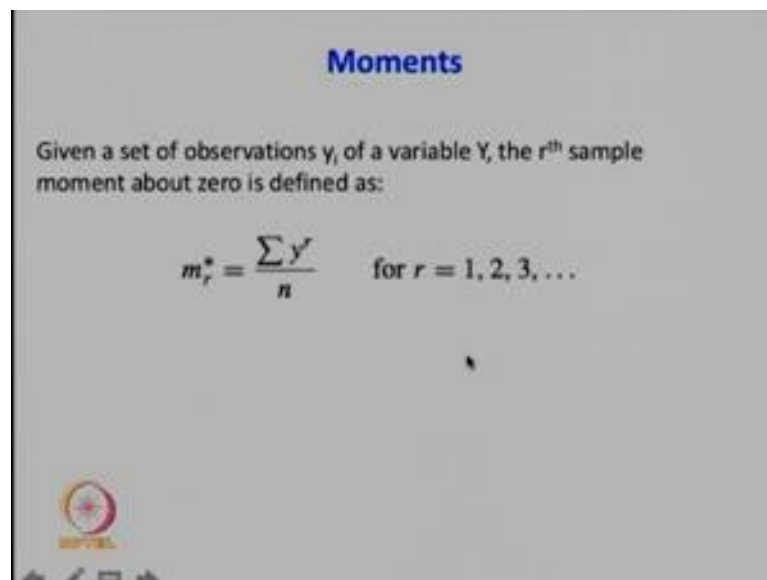


**Introduction to Biostatistics**  
**Prof. Shamik Sen**  
**Department of Bioscience and Bioengineering**  
**Indian Institute of Technology, Bombay**

**Lecture - 08**  
**Kurtosis, R Programming**

Hi. Welcome to today's lecture. I will start from with a brief recap of what we have discussed in last lecture. So, in last lecture, we began with standard deviation, we had a recap over zee score and plotting box plots. Now and last towards the end of last lecture, we had discussed about moments as a way of characterizing data.

(Refer Slide Time: 00:43)



**Moments**

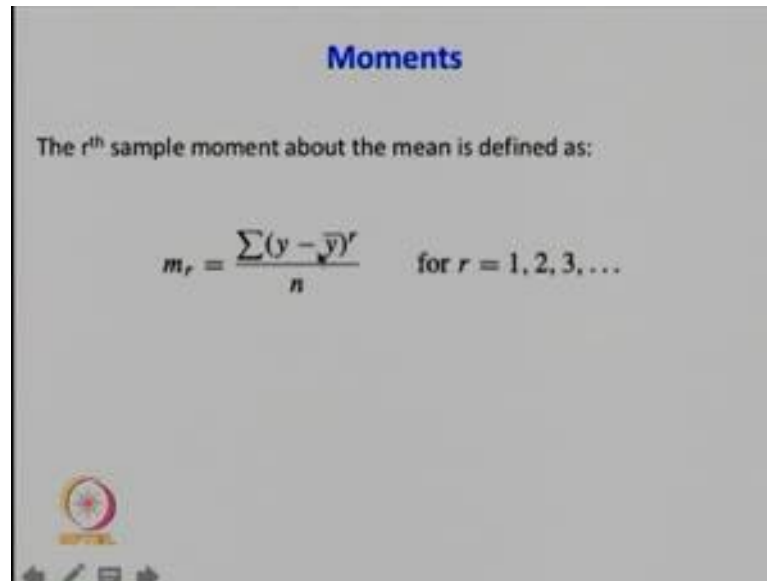
Given a set of observations  $y_i$  of a variable  $Y$ , the  $r^{\text{th}}$  sample moment about zero is defined as:

$$m_r^* = \frac{\sum y^r}{n} \quad \text{for } r = 1, 2, 3, \dots$$

The slide includes a small logo in the bottom left corner and navigation icons at the bottom.

So, the definition of moment as you would recall is in general. So, given a set of observations  $y_i$  of a variable  $y$  the  $r$ -th sample moment about 0 is defined as  $m_r^*$  is equal to summation  $y$  to the power  $r$  by  $n$  for  $r$  is 1, 2, 3, dot, dot, dot. So, clearly if we give; if we set the value of  $r$  equal to 1. So,  $m_1^*$  is summation  $y$  by  $n$  which is nothing but the mean. So, in other words the first moment about 0 of set of observations is the mean of the distribution.

