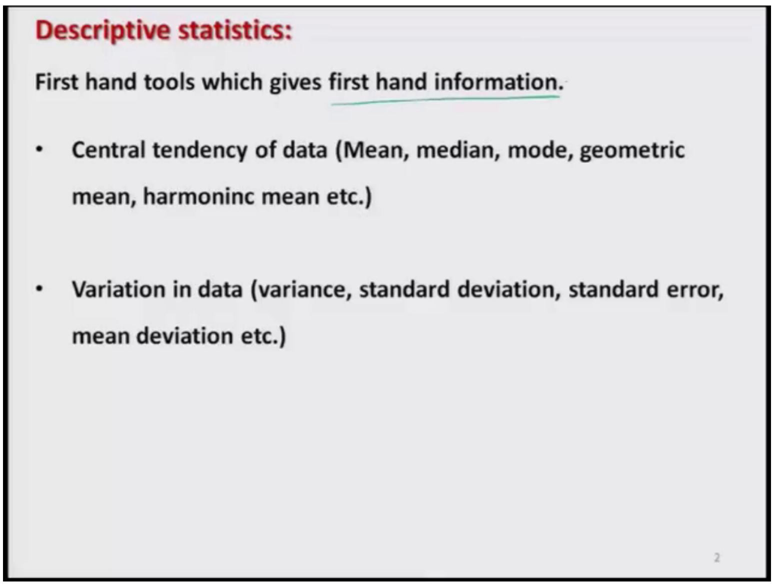
Introduction to R Software
Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture - 37
Central Tendency and Variation

Welcome to that next lecture in the course Introduction to R Software. You may kindly recall that in the last lecture we had discussed some graphical tools to represent the characteristics of a data, that a hidden inside it and with the graphical tools we are trying to take them out for example, we had discussed different types of graphics like a histogram bar plot and so on. Now we are moving towards analytical tool and in this lecture we are going to discuss the tools to measure the central tendency and variation of data. And you already have done most of the things in your earlier classes like as mean median variance standard deviation and so on.

So, essentially we are going to briefly discuss these concepts. And we will try to see how they can be computed using the R software. So, we start our lecture over here and as we had discussed in the earlier lecture whenever we get a data the first hand information is given by 2 types of tool first are analytical tool and second are graphical tool.

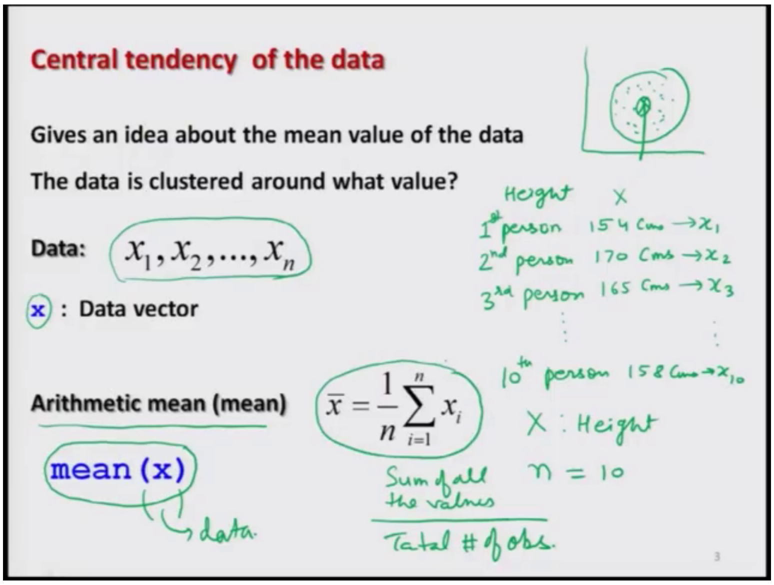
(Refer Slide Time: 01:22)



So, now under the analytical tool there can be different characteristics and here we are going to consider the central tendency of data and the variation in data. There are different tools which are used to measure the central tendency of the data like mean, median, mode, geometric mean, harmonic mean etcetera. And similarly when we want to measure the variation in the data there are different tools like variance standard deviation standard error mean deviation absolute deviation etcetera.

So, first we are going to concentrate upon the different measure for the central tendency. First question comes what is this central tendency and what do we really mean by this. So, whenever we get a data we would like to see; what is the value around which they are concentrated. For example, suppose you get some values and if you try to plot them and suppose this looks like this.

(Refer Slide Time: 02:30)



So, one can see here that this is the point here in the center around which all the values are concentrated. So, this is trying to indicate the measure of a central tendency right. So, there are different measures of central tendency. So, we discuss them one by one. So, first of all let me denote the data. This is here given by x 1 x 2 x n. What do we really mean by this x 1 x 2 x n?

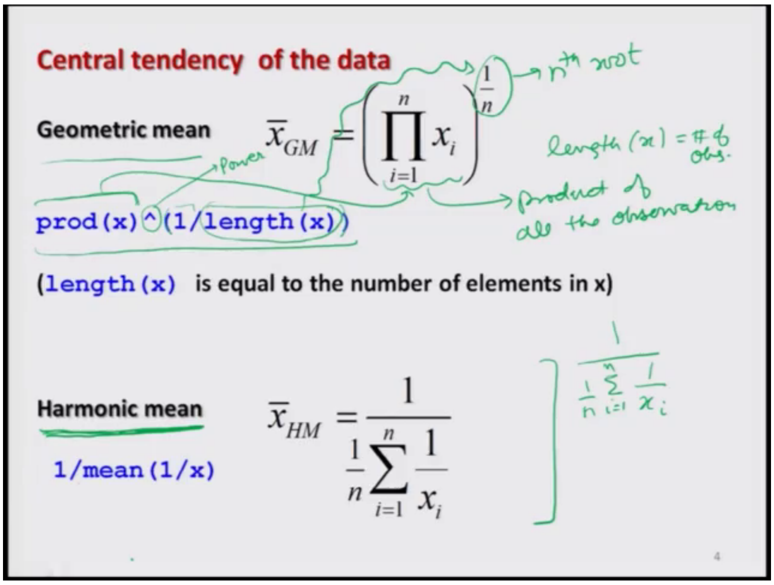
So, let me take a simple example and try to understand what do we really mean by this thing and the same concept will continue in the entire lecture. Suppose I want to measure the height of some persons. And suppose I measure the height of first person and this comes out to be suppose 154 centimeters and then I measure the height of second person this comes out to be 170 centimeters. Similarly, I consider the third person and this height comes out to be 165 centimeters. And similarly I continue and suppose I consider here suppose tenth person and suppose the height of tenth person comes out to be 158 centimeters.

