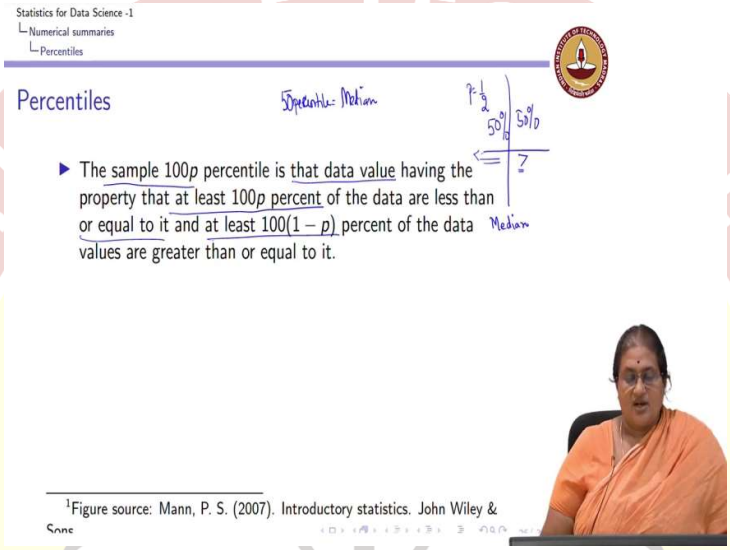
IIT Madras

ONLINE DEGREE

Statistics for Data Science - 1
Prof. Usha Mohan
Department of Management Studies
Indian Institute of Technology, Madras

Lecture – 3.5
Describing Numerical Data – Percentiles, Quartiles, and Interquartile range

(Refer Slide Time: 00:14)

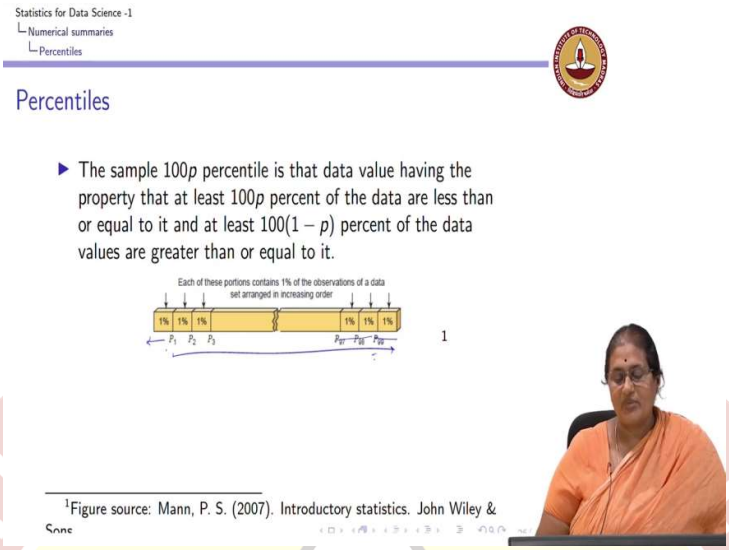


The next important thing which we are going to discuss is the notion of what we call a Percentile. These days with a lot of competitive examinations, most of these course and competitive examinations are reported as percentiles. Percentile is different from percentage.

What is a percentile? The sample 100 percent percentile is that data value that has the property that at least 100p percent of the data are less than or equal to it and at least 100(1-p) percent of the data are greater than or equal to it.

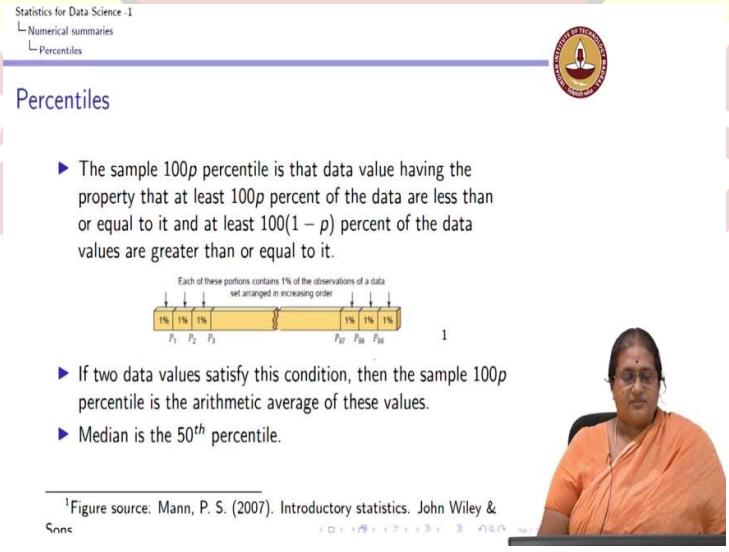
So, now if p = 1/2, 100p = 50% of the data, (1 – p) = 1/2, 100(1 – p) = 50% of the data and we have already seen a measure which says that 50 percent of the data is less than the value and 50 percent of the data is greater than the value and we call this as the median of a dataset. So, the 100 * ½ or the 50th percentile is the median of the dataset.

