Essentials of Data Science with R Software - 1
Professor Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur
Lecture No. 27
Data Based Moments and Variance in R Software

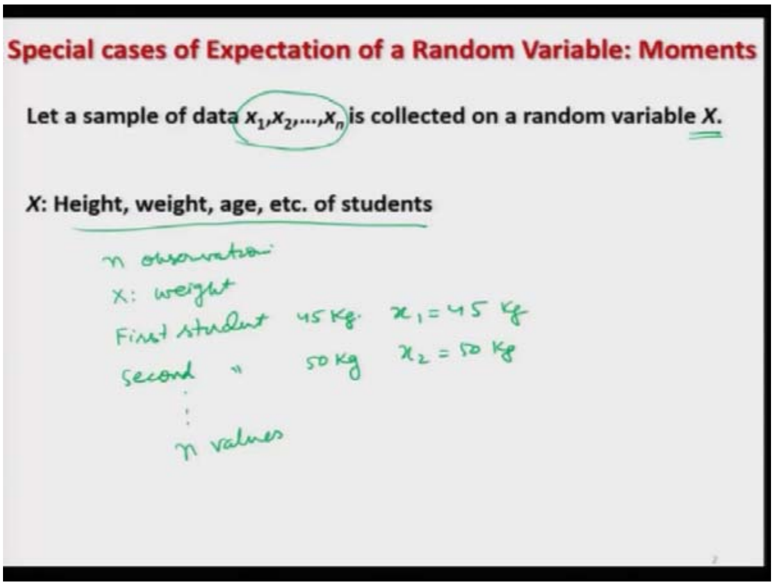
Hello friends welcome to the course Essentials of Data Science with R Software 1 in which we are trying to understand the basic concepts of probability theory and statistical inference. So, you can recall that in the last lecture, we had introduced the concept of moments and we had understood the meaning of first and second moments.

The first moment is the first raw moment actually that is going to give us the information about the mean value or the central tendency of the data and second central moment, which is actually the variance that is the central moment which is measured around the mean that is going to give us an idea about the variation in the data, but those things we have computed in a theoretical framework.

Now, in case if you get the data, how are you going to compute these types of things and how are you going to compute the moments in general that is what is the objective of this lecture. So, we are going to consider here that suppose, you have got a set of data, I am not saying that the set of data means is small or say large when you are working in data science, the data science can be very huge, but that would be some finite number and if you are working in some small samples, the data size might be say only 10, 20, 30 also.

So, in general I will say that if we have a data of size say small n and based on that, how are you going to compute the moments and other types of measures that we will try to see, in this lecture and we will try to implement it in the R software also. So, let us begin our lecture.

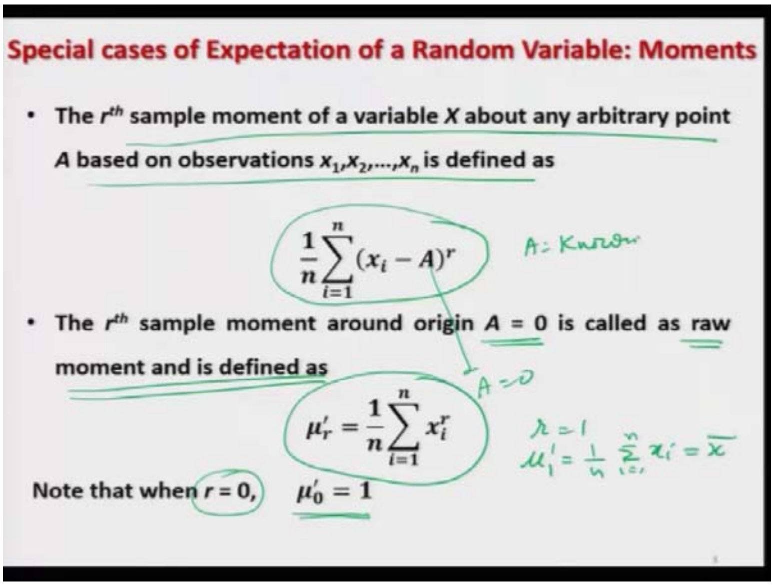
(Refer Slide Time: 02:08)



So, now suppose, we have a sample of data and that sample is collected on a random variable X for example, the X can be height, weight, age etcetera of our students in a class. So, the observations on variables are obtained as $x_1, x_2, \ldots, x_n$. So, what is the meaning of $x_1, x_2, \ldots, x_n$ that means we have got here n observations.