So, this is indicating that we have considered here a variable x whose name is height, and then we are trying to obtain the data on this variable height. So, now, here is my variable x and on which when I try to obtain the first value of the data this is denoted here is a x 1. When I try to get the second data this is denoted as x 2. And when I get the third data this is denoted as here x 3. And finally, I will have here all together 10 values and the last value is the x 10. So, in this case you can see here n is equal to 10. So, this x 1 x 2 x n they all simply indicating some numerical values or there may be some characteristic values also for example, we have taken earlier data set in which we have found the gender of 10 persons and that was indicated as say male and female and that was categorized by 1 and 2 also. So, these are some values.

So, now here I am going to denote by this vector x the data vector. So, whatever data I have collected whatever the values I have corrected they are contained in this data vector x. Now first we try to talk about arithmetic mean. Arithmetic mean is simply sum of all the values divided by total number of observations. And this can be directly computed in the R software using the command mean of x, that is m e a n and inside the argument you have to write the data vector. So, you do not need to write a program to compute this quantity; that means, where you have to first find out the sum. And then you have to divide it by the number of observations R has already developed this program they have compiled all the program inside a function called as mean.

So, by using the function mean I can directly compute the arithmetic mean.

(Refer Slide Time: 06:53)

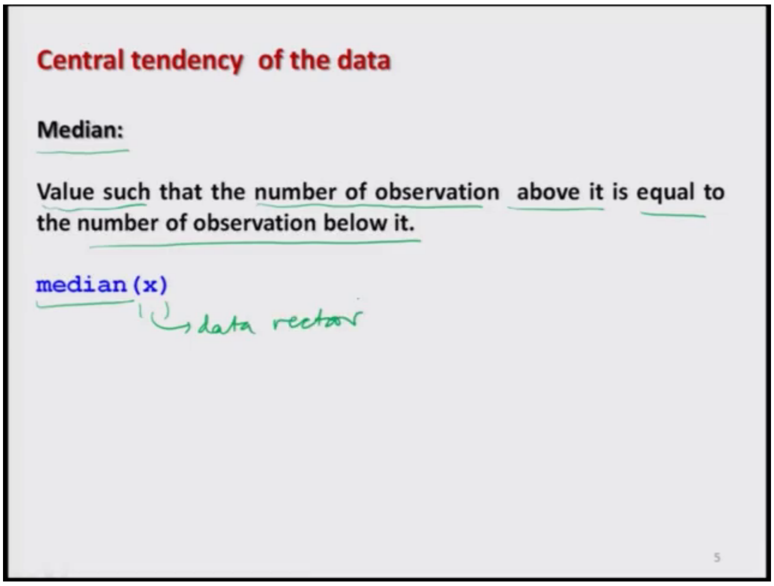


So, now next measure we are going to consider is the geometric mean. We all know that this is defined as the product of all the observations, and then we try to take it is n th root. This is how we compute the geometric mean of the data x 1 x 2 x n. In R package there is no direct function to compute the geometric mean just like arithmetic mean. So, what I want to show you here that this can be computed directly using some other functions. For example, if you try to see here this is the syntax of the geometric mean. First I am going to compute the product. We have learnt in the earlier lecture that this product of x 1 x 2 x n that is x 1 into x 2 into x 3 up to x n this can be obtained by the function p R o d that is a short form of product. So, this part of the syntax is going to give us the product which is this one.

And then this hat as we know that is going to indicate the power; that means, the raise to the power of something. And then I am writing here one upon length of x. We know that length of a vector is given by length of x. So, this is going to be simply the total number of observations what is here n. So, in case if I try to take it is power by using the symbol hat. Then inside the bracket I try to use here one upon length of x, which is denoting this quantity right. So, this is how one can compute the geometric mean. So, now, we are going to consider the harmonic mean. We all know that the harmonic mean is computed by this expression. And here you can see what is really happening. First of all, I am trying to find out the inverse of each and every observation xi and then I am trying to find out it is mean right. And after this I am going to find out it is inverse.

So, there is no direct formula to compute the harmonic mean, or there is no direct function to compute the harmonic mean r. So, what we try to do we use the mean function and we try to devise the syntax for computing the harmonic mean. So, what we try to do here, this becomes simply one upon mean of 1 by x. So, this is how one can compute the harmonic mean in the R software.