(Refer Slide Time: 01:59)



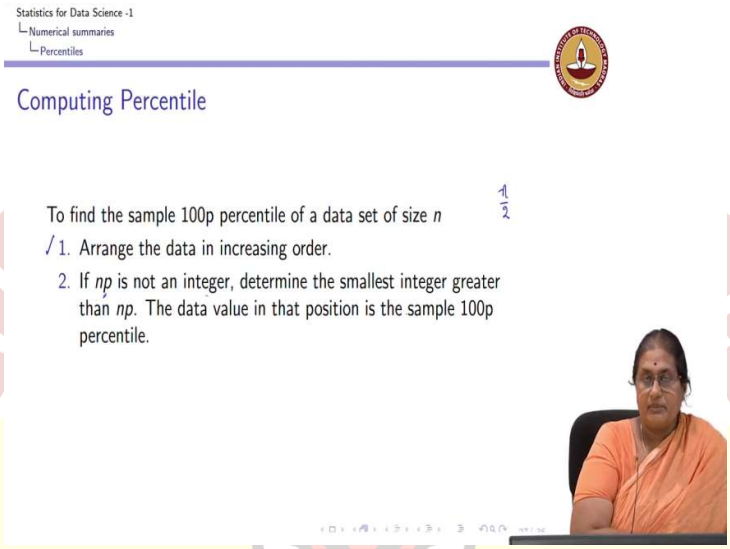
So, to demonstrate it using a figure, you can see that 99th percentile would have 100 * 0.99 that is 99% of the data is less than it, but 1% is greater than it. Similarly, P_1 says 1% is less than it whereas, 99% is greater than or equal to it. So, the concept of the percentile tells us that value in the dataset below which I have 100 * p which are less than or equal and 100 * (1 – p) which are greater than or equal to it .

(Refer Slide Time: 02:38)



Now, if two data values satisfy the same the condition, it is the arithmetic average of these values we have already seen how to compute the median when I have odd or even number of observations.

(Refer Slide Time: 02:55)

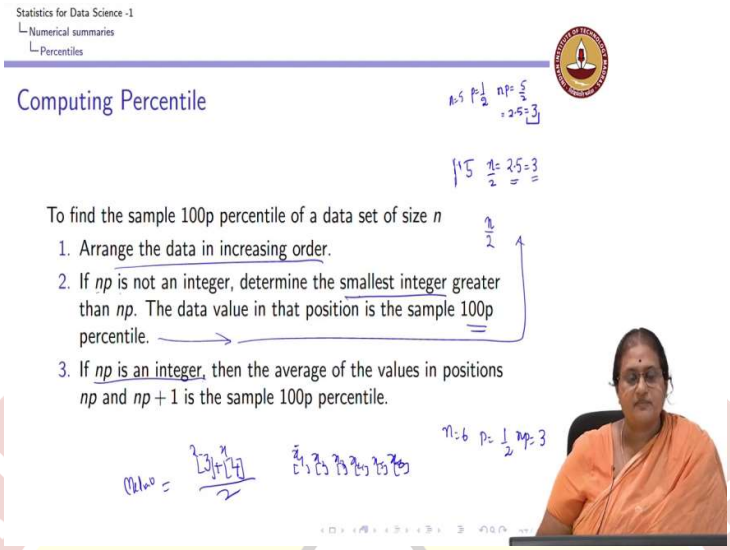


How do I compute a percentile? There are many algorithms to compute a percentile. I now present with a very simple algorithm to computer a percentile which is also very commonly used algorithm.

So, suppose I have a data of size n that is I have n observation, I want to find out what is my sample 100p percentile, what I do is I arrange the data in ascending order that is my first step similar to what we did when we computed a median. Once, I arrange my data in ascending order, I find out what n * p is.

So, now let me give you the analogy of what we did with respect to the median. So, I arrange the data in ascending order, this is something which I did for the median also, remember p = ½ so, I check what is n/2. If n is even, I know n/2 would be an integer. If n is odd, I know n/ 2 is not an integer. So, if np is not an integer, I determine the smallest integer greater than or equal to np that is the value, in that position is the sample 100p percentile.

(Refer Slide Time: 04:21)

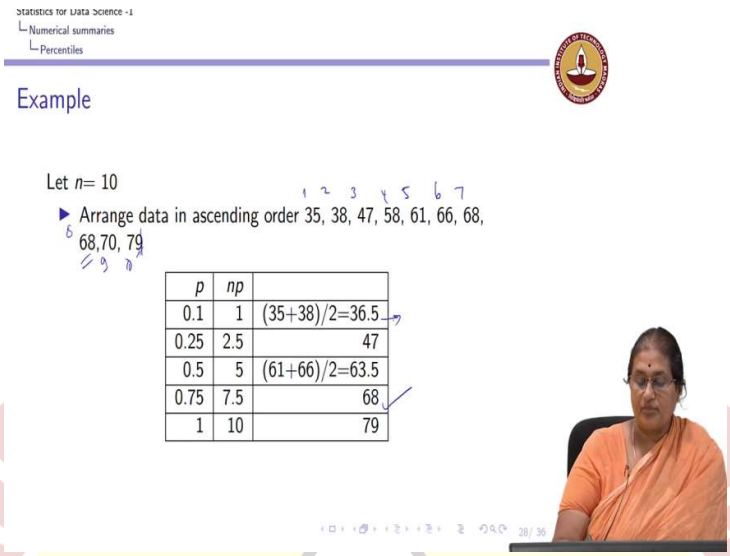


If np is an integer, then I look at the average of these values in positions np and np + 1. So, what did I say if n/np is not an integer for example, if I look at n / 2, suppose n = 5, my n / 2 is 2.5, the smallest integer greater than this is 3 so, you saw that in a dataset of 5 points, the third value would give me my median. So, that corresponds to my algorithm here.