Suppose, if I say X is my here weight suppose, I take the first student and we find out the weight of the students suppose it comes out to be 45 kilogram. So, the quickest is going to be x_1 is equal to 45 and similarly, if I try to take a second student and suppose the age of secondary income, so to be 50 kg then x_2 is going to be 50 kg. So, and so on you can get here n such values.

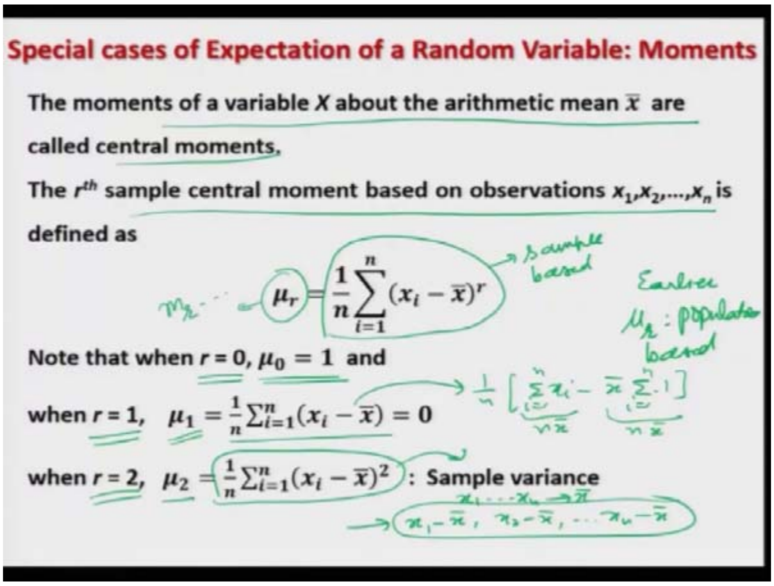
(Refer Slide Time: 03:18)



Now, based on that we are going to work how are we first going to define our moments that we defined earlier based on the probability function. So, the rth sample moment of a variable X about any arbitrary point a based on the observation $x_1, x_2, \ldots, x_n$ is defined at simply here $\frac{1}{n}\sum_{i=1}^{n}(x_i - A)^r$. So, this is how we will try to define it, where A is going to be some known value that we know and in case if you try to choose here equal to 0, then the rth sample moment, sample moment means, the moment which I am trying to compute on the basis of given sample of data.

That is called as raw moment and it is defined here as a μ_r' which will become here simply $\frac{1}{n}\sum_{i=1}^{n} x_i^r$, because you simply substitute here equal to 0 and now, even if you try to take here r equal to 0, then you see what happens, this will give you μ_0' is equal to 1 and in case if you try to take here r equal to 1, then what will it become, μ_1' will become here $\frac{1}{n}\sum_{i=1}^{n} x_i$ which is nothing but your sample mean or the mean of the sample observation.

(Refer Slide Time: 04:46)



So, now, in case if you try to define the central moments. So, what are the central moments, the moments of a random variable X about the arithmetic mean x bar they are called as central moment. So, the rth sample base central moment of the observation $x_1, x_2, \ldots, x_n$ is defined as say $\mu_r = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^r$.

One point I would like to make it here clear that well this is a quantity which is sample based and means, earlier what we had done, we had defined the μ_r which was population based. So, ideally this symbol μ_r which I am using here to indicate the rth central moment with this should be something like m_r or something.

But in practice what I have observed that in data science and other applied statisticians, they are using this μ symbol more often than say any other symbol, μ_1, μ_2, they are although they are indicating officially the population value, but what people are trying to use them as they are the values obtained from the sample. So, that is why intentionally I am not using here any other symbol So, that those people who are from this background, they may not get confused that I am trying to do something new.

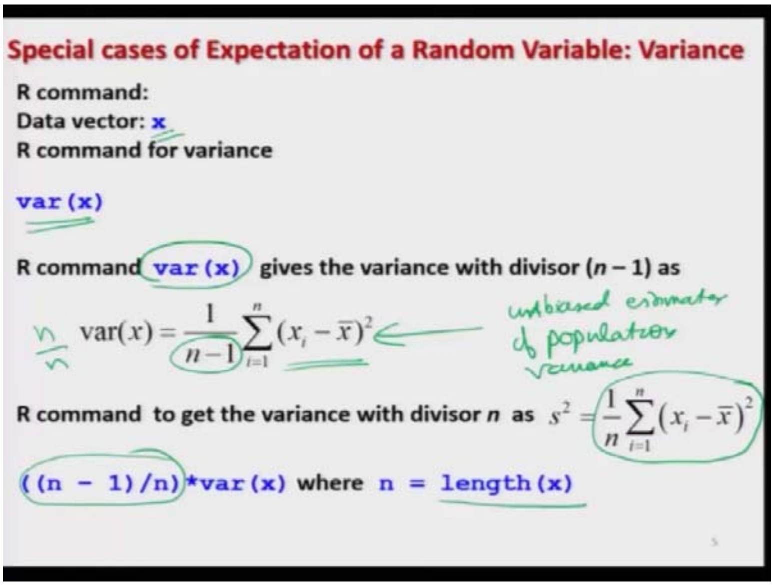
But this is the point where you have to be careful that ideally you must use here some other simple which is indicating that the values are sample based. So, in this case, you will see that in case if you try to take here r equal to 0, then μ_0 also comes out to be 1 and if you try to take it here r equal to 1, then you can see here what is the value of μ_1 that will become here $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) = 0$and if you try to solve it, this will become here $\frac{1}{n}\sum_{i=1}^{n} x_i$ - n$\bar{x}$. So, this quantity is simply your n$\bar{x}$ and this quantity here is n$\bar{x}$. So, this entire quantity becomes 0.