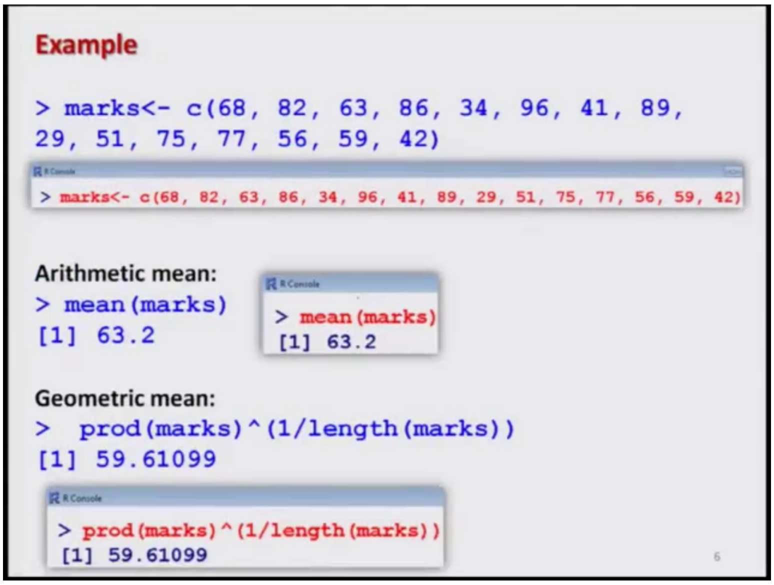
(Refer Slide Time: 10:15)



Next we try to consider another popular measure of central tendency which is median. We all know that median is the value which divides the entire frequency or the total frequency into 2 equal parts. So, this is a value such that the number of observation above it and the number of observation below it they are the same they are equal. And in order to compute the median, we have a direct function in R. And the syntax of this function is median and inside the argument this is the data vector.

So, you simply have to write m e d I a n all are in small letters and inside the argument you have to simply give the data vector.

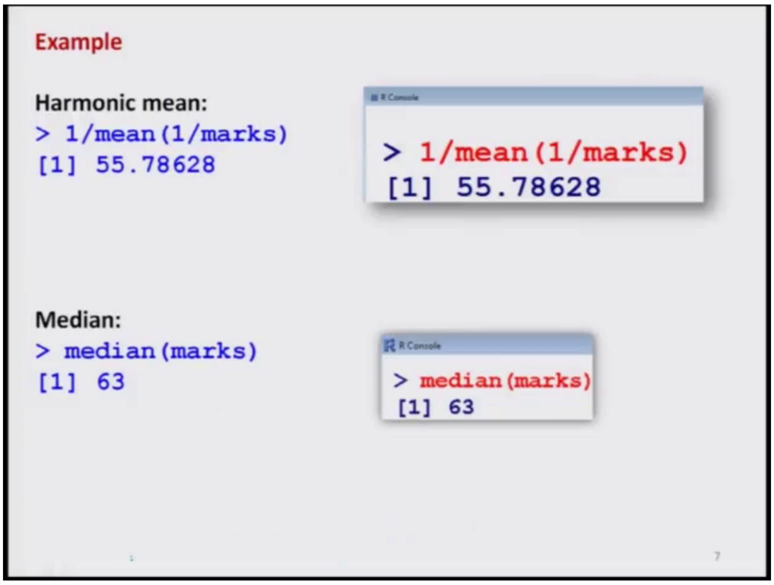
(Refer Slide Time: 11:05)



So, after having these simple functions one or 2 implement them over a small data set and see what really happening. So, now, here we consider a data set of 15 observations. And all these values are combined in a vector here c and these numbers are representing the marks of some students. So, I call it a vector marks and I try to store it in r. So, you can see here it is screenshot of this vector then in order to find out the mean that is arithmetic mean we simply try to use the command m e a n mean and inside the argument the data vector. So, this is here marks. So, you can see here is the value 63.2.

And similarly in order to compute the geometric mean, I try to use the same command that product of marks and then raise to the power one upon length of marks. And then I get here this value and here is the screenshot of the outcome.

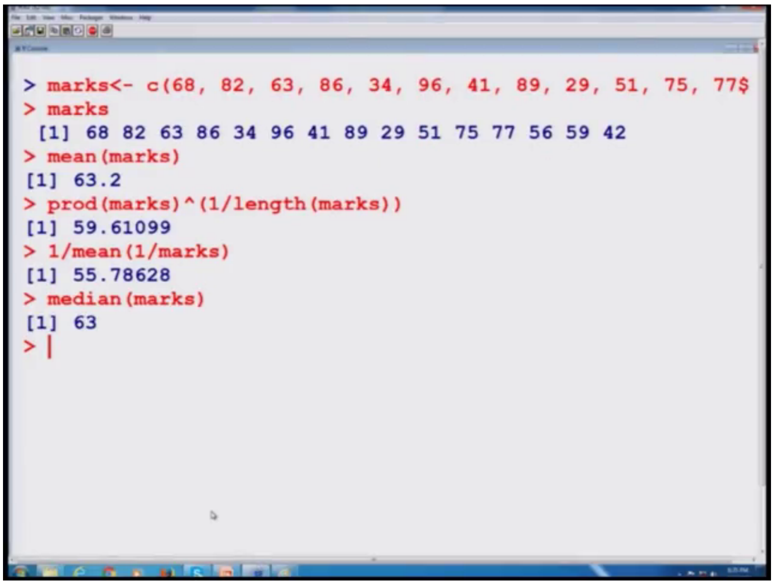
(Refer Slide Time: 12:16)



Next we compute the harmonic mean and you can see here this is the syntax. Marks is my data vector and I try to do here one upon mean of one upon marks and I get this value over here. And here is the screenshot of the outcome. And similarly if I want to find out the median of my data marks then this is m e d I a n median and inside the argument this is the beta vector marks and I directly get this value 63. And here is the screenshot, but one have to do this thing over the R consoles and try to see what do we obtain. So, first let me try to create here the data vector and you can see here this is here marks.

And now if I want to find out the mean of marks then this is 63.2. And similarly if I want to find out the geometric mean, then this is given by the syntax like this. And if I want to find out the harmonic mean this is given by the syntax this it is 55.78.

