are the same which is 64 and therefore, the median which is the average of these two is also 64.

Now we define median not only as the middle value of the sorted, increasing sorted order, the median can also be seen as the 50th percentile of the data and the median is seen as a second quartile of the data. So, this leads us to understanding what is percentile, what is quartile, how many quartiles are there and so on. And we do that first and then we also tried using that we find out what is called the inter quartile range of the data, range of the data per se. Range is the difference between the maximum and the minimum value.

So, since we have sorted this in increasing order, the minimum value comes first which is 32, the maximum value is 94 and the range is the difference between the maximum and the minimum value, which is 94 minus 32 which is 66. Now, we go on to explain what is percentile and what is quartile.

(Refer Slide Time: 17:10)

Percentile and Quartile


Percentile is the value below which a given percentage of observations in a group of observations fall.

$P^{\text{th}}$  percentile of the data is the smallest value in the list (in ascending order) such that no more than  $P\%$  of the data points is strictly less than the value and at least  $P\%$  is less than or equal to that value.

Calculate percentiles .

Find  $\frac{P \times N}{100}$ . If it is a fraction rank = upper integer value (n). If it is an integer,

rank = average of n and n+1 values



There are 4 quartiles. The first quartile is the 25% percentile, the second is the 50<sup>th</sup> percentile (median), the third is 75<sup>th</sup> percentile and the fourth is the last point which is 100<sup>th</sup> percentile

Percentile is the value below which a given percentage of observations in a group of observations fall. So, P-th percentile of the data, is a smallest value in the list in ascending order; such that no more than P percent of the data points is strictly less than the value, and at least P percent is less than or equal to that value.

So, let me repeat, no more than P percent of the data points is strictly less than the value, and at least P percent is less than or equal to that value. So, how do we calculate the percentiles and how do we relate the percentile to the median or how do we relate the median to a percentile. To do that first find P into N divided by 100. There are many methods available, more than one method available to calculate percentiles, we have to use one of them consistently and we are following one of them consistently.

So, the method is to find P into N by 100 and if it is a fraction, rank of that position of the percentile is the upper integer value of N, where n is the calculated value of P into capital N by 100, small n is the rank from which we get the rank, so small n is equal to P into capital N by 100. If small n is a fraction then the rank is equal to the upper integer value of small n. If small n is an integer then the value is the average of the n and n plus 1 values, we will show that using examples.

So, we can find out the 50th percentile, we can find out the eightieth percentile, we can find out the 40th percentile and so on. So, in our earlier example capital N is equal to 50 and if we wish to find the 50th percentile, so P is also equal to 50, 50th percentile of 50

data points. So,  $p$  into  $n$  by  $100$  is  $50$  to  $50$  by  $100$  which is  $25$ . Now, the rank therefore is, the computed value is  $25$ , so the  $50$ th percentile value will be the average of the  $25$ th and the  $26$ th value, because we have said if it is an integer then take the average of the  $n$  and  $n$  plus  $1$  values.

If we want to find the  $25$ th percentile of the data that we have, then the calculated value is  $25$  into  $50$  by  $100$ , capital  $N$  is always  $50$  because there are  $50$  data points. If we wish to find the  $25$ th percentile,  $P$  is equal to  $25$ . So,  $P$   $N$  by  $100$  will become  $12.5$  which is a fraction and we take the upper integer value. Therefore, we will take the  $13$ th value. We will show this as we move along.

So, any percentile we can calculate from  $1$  to  $100$ . The  $100$ th percentile of course, will be the largest of the points. Now the several percentiles are possible, where as we define only four quartiles: first quartile is the  $25$ th percentile, the second quartile is the  $50$ th percentile, the third quartile is the  $75$ th percentile and the fourth quartile which is the last point is the  $100$ th percentile. So, this is the definition of percentiles and quartiles. We now go on to show the computation using the example that we have.

(Refer Slide Time: 21:10)

32	53	62	67	73
38	54	62	67	77
44	56	63	67	78
46	57	63	68	83
47	58	64	68	88
47	59	64	71	88
47	59	66	71	89
48	60	66	72	90
52	62	66	73	92
53	62	66	73	94

Minimum = 32

Ordinal rank of 25% percentile  
Lower quartile = 12.5;  $n = 13$   
25% percentile value = 56


Ordinal rank of 50% percentile = 25  
Median =  $(64 + 64)/2 = 64$

Ordinal rank of 75% percentile = 37.5  
 $n = 38$ ; 75<sup>th</sup> percentile = 72

Five number summary of data  
32, 56, 64, 72, 94  
IQR = 16

Inter Quartile range = 16

Maximum = 94



Now, this is the example in the sorted order, so minimum is  $32$ ,  $25$ th percentile we already explained that we now find  $25$  into  $50$  by  $100$ , the  $25$  comes from  $P$  equal to  $25$ th percentile, capital  $N$  is equal to  $50$  because we have  $50$  data points.