I repeat, if np is not an integer, let n = 5, p =1/2, n * p = 5 / 2 which is 2.5 it is not an integer, the smallest integer greater than this is 3. So, the value or the data which is in the 3rd position because I am already arranging my data in ascending order will give me the sample 100 * p or the 50th percentile or the median.

But if np is an integer. So, now, let n = 6 and p = 1/2, my np is an integer which = 3, then what do I do? I look at the average of the value so, the 3rd + the 4th data value that is what we do and I have six data points x_1, x_2, x_3, x_4, x_5 and x_6 which are arranged in ascending order, then you know that $(x_3 + x_4) / 2$ is my median this is how we define. So, I do the same thing if np is an integer, I look at the average of values in positions np and np + 1.

(Refer Slide Time: 06:27)

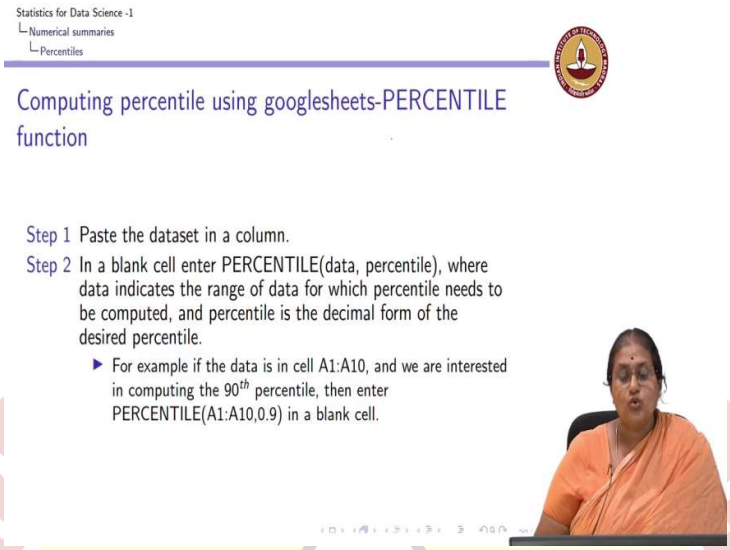


So, let us look at n = 10. I have my data. I again arrange my data in ascending order this is the same data we have been using. So, I have arranged my data, my n = 10. So, for p = 0.1, my n * p is 10 * 0.1 which is 1, it is an integer.

So, going back to your algorithm, I need to look at the value which is np and np + 1, the 1st value np is 35, np + 1 is 38. So, (35 + 38) / 2 is 36.5, this is my 100 so, you look at 0.25, 2.5 it is not an integer so, the 3rd value which is 47 that gives me the 25th percentile.

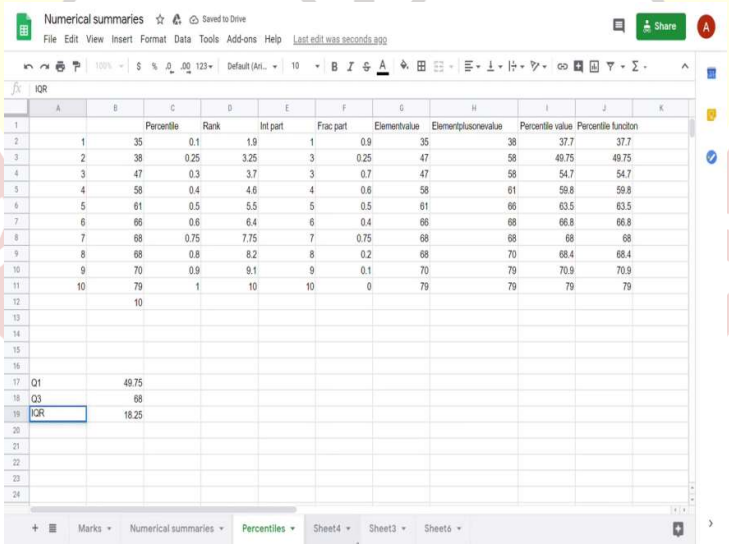
0.5 we have already seen 63.5, again 61 which is my 5th value 1, 2, 3, 4, 5 (61 + 66) / 2 that would give me 63.5. 0.75 is again a fraction so, you can see that when I have 0.75, 7.5 so, the 8th value so, this is the 1st, 2nd, 3rd, 4th, 5th, 6th, 7th and 8th, 9, 10, 8th value is 68, 68 is the 75th percentile, the 100th percentile is the maximum which is 79. So, this is how you compute percentiles and I the example shows how to compute percentile for the given dataset.

(Refer Slide Time: 08:14)



However, googlesheets uses a percentile function and a slightly different algorithm to compute percentiles. So, now we will demonstrate how to compute a percentile in a googlesheet.

(Refer Slide Time: 08:33)



So, let us go back to this google sheet. I have the following dataset the same dataset which has been given in ascending order. This is the dataset which I have written the dataset in ascending order.