(Refer Slide Time: 13:37)

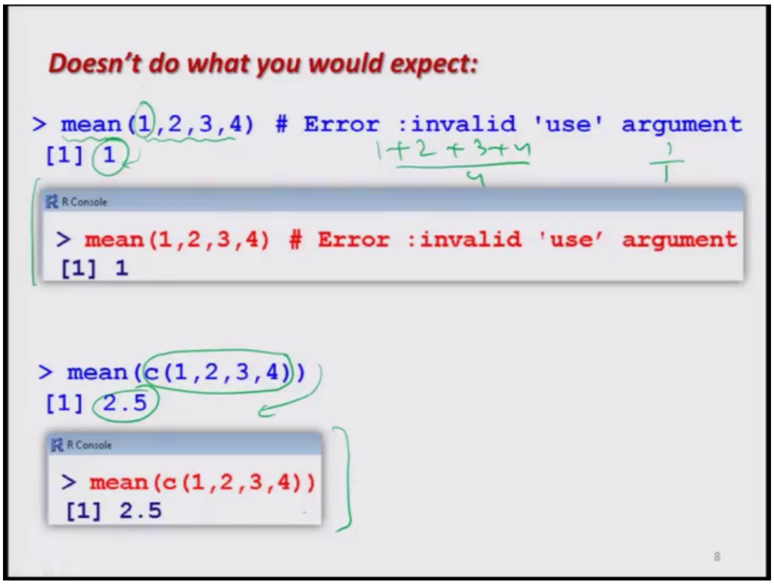


And now if I try to find out the median of these marks then this is given by the same command median of marks and this is 63. So, you can see that is not difficult to compute this these measures, and similarly on this lines you can compute other measures also which are available in the as a built in function in the base package of r. So, now, we come back to our slide and we move further.

Now in this slide I want to show you an example which many people commonly make. I show you this whenever I want to find out the mean I have to give the value in the format of c, that is a vector using the command c.

But suppose if I simply want to find out mean of 1 2 3 4.

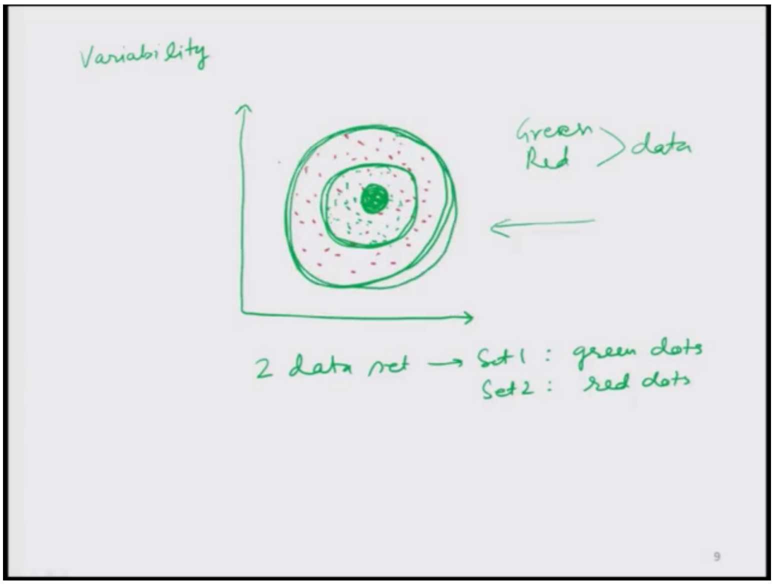
(Refer Slide Time: 14:34)



Then still you will get an answer one, but the mean of 1 2 3 4 is not one because this is 1 plus 2 plus 3 plus 4 divided by 4. So, the reason is that when you are not using here the c command combined function, then it is only considered the first element and it is trying to find out the mean of only the first element that is 1 upon 1 which is equal to here one right and the proper way to find out the mean of 1 2 3 4 is that please try to write down all the values inside the vector by using command c and then you try to find out the mean you will get the correct value and here is the screenshot. So, you can try it yourself, now we move to another aspect we have talked about mean now there is another aspect in the data analysis that is called as variability.

What is this variability first we need to understand?

(Refer Slide Time: 15:37)

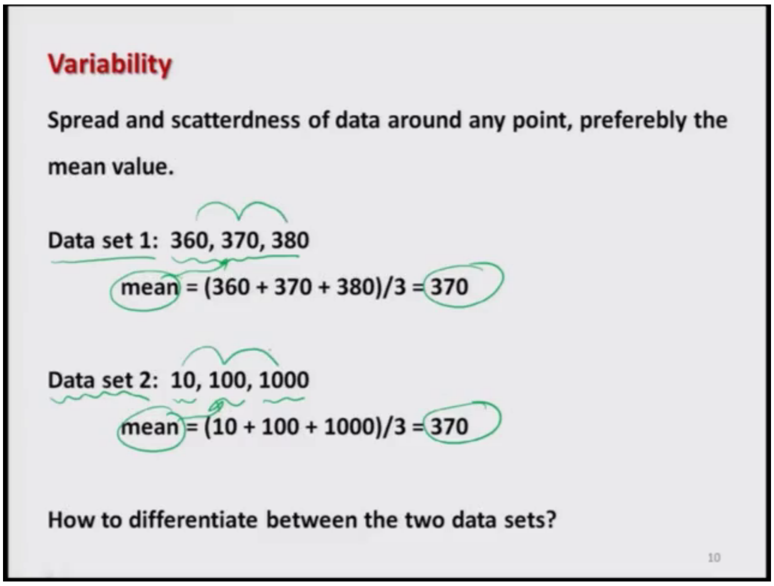


So, now if you try to see in this slide I have created a small picture. This picture is actually having 2 colors of dots. One is here something like green colored dots like this one and another are red colored dot. And this green colored dot and red colored dot they are representing some data. So, what is really happening that I have got here, 2 data sets. Set number one is suppose represented by green dots means I have plotted it and set number 2 values are denoted by red dots. So, now, when I try to plot it what do you observe here, you can see here that the measure of central tendency of green dot and red dot that is somewhere here, but the values which are denoted by green dots they are concentrated only in this area. And the values presented in the red dots they are concentrated in this area.