So, this is the very important result that you have to always remember that the sum of the observations which are measured as the deviation from the arithmetic mean is always 0. That means, in case if you are given some observations x_1, x_2,…, x_n and in case if you try to transform them that you try to find out their arithmetic mean and then try to define them say x1 as say x1 -$\bar{x}$, x2 as x2 - $\bar{x}$, and so on, x_n as x_n - $\bar{x}$ try to transform them.

Then the mean of these sets of new transformed observation will always be equal to 0 and that is a property which is very frequently used in statistics many times whenever we want to make the mean of certain variable to be 0, we always try to subtract the values of the random variable by its mean and remember one thing μ_1 is always equal to 0 whereas, μ_1' is actually sample mean, sample arithmetic mean.

Whereas, in case if you try to take here r equal to 2, then this μ_2 will become here simply $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ which is the sample variance. So, sometimes people say that μ_1 is mean and μ_2 is various which is not correct, μ_1 prime is mean and μ_2 is the variance that is trying to indicate the properties of central tendency and variation in the data.

(Refer Slide Time: 09:09)



Now will let us try to see how you can compute this variance and moments in the R software. So, in our software if you want to compute the variance, for some given data vector here x, then the command here is simply var and you have to write var inside the parentheses you have to write x. Well, one thing I can share with you that whenever I am trying to use here the R commands, I am using them in the most simple format assuming that all the observations are available to us and there is no issue, but this variance command has many options also, that in case if some data is missing, then how to find it out and there are many other things.

So, my request will be today and for all the lectures that I will be handling here the most simple case because I want to make you understand, but once you understand you must look into the help menu of the variance command or any command and then try to see what are they trying to inform you and you will see that that when you try to use those options, they will be very helpful in data sciences, because in the data sciences many time the observation for example are not available they are missing.

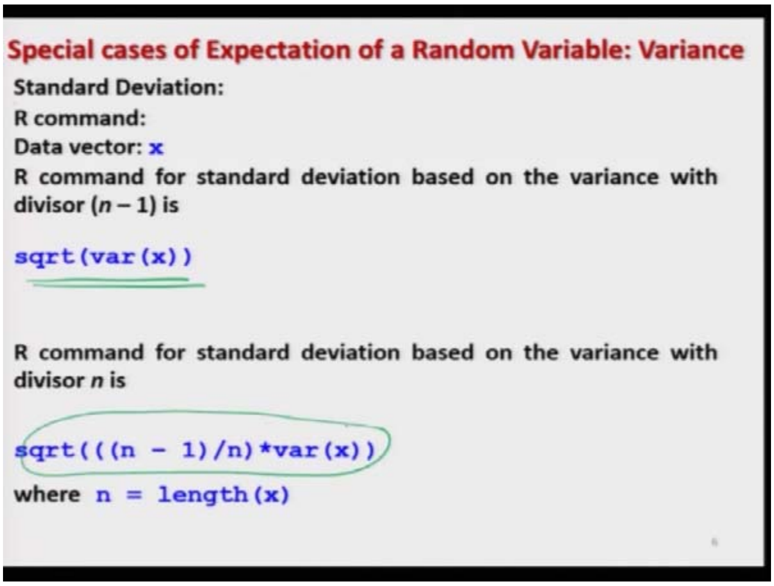
So, under those cases, this command variance of x will not actually work, will not give you the correct information, but if you try to use this variance commands with the possibility of excluding the missing observation, then possibly this is going to work. So, one thing which I would like to emphasize here that when you are trying to compute the variance using the command var in the R software, then this gives you the value of the variance which has a divisor n minus 1.