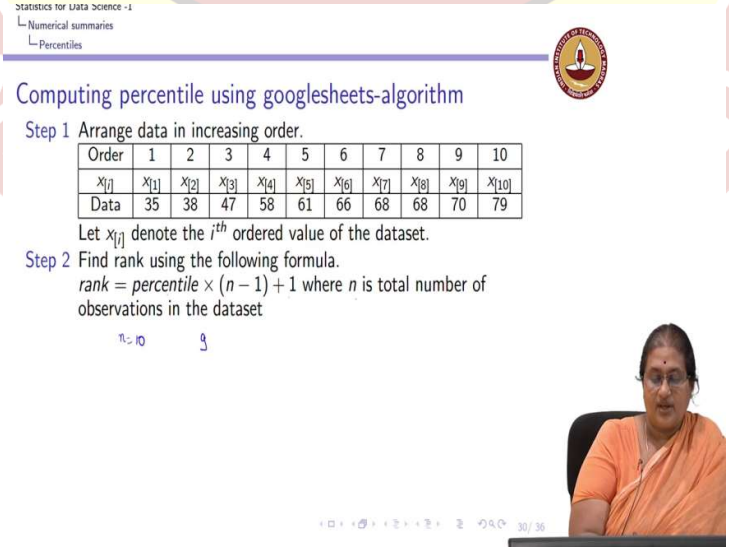
In a blank cell enter percentile and the data. So, I just go here, I type out what is the percentile. So, if I am looking at this dataset, I look at the percentile, so, B2 to B11 that is my dataset with C2, C2 is 0.1 so, this gives me what is a percentile so, for example, if the data we are interested in 90th percentile, I put a 0.9, if I am interested in the 10th percentile, I put a 0.1. So, you can see that 37.7, the 90th percentile I am putting it, here is 70.9, the 100th percentile is 79. So, you can see this is how we have computed or the google sheet computes it.

Again I repeat, the way google sheet you do is choose the data that is B2 to B11 that is the data which I want to compute the percentile for and C2, C2 is 0.1, 0.25 will give me the 25th percentile this gives me the 10th percentile, 54.7 will give me the 30th percentile, 0.9 gives me 70.9 is the 90th percentile and 79 is the 100th percentile.

Notice that the percentiles need not be part of the dataset. What you will notice immediately is these percentiles which google sheet percentile function gives us is different from the for the same dataset it is different.

Here I got my 10th percentile to be 36.5 whereas, google sheet gives me 37.7. So, the algorithm that google sheet uses, it is not that this is wrong and that is right, but the algorithms use as I mentioned earlier the algorithms used are different.

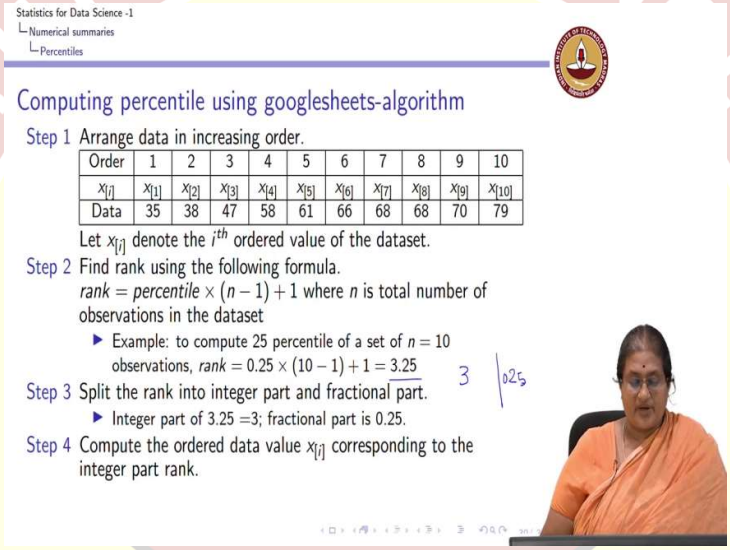
(Refer Slide Time: 11:04)



So, what is the algorithm a google sheet uses to compute percentile? So, first it arranges the data in ascending order. So, I have the data these are x_1, 2, x_3 these are the ranks the data is arranged in ascending order that is what we have done here the data is arranged in ascending order.

The second step is for any observation, x_i is denoting the ith order value, find the rank using the following formula. So, the rank is percentile * (n – 1) + 1 where n is the total number of observations.

(Refer Slide Time: 11:54)



So, let us look at it n in this case is 10, 10 - 1 is 9 ok, percentile *(n – 1). So, for example, if I want the 25th percentile, I put 0.25 * 10 - 1 so, I have 0.25 * 9 + 1. I get a 3.25, is it clear.

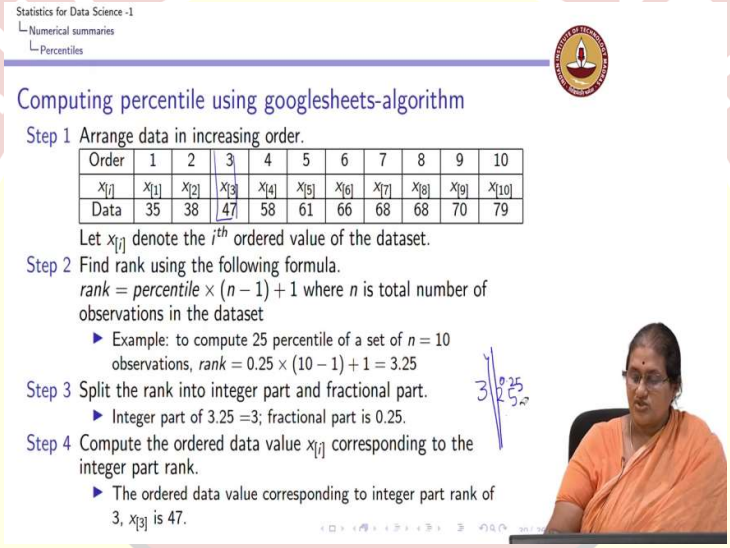
So, you can see that here what I have done is the computed the rank using that same formula. So, what is this formula? It is C2, C2 is your percentile *(n – 1) which is my 9 + 1. So, this is the data. So, percentile into so, I am computing the rank of each of these datasets. So, the rank here is 1.9, 3.25 we have already demonstrated how I got this 3.25.