So, now if you try to find out the mean of both the data sets possibly they come out to be the same, but they have another property that the values in the green and red data set they are scattered around the mean which is here like this a big green dot, but their scatteredness is more. For example, all the values in the green dot they are scattered only inside this circle whereas, the values in the red dots they are contained in this bigger circle. So, this is the property which is called as scatteredness. And this is the property which is called as scatteredness of the data. And we would like to quantify it. And in order to quantify it we have different types of function and by those function we try to measure the variability right.

So, now I try to take some values in the 2 data sets.

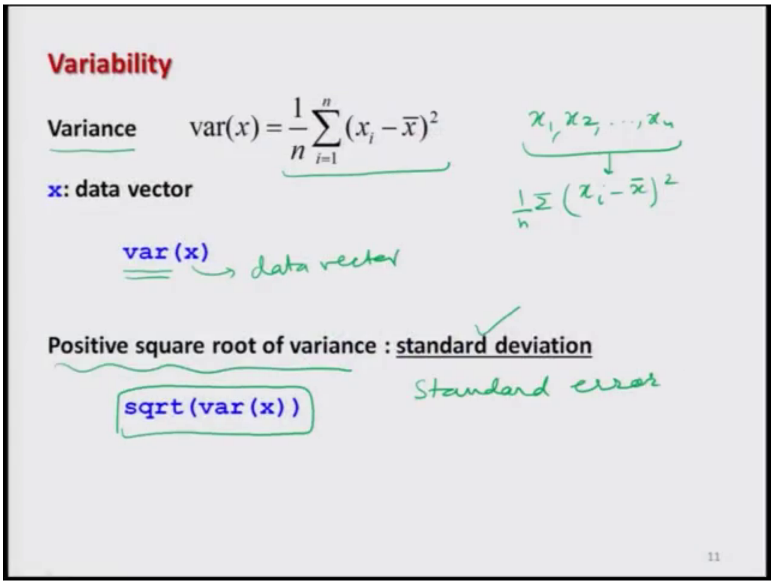
(Refer Slide Time: 18:02)



Suppose this is my data set one having 3 values 360, 370, 380, and second data set having 3 values 110 and 1000. You can see here when I try to find out the mean of these 3 values and mean of these 3 values this comes out to be the same 370. So, now, my question is how do I differentiate between the 2 data set. They have some properties which I want to quantify. And this is about the variability of the data. You can see here that 3 6370 and 380 they are very close in comparison to the 3 value 110 and 1000. So that is a property which we try to measure with the different measures of dispersion. And there are different measures of dispersion that we talked about, variance standard deviation mean deviation range interquartile range and so on.

So, now we are going to discuss those measures one by one, and we would also like to see what are the different syntax on our functions available in R package first quantity I am going to discuss is variance.

(Refer Slide Time: 19:19)

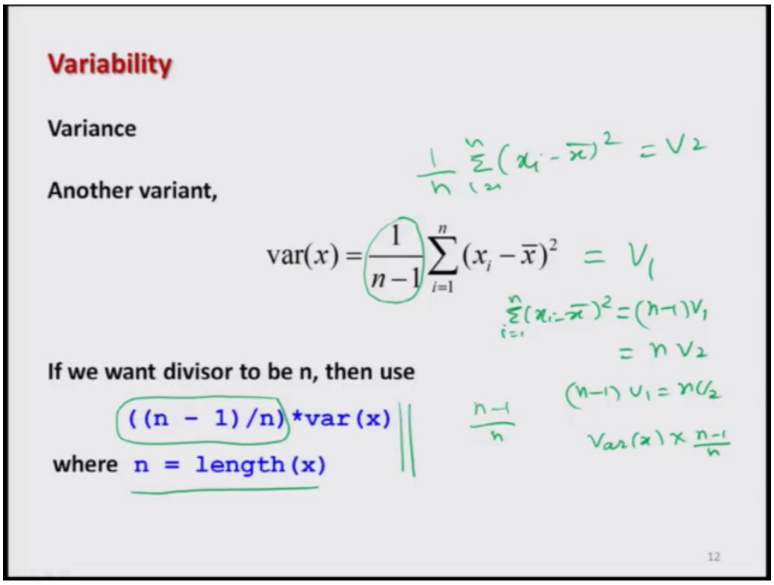


The variance of data set here x is again I have got here value x 1 x 2 x n. The variance of these values x 1 x 2 x n is defined by this. One upon n summation I goes from 1 to n xi minus x bar whole square. So, it is simply trying to say take each value x 1 x 2 x n then try to find out the difference from their arithmetic mean. And then try to square them then try to take the average of these squared values and this function is called as variance. So, in R this variance can be computed directly by the built in function v a r.

And the syntax is var then inside the argument we have to give the data vector. And when I try to take the positive square root of this variance this is called as standard deviation. In some places they talk about standard error also. Standard error and standard deviation have some differences from the conceptual point of view, but here I am not going into those details, but from the layman language, I would say that you try to treat a standard error and a standard deviation in a simple way. They are more or less like the same thing for us at this moment, but definitely I would repeat there is a difference they are not the same. In order to compute the standard deviation, we simply have to take the square root of the variance of x and the syntax is like this sqrt and then the square root of var and then the x data vector.

So, it is as simple as that finding out the square root of variance whatever you have found. One thing which we have to be cautious that in statistics there is another variant of variance and in this variant the divisor here is one upon n minus 1.