So, the ordinal rank of the 25th percentile will become 12.5, which has 25 into 50 by 100 and since it is a fraction or it has it is a decimal we take the upper integer value, n becomes 13. And therefore, the 25th percentile value or the lower quartile as, it is called is the 13th value which is 56 the median is the 50th percentile.

So, the rank of the 50th percentile based on the calculation is 50 into 50 by 100 which is 25. Since it is an integer we take the average of the 25th and the 26th value, both happened to be 64 and the median is therefore, 64. The rank of the 75th percentile will be 75 into 50 by 100 which is 37.5, so n equal to 38 because it is a decimal and the value is the value in the 38th position which is 72.

So, the lower quartile is 56, the median is 64 the 50th percentile which is a second quartile, which is also the median is 64. The third quartile or the upper quartile or 75th percentile is 72 for this data, and the inter quartile range is the difference between the third quartile and the first quartile 72 minus 56 which is 16. And now we define what is called a 5 member summary of data, it starts from the minimum, the first quartile the median the upper quartile and the maximum with inter quartile range is equal to 16

(Refer Slide Time: 23:26)


Exercise

Pay package in lakhs for 50 students is given below:

18	11	10.2	8.5	7.7
17.4	11	10.2	8.5	7.7
16.5	10.6	9.9	8.4	7.7
15.9	10.6	9.9	8.4	7.7
11.2	10.6	9.9	8.4	7.7
11.2	10.6	9.6	8.4	7.7
11.2	10.6	9.6	8.4	7.7
11.2	10.5	9.6	8.3	6.9
11.2	10.5	9.3	8.2	6.9
11	10.5	9.3	7.7	6

Compute the Five number summary of data, IQR and range?

Minimum = 6  
 Lower quartile = 8.3  
 Median = 9.75  
 75<sup>th</sup> percentile = 10.6  
 Maximum = 18  
 IQR = 2.3  
 Range = 12



So, we now do an exercise where we compute the 5 number summary of the data, the IQR and the range pay package in lakhs for 50 students is given below. Now when we compute the 5 number summary of data, we should sort the given data in the ascending order. Now the data is given in the descending order and therefore, we approach it from

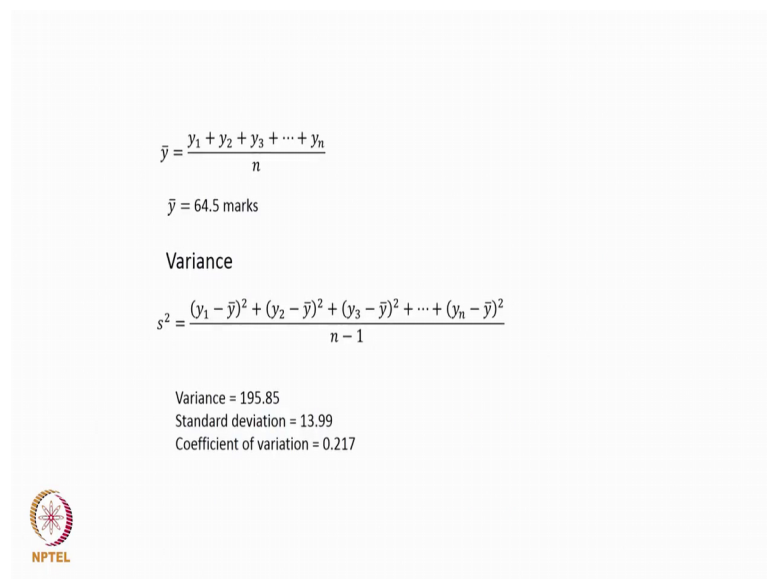
this side which is the ascending order. We could alternately sort it in ascending order and start doing as we did in the earlier example.

Now, from this we observe that the minimum is 6 which is now the last number. The lower quartile we already saw from the earlier example that  $n$  is equal to 13, because 25th percentile will give us 12.5, the upper integer value is 13. So, the 13th value is the 25th percentile, so the 13th value is here 11 12 13, so 8.3 is the value of the 25th percentile. Median is the 50th percentile. And since we have 50 points  $n$  is equal to 25, which is an integer and therefore, we take the average of the 25th and the 26th values which happened to be 9.6 and 9.9 and the median is 9.75.

We should also note that in these computations, we can have a median which is actually not a data point, where as a mode will have to be a data point. The 75th percentile we have already done the calculation  $37.5 \times 75$  into 50 by 150 data points 75th percentile, the calculation gives us 37.5 which is a fraction and therefore, we take the 38th value. The 38th value is here which is 10.6 the maximum is 18.