So, it is essentially computing $\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and but whereas, what you have learned that quantity has a divisor and like as $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ and the question comes that here what is the difference between the two quantities? Well, in this lecture definitely I am going to talk about it and here I can just share with you that this is related to the unbiasedness property of the estimator.

This one which you are obtaining here this is an unbiased estimator of population variance, but differently as soon as I write this sentence your mind takes 180 degree turns, what is this unbiased and what is this estimator etc., etc., but I simply say do not worry just take my words here and when I will be talking about the unbiased estimation etc, then surely I will try to come back on this aspect and I will explain you.

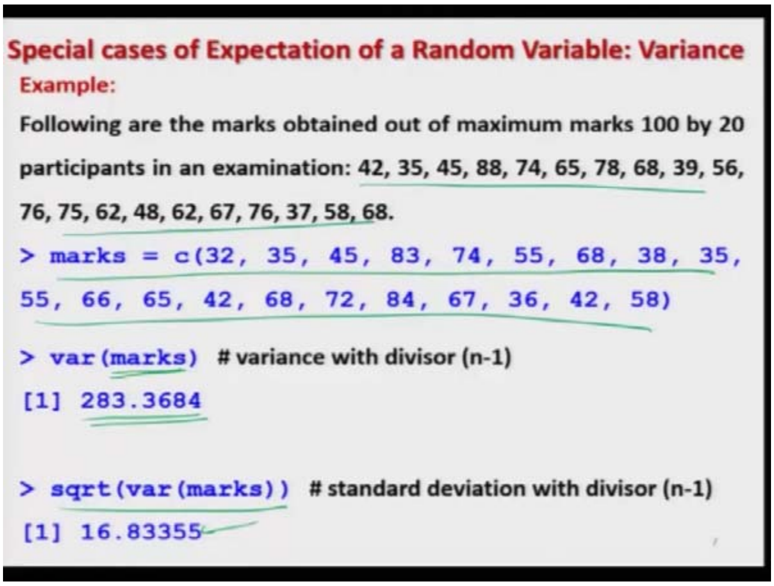
But, surely in case if you want to compute the quantity like $\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, then it is very simple, you simply try to multiply and divide this quantity by here n and then you will see here that this will become simply here is n minus 1 upon n into variance of x. Where n can be obtained as the number of observation using the command length of length of x l e n g t h. So, that is pretty simple.

(Refer Slide Time: 12:56)



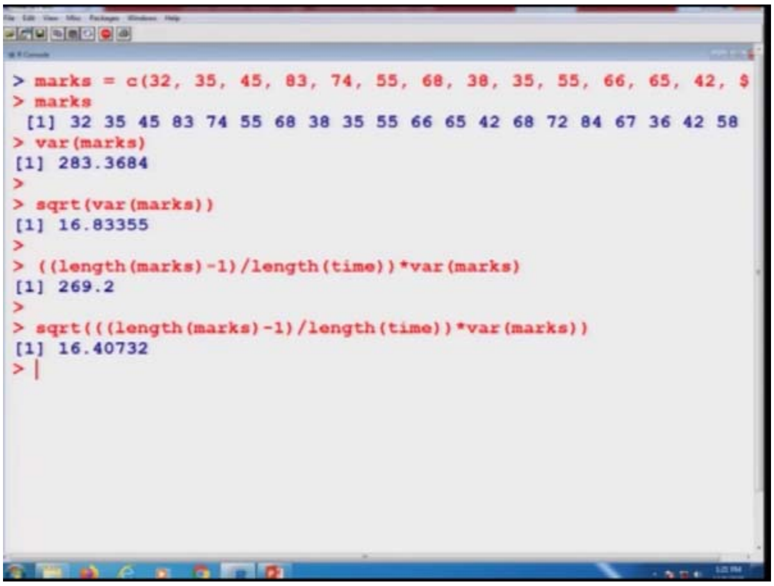
So, and then in case if you want to find out the standard deviation or standard error whatever you want to call, but definitely the divisor is going to be n minus 1 then in that case, you simply have to find out the square root of the variance and for if the square root you know the command is sqrt and if you want to find out the standard deviation or standard error based on the divisor n, then you simply have to find out the square root of the variance that you have obtained here. As simple as that very simple and straightforward concept.