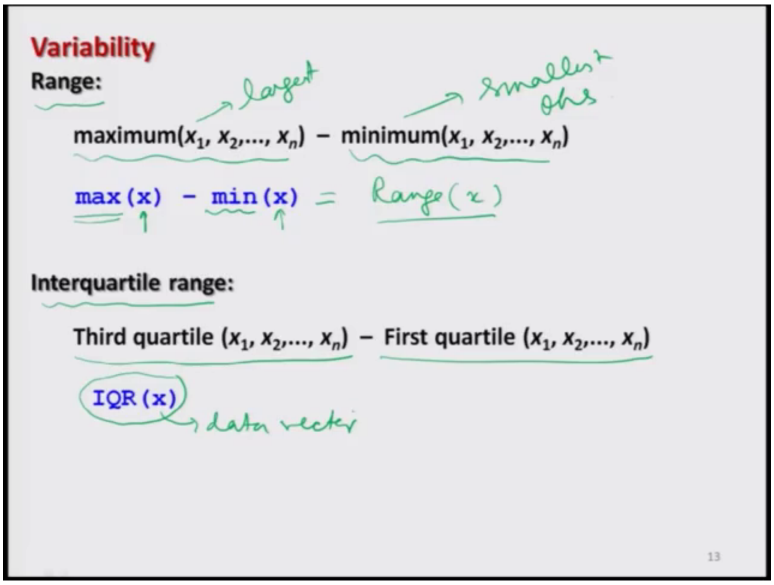
(Refer Slide Time: 21:26)



Whereas, in the earlier case this was one upon ni goes from 1 to n xi minus x bar whole square. So, in many situation we want to use this format. So, these formats if you want, then you simply have to multiply the variance of x by n minus 1 upon n. Why if you try to see here if I say this is my here see here v1 and this is here v 2. Then I know that summation xi minus x bar whole square i goes from 1 to n this is equal to here n minus 1 times v1 and this is same as n times v 2, right.

So, this transformation n minus 1 times v1 is equal to n times v 2 is used and we simply obtain that in case if I am trying to find out the variance of x, then this has to be multiplied only by a vector n minus 1 upon n. This is the length of x vector which is the total number of observation available in the x vector. So, this is how I can also compute this variant and we have to be careful whenever we are using a software. And we have to a certain that which of the form the software is using. Similarly, there is another variant what is called as range.

(Refer Slide Time: 22:55)

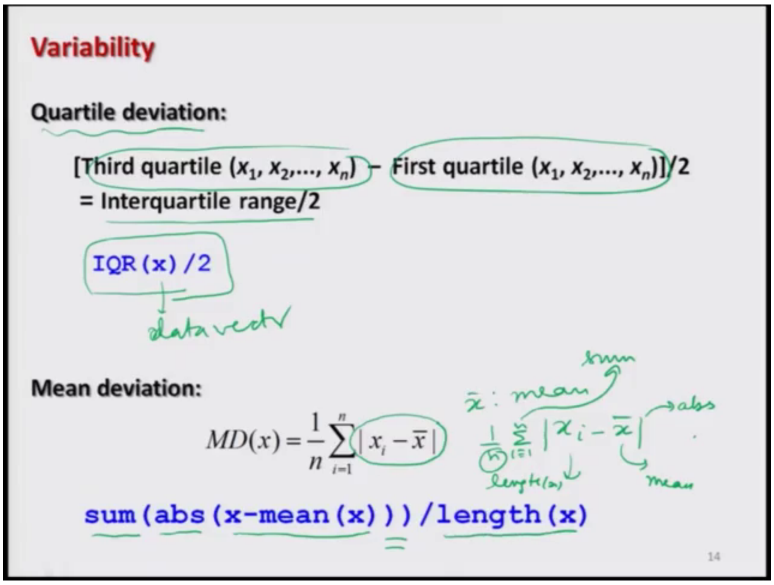


The range is defined as the difference of maximum value of the data and the minimum value among the data set; that means, I simply have to take the observation, and then I have to find which is the largest observation and which is the smallest observation and then I have to find their difference.

So, this range can be very easily computed by using the built in function max and min. So, m a x gives the maximum among the values which are contained in the data vector x and min gives the minimum value that is contained in the data vector x. So, this difference maximum of x minus minimum of x this is going to give us the value of the range for the data vector here x right. The next popular measure is interquartile range. The interquartile range is simply defined as the difference of third quartile and first quartile. We had discussed in the earlier lecture; what are quartiles means the entire frequency is divided into 4 equal parts: first quartile second quartile, third quartile, 4th quartile. Second quartile is the median.

So, I try to take the difference of third quartile minus first quartile. And this quantity can be directly computed in the R using the function IQR which means interquartile range of the data vector x.

(Refer Slide Time: 24:42)

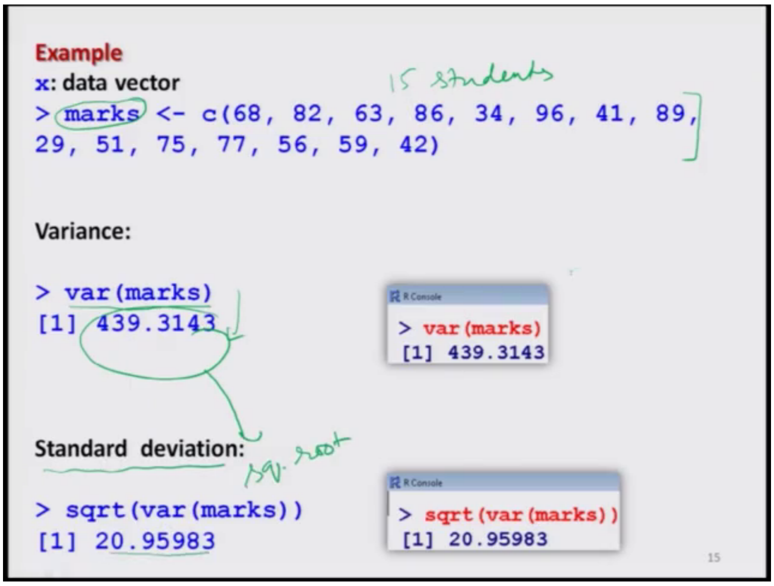


Next measure is quartile deviation. This quartile deviation is another variant of interquartile range in the sense that it is defined as the difference of third quartile of the data minus first quartile of the data divided by 2. So, this is essentially the half of the interquartile range. So, there is no need to have a separate function for computing the quartile deviation. Once we have a function for computing the interquartile range I simply have to make it half. So, the syntax to compute the quartile deviation is simply IQR which is interquartile range of the data vector x divided by 2.