The inter quartile range is the difference between the 75th percentile, the upper quartile and the 25th percentile which is the lower quartile and this works out to be 2.3. The range is the difference between the maximum and the minimum and works out to be 12.

(Refer Slide Time: 26:01)




$$\bar{y} = \frac{y_1 + y_2 + y_3 + \dots + y_n}{n}$$

$\bar{y} = 64.5$  marks

Variance

$$s^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 + \dots + (y_n - \bar{y})^2}{n - 1}$$

Variance = 195.85  
Standard deviation = 13.99  
Coefficient of variation = 0.217



We next look at another measure which is called variance.

(Refer Slide Time: 26:06)


Measures of variation (dispersion)

Inter quartile range	94	73	66	62	53
	92	73	66	62	52
Variance and Standard deviation	90	72	66	60	48
	89	71	66	59	47
Coefficient of Variation	88	71	64	59	47
	88	68	64	58	47
Median = 64	83	68	63	57	46
Mode = 66	78	67	63	56	44
Lower quartile = 72.5	77	67	62	54	38
Upper quartile = 56.5	73	67	62	53	32
IQR = 16					

Sorted data

Minimum = 32  
 Lower quartile = 56.5  
 Median = 64  
 75<sup>th</sup> percentile = 72.5  
 Maximum = 94

Five number summary of data



(Refer Slide Time: 26:11)


Measures of central tendency

Mean, Median, Mode	94	73	66	62	53
	92	73	66	62	52
	90	72	66	60	48
	89	71	66	59	47
	88	71	64	59	47
	88	68	64	58	47
	83	68	63	57	46
	78	67	63	56	44
	77	67	62	54	38
	73	67	62	53	32

Sorted data

Sample mean  $\bar{X} = \frac{\sum_{i=1}^n X_i}{n}$

Mean = 64.5  
 Median = 64  
 Mode = 66



We have already seen that for this data in the earlier slide, for this data the mean was 64.5 which is the sum of all the 50 values divided by 50.

So, in this case we have again shown the computation of this, which is  $\bar{y}$ . one can use  $\bar{y}$ , one can also use  $\bar{X}$  and so on. So  $\bar{y}$  now represents the marks that we have for these 50 students, so  $\bar{y}$  is the sum of all these 50 marks divided by 50 which gives us 64.5. We now define this measure called variance and variance is  $\sum (y_i - \bar{y})^2$

whole square plus  $y^2$  minus  $y$  bar the whole square and so on up to  $y_{50}$  minus  $y$  bar the wholes divided by  $n - 1$  which is 49.

So,  $S^2$ ,  $S$  is used to represent sample, so when we compute sample variance we divided by  $n - 1$  and when we compute population variance we divide by  $n$ . So, in this example we have taken a sample of 50 students, so to calculate the sample variance we divided by 49. So, we take the first value and subtract 64.5 and square it and do that for all the 50 values and then divided by 49.

Since each of these terms in the numerator is a positive quantity or a 0, it is a non negative quantity, because its squares a difference between two numbers and then be some 50 such non negative quantities and divided by 49. So the variance is always a non negative or a positive quantity. So, in this example the variance is 195.85, the standard deviation is the square root of the variance, it is a positive square root of the variance and therefore, standard deviation is 13.99 in this example.

And we have also listed another term called coefficient of variation which is  $\sigma$  by  $\bar{y}$  13.99 divided by 64.5, which is 0.217. So, we have introduced several measures of spread or dispersion of data. In the previous slides we saw the inter quartile range and in this slide we saw measures; such as variance standard deviation and coefficient of variation.

Now, variance is a quadratic kind of a measure, because we take the deviation and square it. Whereas, in the earlier when we did the quartiles, we did not square it we only looked at the differences. Now variances square measure, standard deviation is root of that and coefficient of variation also depends on the mean and the standard deviation. So, these three are interrelated and we will now also see some situations in which we use this and try to understand where each one is actually applicable.




(Refer Slide Time: 29:32)

Six months earnings of a businessman is given: 5.4, 7.3, 10.9, 3.2, 4.7, 11.4. Find the mean and variance?

Month	Earnings	Deviation	Squared
1	5.4	$5.4 - 7.15 = -1.75$	3.0625
2	7.3	$7.3 - 7.15 = 0.15$	0.0225
3	10.9	3.75	14.0625
4	3.2	-3.95	15.6025
5	4.7	-2.45	6.0025
6	11.4	4.25	18.0625
Sum	42.9	0	56.815

$Mean = \frac{42.9}{6} = 7.15$        $variance = \frac{56.815}{5} = 11.363$



Now, let us look at some data and also try to show the computation of the variance and the standard deviation. So, let us assume the 6 months earnings of a businessman is given by 5.4 7.3 etcetera, you can assume that these earnings are in lakhs and find out the mean and the variance. So, reasonably straightforward computation so the mean is the sum of these 6 earnings which comes to 42.9 divided by 6 which is 7.15.