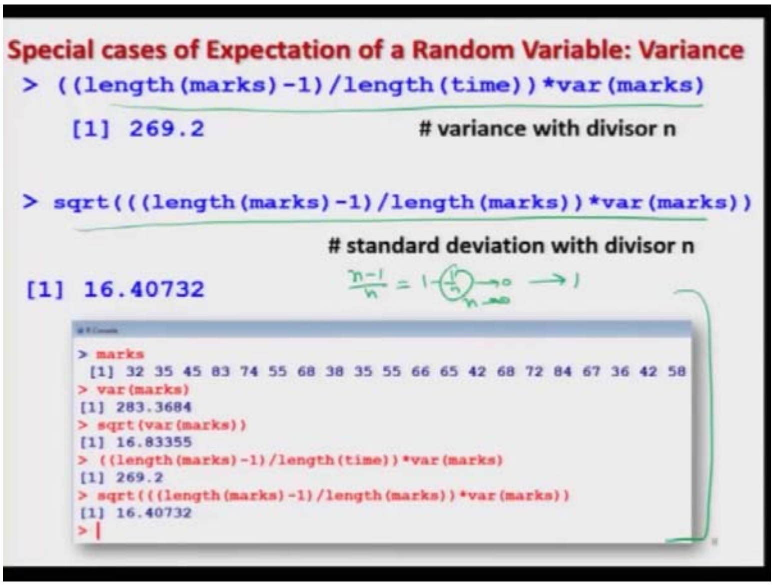
(Refer Slide Time: 13:30)



Now, let me try to show you that how are you going to do it on the R console. So, suppose there are 20 participants in an examination and who have got the marks out of 100, whose marks are reported here like this and these marks are stored in the variable marks, marks like this one and suppose we try to find out the variance of this data vector marks, then this comes out to be 2803.3684 and if you try to find out the standard deviation or standard error, whatever you call the square root of the variance of marks comes out to be 16.83355. So, let me try to show you these things on the art console, so that you get more confident. So, let me try to copy this data.

(Refer Slide Time: 14:20)





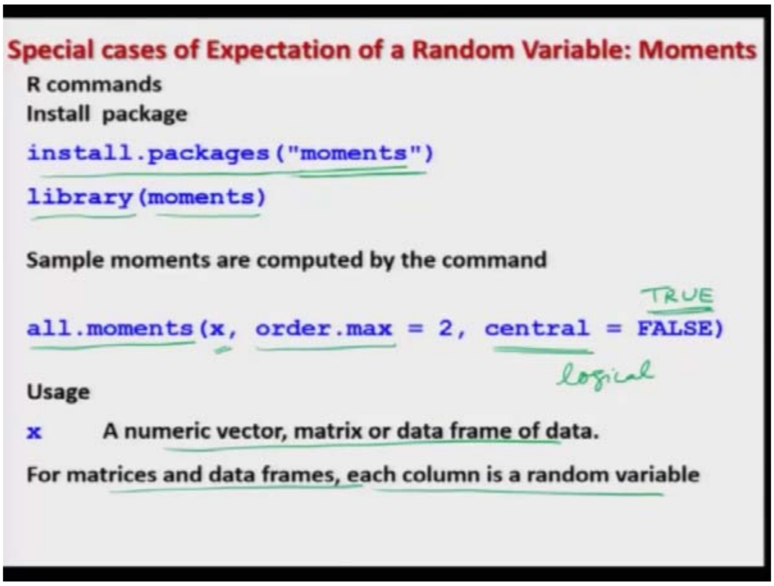
You can see here this is the dataset marks and now, in case if you try to find out the variance of marks, this comes out to here like this 2803.3684 and if you try to find out the square root of this quantity, you simply have to use the function sqrt and you can obtain it here 16.83355. So, that is pretty simple.

And if you really want to find out the variance in terms of the divisor n, then this command is going to be simply here say n minus 1 upon n into variance of marks and if you want to find out the standard error, this can be simply the square root of this, of the same quantity. So, and this is here the screenshot of the same operation, but let me try to show you these operations on the R console also. So, that you get more confident that yes, these things are working.

And I would like to show you here one thing actually here more and that the variance is like this and if you want to find out the standard deviation or standard error, just try to take the square root of this quantity. This is coming out to be 16.40. So, you can see here that although there are two forms of this variance in the divisor 1 upon n minus 1, it is going to give you the value 283 and in the case of divisor 1 upon n this is going to give you the value 269 that is obvious means when you are trying to divide the same quantity in both the expression by n or n minus 1, but definitely this difference will be more important when you when your sample size is small.

In case if the sample size is extremely large, then possibly the quantity is small n minus 1 upon n that will become nearly the same as here 1 and it will not practically make any difference whether you are using the divisor n or n minus 1. So, you can see here that n minus 1 upon n that is equal to 1 minus here 1 upon n and if your n is going to infinity this quantity will go to 0 and the entire quantity will go to 1. So, that will not make any much difference here.