For each of these ranks, I have an integer part and I have a fractional part. If the rank is 1.9, the integer part is 1, the fractional part is 3.9. For 3.25, the integer part is 3, the fractional part is 0.25 ok.

So, for each one of them, I have an integer part and a fractional part. So, in the 3rd step, I split it into an integer and fractional part and that is what we have done here. For the rank, I have split it into an integer part and the fractional part. For every rank, first we compute the rank, then I split the rank into integer and fractional part.

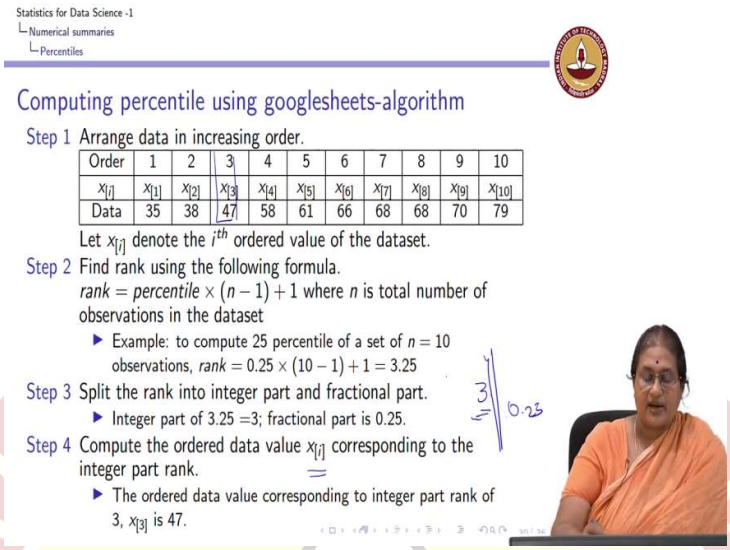
Once that is done, I look at what is the I compute the ordered data value corresponding to the integer part rank. So, what do I mean by this? For 3.25, the integer part is 3, the fractional part is 0.25.

(Refer Slide Time: 14:05)



Compute the ordered data value corresponding to the integer part rank. The integer part rank is 3 and you can see that the ordered data corresponding to this integer part rank is 47.

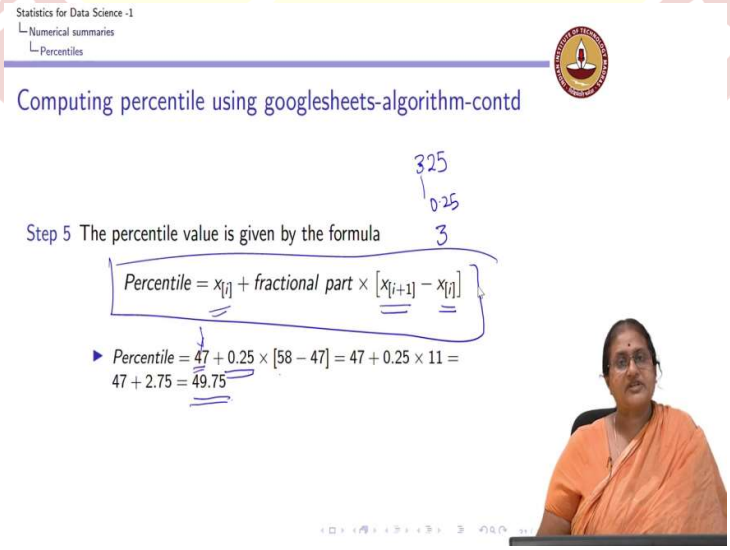
(Refer Slide Time: 14:26)



Again, for this 25th percentile 3.25, integer part is 3, the fractional part is 0.25. So, what I do is I look at computing or I look at that value which is corresponding to the integer part rank which is 47 that is what we have done here ok.

So, I have this element value. So, 1, 35 the integer value corresponding to 1 is 35, corresponding to 3 is 47, 4 is 58, 5 is 61, 6 is 66, 7 is 68 again here, I have the integer part which is 8, again it is 68, with 9 it is 70 and with 10 it is 79. This is the value which corresponds to the integer. So, 1 you can see its 35, 3 it is 47 and so forth.