Find out the variance, now 5.4 minus 7.15, 7.3 minus 7.15, 10.9 minus 7.15. So, these are the deviations these are the squared deviations and this is the sum of the squared deviation, again divided by n minus 1 to get variance of 11.363 when we did just to understand the unit of variance.


(Refer Slide Time: 30:31)

Variance

$$s^2 = \frac{(y_1 - \bar{y})^2 + (y_2 - \bar{y})^2 + (y_3 - \bar{y})^2 + \dots + (y_n - \bar{y})^2}{n - 1}$$
$$s^2 = \frac{9596.5}{49} = 195.85 \text{ marks squared}$$
$$s^2 = \frac{310.7402}{49} = 6.342 \text{ lakhs squared}$$

How do I understand lakhs squared?

Take the square root so that the unit of measurement is the same

$$s = \sqrt{s^2} = \sqrt{195.85} = 13.99 \text{ marks}$$
$$s = \sqrt{6.342} = 2.518 \text{ lakhs}$$


Now, the term variance is a squared measure and in the case of the marks example when we looked at 50, marks of 50 students the variance has a unit called marks squared. When we calculated the variance of the salary, the salary variance has a unit of rupees squared or lakhs squared as the case may be, so variance does have a unit. Standard deviation which is a positive square root of the variance in the marks case would still be marks.

And in the salary case would still be either rupees in lakhs or rupees as the case may be. So, both variance and standard deviation have a unit, as much as mean median and mode also have units. So how do I understand lakhs squared, when I say the variance is lakh squared. So, it becomes slightly difficult to understand the unit of variance as lakhs squared; therefore, the standard deviation comes, take the square root so that the unit of measurement is the same and therefore, we get, for the marks we get 13.99 marks and so on and 2.518 lakhs as the case may be.

(Refer Slide Time: 31:55)

#### Role of standard deviation

100 chocolate balls were weighed and the mean weight was found to be 2.54 grams. The standard deviation = 0.022  
How many pieces are in a 50 gram packet?

Mean = 2.502, number =  $50/2.54 = 19.98 = 20$ .  
Due to standard deviation some packets may either weigh less than 50 g if you put 20 chocolates or we have to pack more than 20 to take care of variation



Reduce process variation. Introduces to the concept called 6 sigma

Now, how do we understand the role of the standard deviation? Now let us look at this simple example, let us assume we take some small chocolates and these were weighed and the mean weight was 2.5 grams, and the standard deviation was found to be 0.022. So, let us try to find out how many pieces are there in a 50 gram packet.

Now, since the mean is 2.502, the number of small chocolate or chocolate balls in a packet would be  $50 \div 2.54$  which would be 19.98 which would be 20. So, due to standard deviations some packets may have a weight slightly less than 50 and some of them may in. In some instances you may have to put the 21st piece to make it slightly more than 50, so that the actual weight of a 50 gram packet would be very close to 50, but it could be on either side, if we actually measure it extremely accurately.

So, there are two aspects to it, the standard deviation creates a situation, where the actual weight can need not exactly be 50, but could be a small number lower or higher than 50, which shows that there is a variation, which also leads to a very important phenomenon that we have to reduce this process variation. And this leads to a very important concept called 6 sigma in manufacturing, where we concentrate a lot on reducing the process variation. So, standard deviation represents a method of dispersion and also helps us to understand that there is variation and there can be inherent variability and so on.


(Refer Slide Time: 33:41)

Role of standard deviation – Calculating Risk

Year	Stock A	Stock B
1	10.8	9
2	12	14.2
3	13	16
4	12	8.3
5	12.2	12.5
Average	12	12
Std Dev	0.787	3.308

Mean = 12 for both the shares. Share B has a higher standard deviation than share A. It has higher risk.

Variance (or standard deviation) is a measure of risk



What happens when the averages are different?

Now, we will look at one more aspect of standard deviation in this lecture. So, standard deviation also helps us in calculating risk. Now let us look at an example that is shown here, we have two stocks; stock A and stock B and we have 5 years data of the returns on stock A and stock B.

And we observe that the average or the mean return is 12 percent let us say, for both stock A and stock B. But if we find the standard deviation, we realize that stock A has a lower standard deviation than stock B, which has a standard deviation of 3.308 while standard deviation of stock A is 0.787. So, mean is the same for both, but share B has a higher standard deviation than share A or stock A.

Now, stock B has higher risk. So, variance or standard deviation is a measure of risk, and we have to ask a question what happens when the averages are different. Now in this case the averages were the same and therefore, we said stock B has a higher amount of risk. Now what happens when the averages are different, then we introduce a third measure called coefficient of variation, using which we try to calculate which one is better.

Now we would see that aspect in the next lecture.