(Refer Slide Time: 16:46)



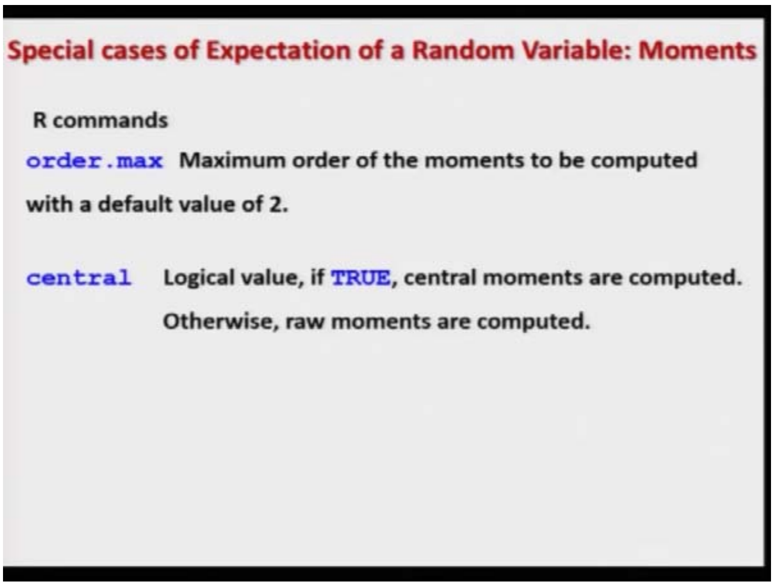
So, now I come back on say another aspect that I would like to give you some idea that how are you going to compute the moments, in order to compute the moments you need a special a package that package is not included in the base package of your R. So for that, you need a package whose name is moments, moments and so, you simply have to use the command here install dot packages and within the parentheses within double course you have to write moments and after that, you need to load the package by using the command library moments.

So, now in case if you try to compute the sample moments on the basis of given set of data, then we have the same command all dot moments for computing raw moment central moments and some other type of moments, you can find the details in the help, but these moments are just obtained by putting different choices inside the parentheses for different parameters. So, we try to see here how they are going to be obtained.

So, for example, if you try to see in this command all dot moment, there is one value inside the parentheses which is here x, so x is going to be a data vector for whichever you want to find out the moments. So, this can be a numeric vector matrix or a data frame. In case if you are trying to consider the data frame or the matrices as a input data, then each column is going to work like as a random variable and means there is a command here order dot max. So, that is going to indicate that what is the order of the moments that you want to compute.

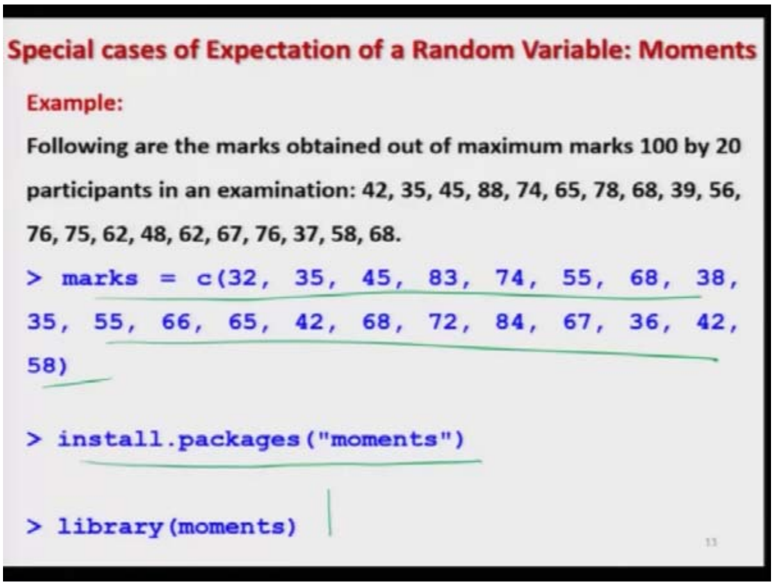
And there will be option here central, central this is a logical variable and its value can be either TRUE or FALSE. So, this can be here TRUE or FALSE and if you try to see here, if you try to understand the meaning, you are simply trying to say here central is equal to TRUE or FALSE. So, central is indicating the presence of central moments. So, when you are trying to say here central is equal to FALSE that means you do not want a central moment that means you want a raw moment and if you try to say here central is equal to TRUE that means you want the central moment.