(Refer Slide Time: 15:22)

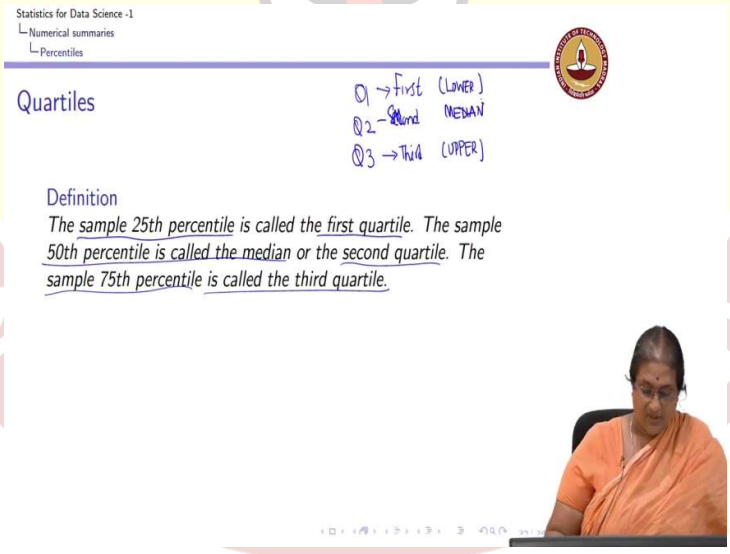


Once that is done, in the 5th step you find out what is the percentile value. The percentile value is the x_i which is 47 + the fractional part, remember the part 25th percentile was 3.25, the fractional part was 0.25, the integer part was 3, the value corresponding or the ordered corresponding to x_3 was 47 I take the fractional point which is 0.25 and I look at x_i + 1. What is x_i + 1? It is 58.

So, I look at x_i + 1 which is 58, - x_i which is - 47. So, 47 + 0.25 * [58 – 47] is 49.75 and that is what I have here. In the first thing, I have again element value is 35, element value + 1 is 38 which is x_2, x_2 - x_1 is what I have is 2, I multiply that with my fractional part that is what you do here which is 0.9. I add that and I get a percentile function which is 37.7. So, I have 33 * 0.9 which is 2.7, 35 + 2.7 is 37.7.

So, you can see that this column I gives us the percentile value computed using this equation whereas, column J gives the percentile value computing the percentile function of the google sheets and we can see that both of them are exactly the same.

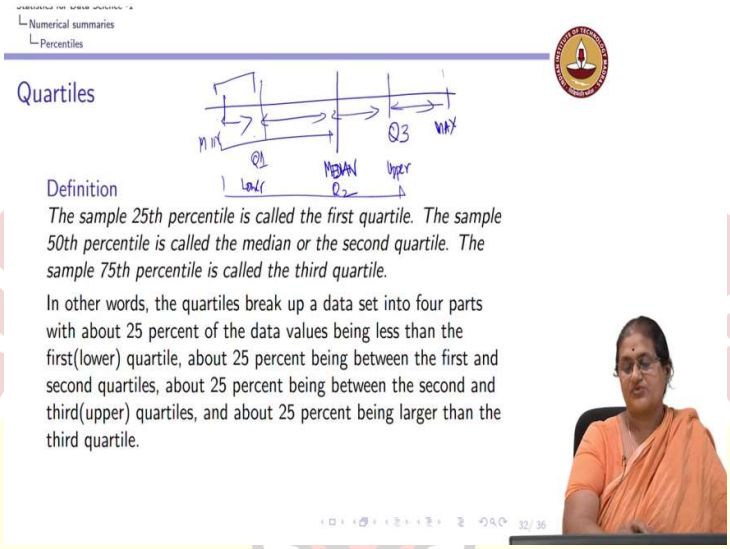
(Refer Slide Time: 17:31)



So, why are percentiles mentioned when we talk about the measures of dispersion? As I earlier mentioned, a very important measure is called the quartile. The sample 25th percentile is called the first quartile, the 50th percentile we already know is the median or the second quartile and the sample 75th percentile is called the third quartile.

So, I have Q1, Q2 and Q3 this is referred to as the first quartile or in some books as the lower quartile, this is the third quartile or referred to as the upper quartile, this is the median or the second quartile which is already we have seen is the median ok.

(Refer Slide Time: 18:22)

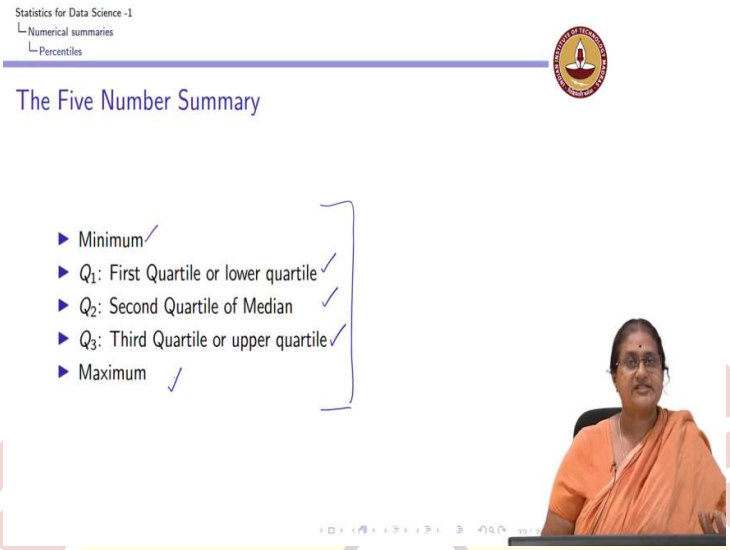


Now, when I have the quartiles, the quartiles in fact, break up a dataset into four parts. So, if I have this as my dataset, I know already the median breaks up a dataset into two parts, the first quartile.

The first quartile or the lower quartile I have a upper quartile this is Q2, this breaks up so, I have the least, this is the minimum, I have the maximum, I have the first quartile, the second quartile, the third quartile also referred to as a lower quartile and the upper quartile ok. It breaks up the dataset so, I have part 1, part 2, part 3 and part 4.