So, that is pretty simple and straightforward. Similarly, the mean deviation is defined as the average of the absolute deviation from the arithmetic mean, x bar here is this here mean. And what we are trying to do we are trying to take every observation say ith observation. And I would like to take it is deviation from the arithmetic mean and instead of squaring it. I would like to take it is absolute value and then I will try to find out the arithmetic mean of all the absolute values this is called as mean deviation. And this mean division is least when it is measured around median, but here we are trying to compute it from there x bar. There is no direct function to estimate the mean deviation, but we have to just use the built in function and then we have to write down the syntax for computing the mean deviation.

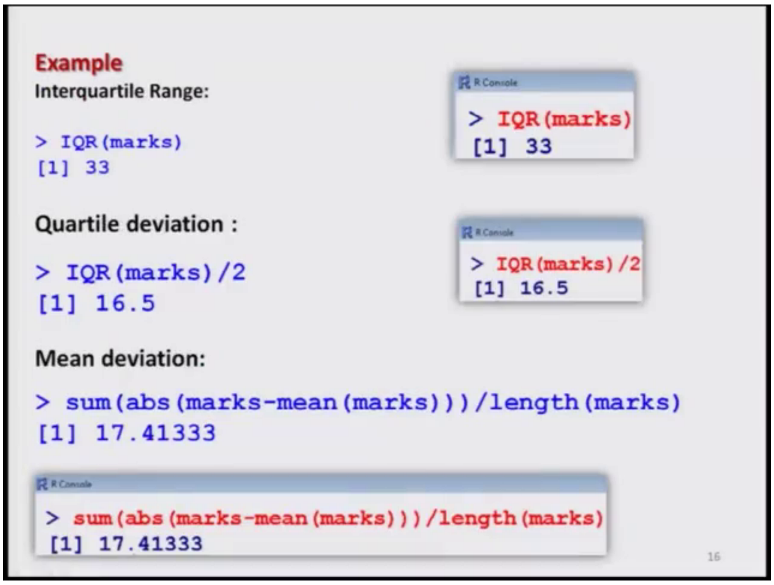
So, now if you try to look at this structure, it is pretty simple. We know how to compute the arithmetic mean that is mean of x. And we also know how to compute the absolute value the function is abs. And we also know what is hear n this is length of x and then we also know how to find out the sum. So, I simply have to write the syntax here sum of absolute value of x minus mean of x. Just be careful about the number of arguments means left hand side and right hand side argument should be matching and divided by the length of x. And this is how I can directly compute the mean deviation using the built in functions.

(Refer Slide Time: 27:29)



So, now we consider an example here, this is the same data set which we had used in the earlier example where I had collected the observations on marks of 15 students. And they are combined inside a vector marks. So, I try to expose all this function whatever we have done on this data set marks. So, in case if I want to find out the variance of marks I simply have to write down var inside the argument I have to write marks. And as soon as you enter you get this value. And similarly if you want to find out the standard deviation you simply have to take the square root of this value mark. And as soon as you operate this function you get here the value 20.95 and here the screenshot of the operation.

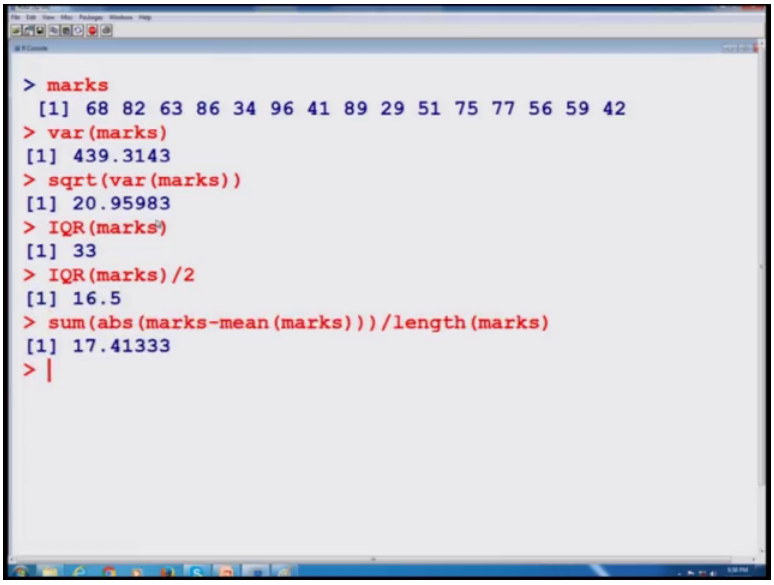
(Refer Slide Time: 28:22)



Now, we compute the interquartile range.

So, we know that the function for the interquartile range is simply IQR in capital letters and the data vector here marks and as soon as you enter it you get here the value 33. And similarly if you want to compute the quartile deviation you simply have to divide the interquartile range which you have obtained here by 2 and you get here this value 16.5. And in case if you want to compute the mean deviation you simply have to use this expression which is simply one upon n summation i goes from 1 to n absolute value of xi minus x bar. And this gives us this value and here are the screenshots. So, one have to do all these things on the R console and try to see what do we obtain? Now we come to our R console and try to do all these things over here. So, you can see here we already have there is marks data set.