(Refer Slide Time: 19:28)



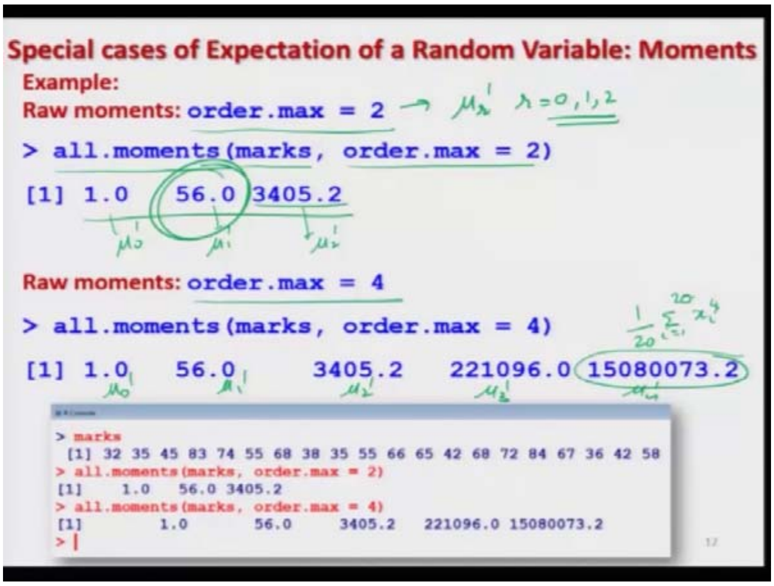
So, just by controlling this central as TRUE and FALSE computer raw and central moments. So, that is what I have written on this slide.

(Refer Slide Time: 19:36)



Now, can we try to take the same example and we try to compute this moment. So, we have the same data vector on which we have just computed the variance, which is stored in the data vector here marks, so we try to install the package moments and then we try to load it.

(Refer Slide Time: 19:57)



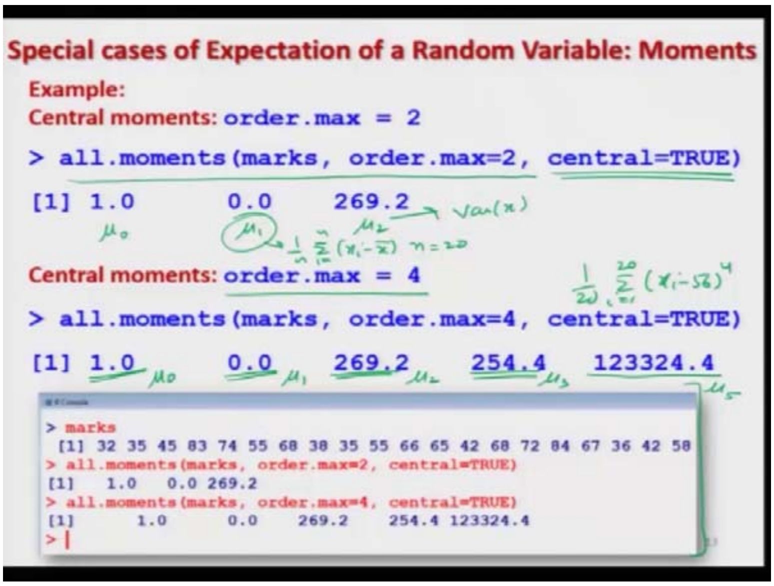
And after that, we try to use here two commands just to show you that if you want to find out the moments up to order 2 or the moments up to order 4 and how you have to be careful, how you have to interpret the values. So, suppose if I use the command here order dot max equal to 2 and we assume that we are interested in finding out the moments of 2 second order. Now, here you have to be careful that when you say order dot max is equal to 2, then what is the starting point? The starting point is from 0 or from 1 to 1.

So, whether you want to compute the first moment and second moment or you also want to compute the zeroth moment that is μ_0'. So, in R this start from r equal to 0, when you say order dot max then that means μ_r', r is equal to 0, 1, 2 this is how it is going to be used, we have to be careful many times people is think that this is going to give us only 2 values or if they try to look only in the outcome, they believe that this is going to give them 3 value which are the values of firstly moment So, be careful.

So, in case if you try to find out the moments of this data set for order dot max equal to 2, then we use the command here all dot moments inside the parentheses I will use the command marks, order dot max is equal to 2 and the output will be like this 1, 56.0, 340.2 and so, this is going to give you the μ_0' prime, this is the value which is the mean value μ_1' and this is the value of here in μ_2'.

And similarly, if you try to use here the option order dot max equal to 4, then you will get here the , $\mu_0', \mu_1', \mu_2', \mu_3', \mu_4'$. So, because of the for example sample. This is the value of here $\frac{1}{20}\sum_{i=1}^{20} x_i^4$. So, this value will come out to be like this. Now, so this was about the calculation of the raw moments.