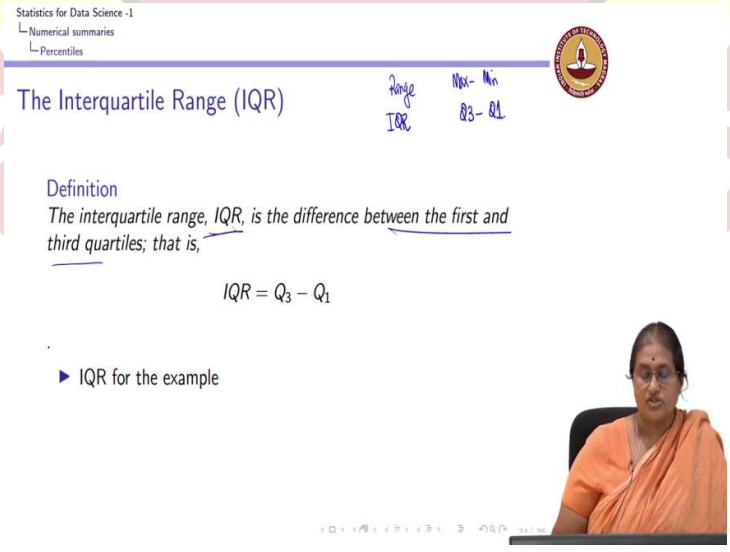
So, you can see that the quartiles break up an entire dataset into four parts. 25% of the data lie here, 50% lie here, 75% lie here and entire dataset lies between minimum and maximum.

(Refer Slide Time: 19:34)



So, why are these again an important measure? If you look at it any descriptive statistics most of the summaries are given / what we refer to as a five-number summary. The five-number summary includes the minimum, the first quartile, the median, the third quartile and the maximum as we have given here. So, this five-number summary is a very good way of summarizing a dataset.

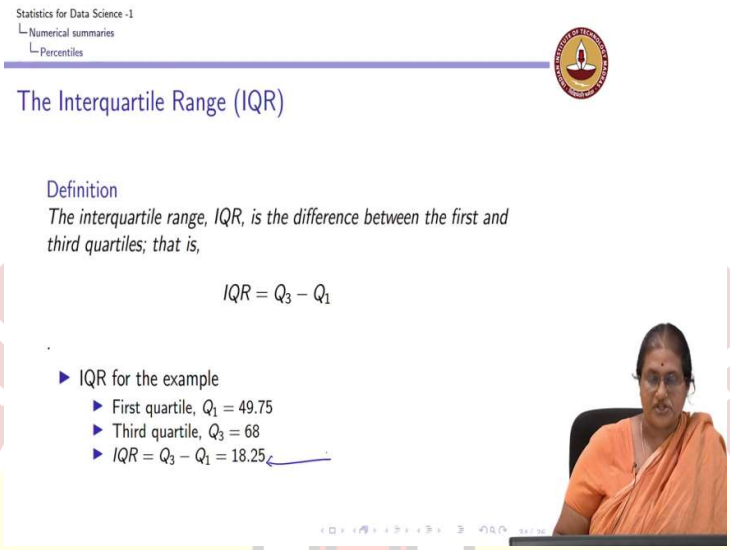
(Refer Slide Time: 20:05)



The interquartile range is a very important measure of dispersion which you have we have already seen the range is the difference between maximum and minimum. The

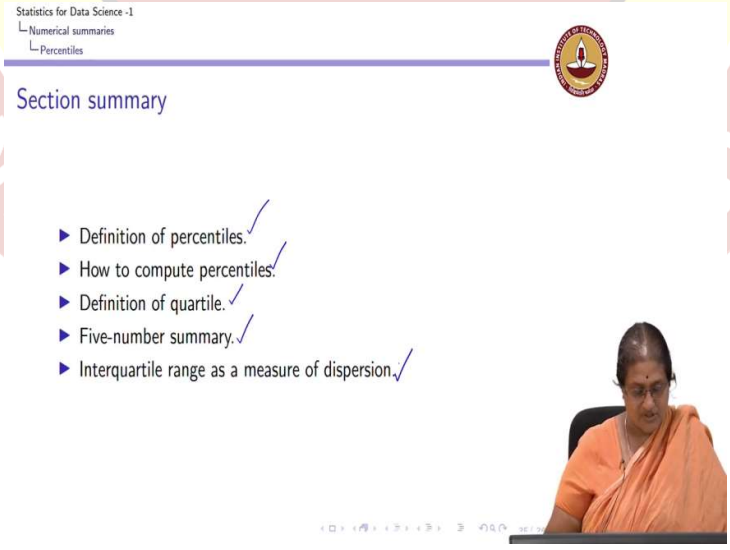
interquartile range is the difference between the third quartile and the first quartile and this is referred to as the interquartile range.

(Refer Slide Time: 20:37)



So, for our example, the first quartile was 49.75, the third quartile was 68, the interquartile range is 18.25. So, the interquartile range is also a measure of dispersion.