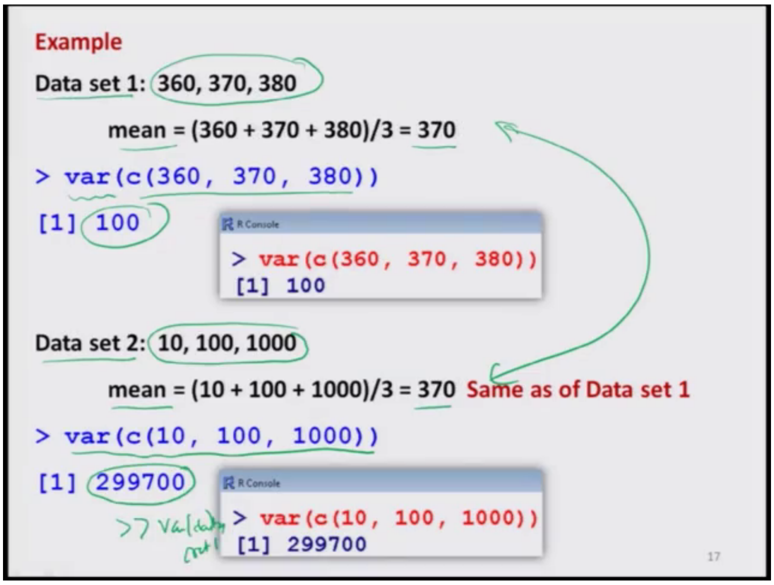
(Refer Slide Time: 29:21)



So, now first I try to find out the variance of marks you can see here you have to use the function var as soon as you enter you get it is variance. And in case if you want to find out the standard deviation you simply have to find out the square root of variance of marks and the positive square root is called as standard deviation. So, this is the standard deviation of the data vector mark; and now if you want to find out the interquartile range. So, you simply have to type IQR and the data vector here marks this gives here 33; and simply if you want to find out the quartile deviation which is the half of the interquartile range. So, you simply have to write down IQR marks and divided by 2 this gives here this thing and in case.

If you want to find out the mean deviation this is here given by this thing and then you can see here, this is given by like this value 17.41. So, you can see that it is not very difficult to compute this value directly from the R software. So, now, let us come back to our slides and try to see what we can do more. Now we try to consider the earlier example where I had considered 2 data set data set one and data set 2 data set.

(Refer Slide Time: 30:44)

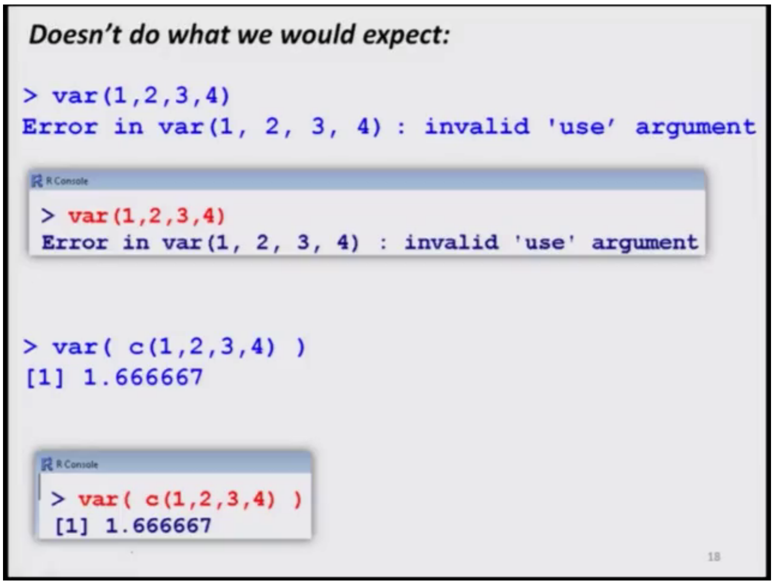


One has value 3 6370 380 which are very close and the data set 2 has value 110 and 1000 which are quite far away from each other and then I try to find out the mean of data set one this comes out to be 370.

And the mean of data set 2 this comes out to be 370. And both these values are same then the question was what is the difference between the 2 data set or how to discriminate between the 2 data set. So, now, we know we simply want to compute the variability. For example, we try to compute the variance of data set one, this comes out to be hundred and then I try to compute the variance of data set 2. This comes out to be 2 9 9 7 0 0 which you can see here this is much larger than the variance of data set 1. So; obviously, now you know that there is another feature which is contained inside the data set which is given by the variance and now you have quantified it.

So, you can choose a data set which has got the lower variance that is always preferred.

(Refer Slide Time: 32:02)



Now, a word of caution; suppose you want to find out the variance of 4 values 1 2 3 and 4. And suppose you use the variance function var directly on these values 1 2 3 4 which are say given inside a argument then you will not get anything, but it will give you that there is some problem in this command and this is an invalid command. Why because you have not given the values in the form of a vector. So, in order to find out the variance of one 2 3 4 first we should combine them in a vector using the c command and then use the variance function and this will give us the correct value. And here are the screenshots of the outcome.

So, you can see and you can try yourself. Now we stopped in this lecture. And we have learnt how to measure the central tendency of the data and how to measure the variation in the data using some popular functions. And you need not to write a individual program to write down these quantities, but R base function is giving all this function directly they are the built in functions. And that is the advantage of using R, that you can compute all these values directly without any trouble.

So, I would request you that you please try to do some practice using some data sets. And we will see you in the next lecture with some more topics, till then goodbye.