(Refer Slide Time: 22:30)

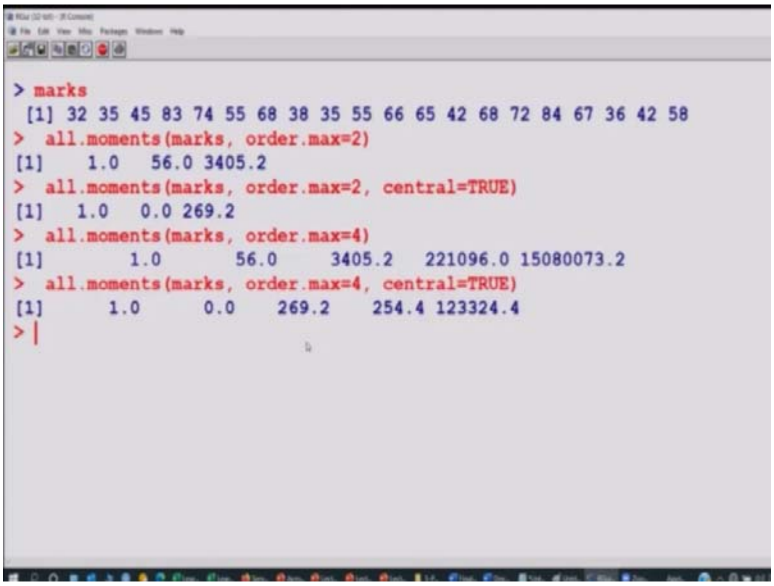


Now, in case if you want to compute the central moments, I will use the same command but I will simply make here the parameter central to be TRUE that is all I will simply add here 1 disk parameter. Central is equal to 2 and all other interpretation they will remain the same. So, I use here the command all dot moments marks, order dot max equal to 2. So, this is going to give us the value of here μ_0 and then here μ_1 and then here μ_2.

So, you can see here this mu1 is coming out to be 0, because this is the value of $\frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^{\square}$ and then n is equal to here 20 and μ_2 is the value of the same command that will come from the variance of x and similarly, if you want to find out the first 4 moments, so for that you have to use the command here order dot max where the first value will be coming out to be here the value of μ_0, then the second value will be of μ_1, third value will be of μ_2, fourth value we will be of μ_3 and the last fifth value will be of your μ_5.

So, in this case this μ_5 is going to be $\frac{1}{20}\sum_{i=1}^{20}(x_i - \bar{x})^{\square}$. Now what is the value of here mean that you can see here, this is here 56.0. So, this xi minus 56 is the power of here 4. So, you can see here that this R helps you in the quick computation of this problem problems and now, this is the screenshot of the same outcome that I have just shown you. Now, I would like to show you these outcomes on R console. So, let us come back to what could R console here.

(Refer Slide Time: 24:33)



And we clear the screen here Ctrl L. So, now you can see here we already had enter the data in terms of data vector marks and if you try to find out its moment by the command here like this means all dot moments marks dot order dot max equal to 2. So, that will give you the moments up to the order 2 which are the raw moments and if you try to add here central is equal to here TRUE, then you will see here that you are getting here the central moments.

So, this is your first central moment which is always 0 and this is your here second central moment and if you try to simply find out here the moments have to order here the moments up to the order here 4, these are here these values are the first raw moment, second raw moment, third raw moment and this is here fourth raw moment and if you want to find out here, the first for central moment, then you have to simply type here say central is equal to TRUE and you will get here the values of central moments.

So, you can see here that these are not very different things and competition of these moments is not difficult in the R software. But one thing, what are they going to indicate that is very important moments are simply the only the tools. For example, we have seen the interpretation of first raw moment and the second central moment, they are trying to depict or they are trying to take out the important information which is not visible from the eye, eyes, but it is hidden inside the data and these tools are going to give you that idea.

For example, the first raw moment is giving you an idea about the central tendency of the data and second central moment that way which is variance that is going to give you an idea about the spread of the data around the central value and these are very important properties which we try to study in terms of any dataset and we try to categorize it. So, for example, in this example, you can see that we have found moments have to order 4.

So, the first question comes that we have now understood what is the meaning of first and second moment what this third and fourth are trying to do? And what means the higher order moments like as fifth, sixth etcetera, they will try to indicate us. So, that is a topic which I am going to take in the next lecture.

So, try to have a quick revision try to take a small dataset and try to find out the value of the moment manually using your own hand manual calculations and then try to verify them in the R software also and then be confident. So, try to practice it and I will see you in the next lecture till then goodbye.