(Refer Slide Time: 20:55)

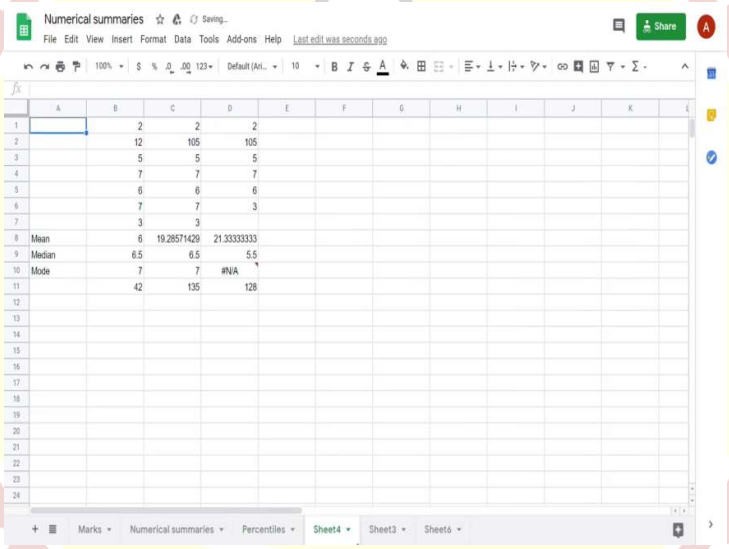


So, what we have seen so far was how to define percentiles, how to compute percentile, what is the definition of a quartile, the five-number summary which is very very

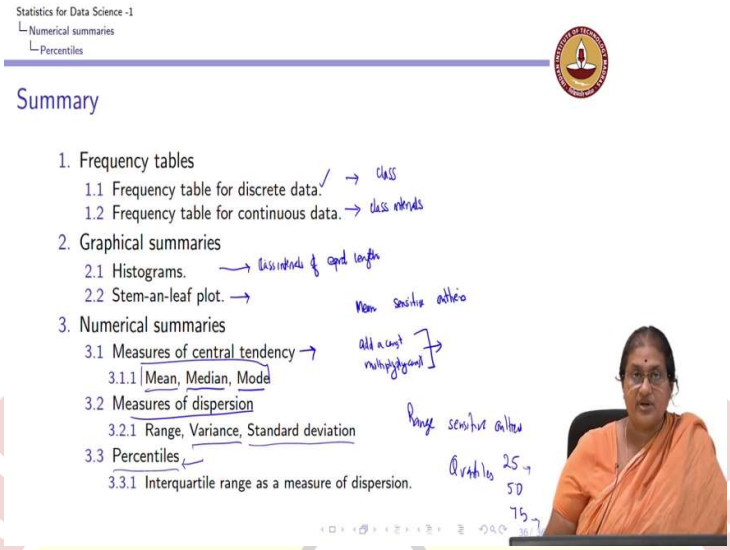
important in many descriptive statistics and interquartile range which is a measure of dispersion.

The interquartile range so, this if you look at googlesheet, the quartile function with the array and 1 the lower quartile is the first quartile gives me the lower quartile. The quartile function with 3 gives me the upper quartile and the difference between the upper quartile and the lower quartile gives me the interquartile range and we can see that the quartile function if I put 2 gives me the median which you have already seen is 63.5, 3 gives me the upper quartile. So, this is my Q1, this is my Q3 and this is my IQR or my interquartile range.

(Refer Slide Time: 22:08)



(Refer Slide Time: 22:10)



So, a summary of the this module where we summarize numerical data. What you should be knowing at the end of this module is we first started by looking at frequency tables, we looked at discrete data where each data point was considered as a category or a class, then we looked at group data, we defined what were class intervals, we defined what was lower class limit, upper class limit and we saw how to construct the frequency tables for both discrete and grouped data.

We followed it with histograms. Again, here we assumed class intervals were of equal length and we saw how to construct histograms using google sheets. The stem-and-leaf plot, we also demonstrated how to come up with a stem-and-leaf plot for an example data.

One then, we moved on to numerical summaries. We started with the measure of central tendency although we have looked at mode and median in the earlier categorical data module, but then now we introduce a very important measure called the mean.

What we observed was mean was very sensitive to outliers is something which we saw and then, we saw what would happen to each of these measures of central tendency if you add a constant and if you multiply with the constant. The reason is it is always helps us to know what would happen to the measures whenever we manipulate the data and the way we manipulated the data here was to add a same value or multiply it with the same value.

Then, we moved on to look at measures of variability or dispersion or spread. We started with the measure range, again we saw that the range is extremely sensitive to the outliers we showed this through illustrated it through an example, then we talked about variance.

We also defined the population variance and the sample variance, but we stuck on to the sample variance. At this point of time, we said that all these measures take the same units of your original data whereas, variance takes the units of squared units of the original data hence, we define a measure which is standard deviation which is square root of the variance which takes the units of your original data.

Then, we went on to percentiles. When we discussed about percentile, we introduced what were percentiles, then we introduced important percentiles namely the 25th percentile, the 50th percentile and the 75th percentile.

We call this the first quartile or the lower quartile, the 75th percentile is the 3rd quartile or the upper quartile and the 50th percentile is what we already have seen as the median. The difference between the third and the first quartile is what we refer to as the interquartile range and we see that interquartile range is a measure of dispersion.

So, with this, we come to the end of the discussion on how to come up with numerical summaries for a single variable. What we are going to see next is the association between two variables. So, far we have looked both in the categorical case and the numerical case, measures of graphical summaries and numerical summaries of a single variable.

What would happen when I have more than one variable. So, we look at having two variables, we start with having two categorical variables look at how we, look at the association between two categorical variables, we look at association between 2 numerical variables and then, we looked at association between a categorical and a numerical variable. So, this is what we are going to do next.

Thank you.