(Refer Slide Time: 01:18)



So, we can next go define in a more general sense the r-th sample moment about any particular value and in particular we want to know the r-th sample moment about the mean. So, the r-th sample moment about the mean is defined by summation y minus y bar to the power r by n for r equal to 1, 2, 3, dot, dot, dot. So clearly, when you say about the mean if we are to generalize it about a value a instead of y bar, we will put a value of a. Again as before we did if you put r equal to 1. So, first moment about the mean is summation of y minus y bar whole to the power 1 by n and summation y minus y bar is going to give you a value of 0 as we had determined in last lecture.

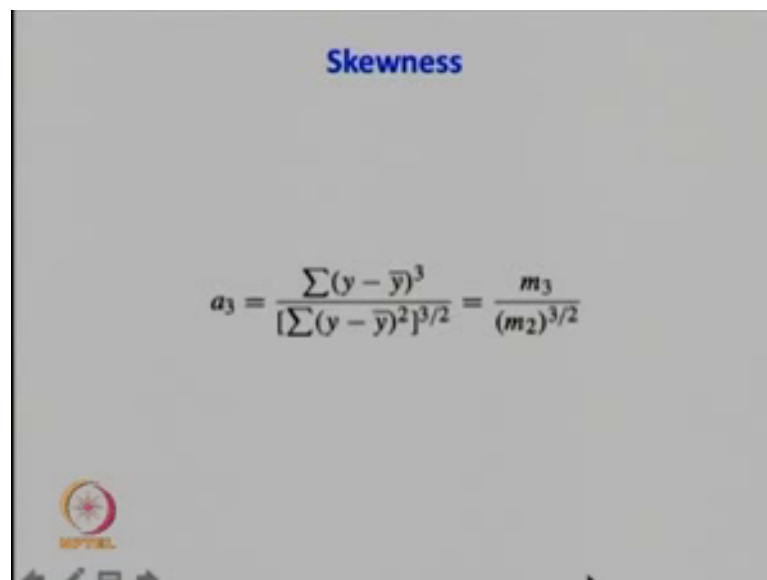
So, first moment about the mean is 0. What about the second moment about the mean if let us say, n is reasonably large. So, your m 2 is nothing but summation y minus y bar whole square by n and you can clearly see that this is nothing but very close to what is our definition of the variance. As opposed to divide it by n minus 1 we have divided by n, but for n large your m 2 is nothing but the sample variance. Again we can use it for getting higher moments like third moment about the mean y minus y bar whole to the power 3 by n.

So, one aspect that we discussed in last class was depending on the nature of the distribution all odd moments about the mean. So, in which means that m 1, m 3, m 5, m 7 so on and so forth, will return you a value of 0. This is because for every value of y which is situated to the left of y bar. So, there is another value of y which is situated to

the right of  $\bar{y}$  and their frequencies of these 2 values are equal which means that for every negative value that you accumulate for let us say  $y_1 - \bar{y}$  there is a corresponding  $y_2 - \bar{y}$  which is positive and equal value. So, these will cancel each other out eventually giving you a value of  $m_3$  or  $m_5$  which will be equal to 0.

But of course, for a non 0 for a asymmetric distribution, this value is not going to be 0 it will have some value. Now depend the way these moments are defined, if you have a value of  $y$  which has a given unit then this  $m_r$  will not return your value which is unit less with rather it is unit which it has some units. So, you of course, want to eliminate that I will you know that aspect of dimensionality in your measurements.

(Refer Slide Time: 03:50)



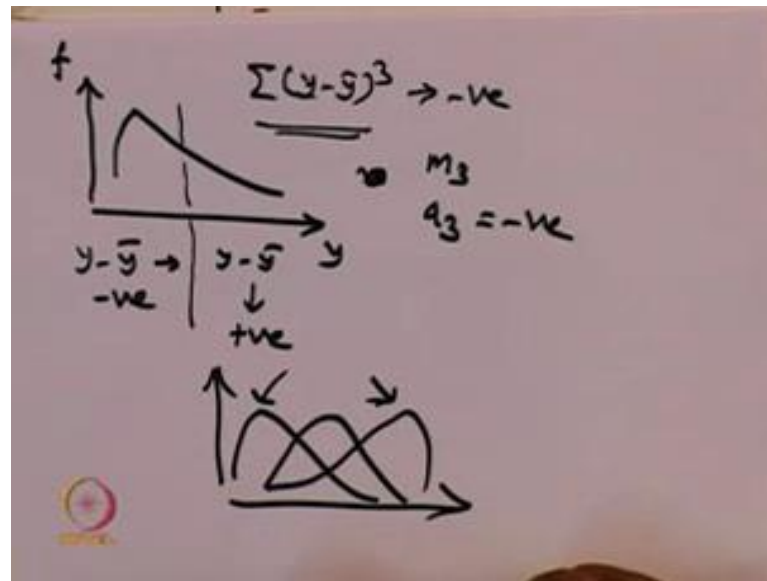
The slide is titled "Skewness" in blue text at the top. Below the title, the formula for the coefficient of skewness is displayed as follows:

$$a_3 = \frac{\sum(y - \bar{y})^3}{[\sum(y - \bar{y})^2]^{3/2}} = \frac{m_3}{(m_2)^{3/2}}$$

At the bottom left of the slide, there is a small logo for "BYJU'S" and some navigation icons.

And for that purpose what you typically do is you divide by another moment which is raised to some other powers so that the units are the same. So,  $a_3$  is one such measure, it is defined as summation of  $y$  minus  $\bar{y}$  whole cube by summation of  $y$  minus  $\bar{y}$  whole square whole to the power 3 by 2. So, it is nothing but  $m_3$  by  $m_2$  whole to the power 3 by 2. So, this is of course unit less as you can see from the definition this is called Skewness.

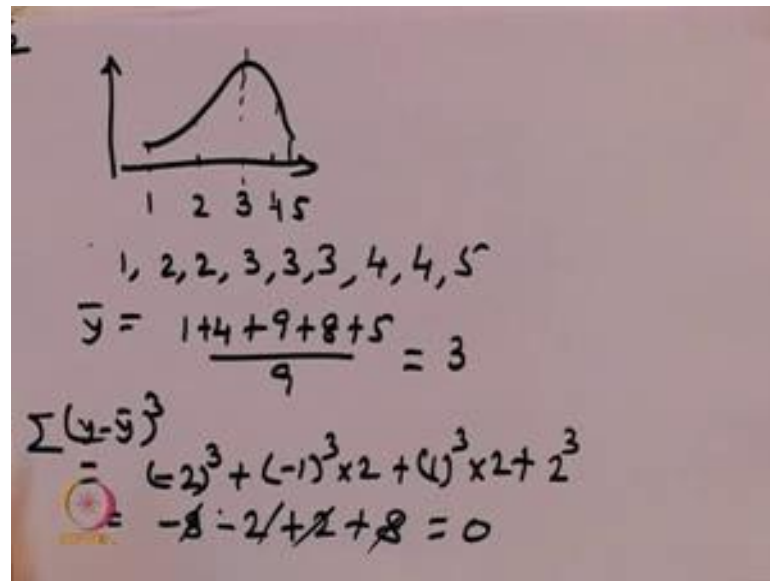
(Refer Slide Time: 04:22)



So, we had worked out. So, logic we had reasoned that if you have a distribution which one like this, this is  $y$ , this is frequency. So, your mean is somewhere here and so all these values so because there is a precedence of values which are to the left. So, all these  $y$  minus  $y$  bar values to then this domain will give me negative and in  $y$  minus  $y$  bar in this domain will be positive as a consequence of which there is a possibility that when you compute  $y$  minus  $y$  bar whole cube summation this might turn out to be negative. So, there is a greater chance that in this case that you the when you calculate  $m_3$  or  $a_3$  you get a value which is negative, because  $m_2$   $m_4$   $m_6$  are always positive because they have  $y$  minus  $y$  bar whole square whole fourth; whole 6 those are always positive. So, all odd moments for asymmetric distributions may be either negative or positive depending on how the data is biased.

So another measure; so Skewness of a data is basically to see differentiate it in a symmetric distribution with a non symmetric distribution either which is biased in to the left or biased to the right. So, these 2 are asymmetric and these will give me different values. So, we had worked out an example in last class where we tried to find out what is the Skewness measure for this particular population we anticipated it would be negative we turned out with a value which is slightly positive, but let us work out another example where let us say where we take a data which is biased to the right. So, in that case let me have these values.

(Refer Slide Time: 06:04)




So, let us say our mode is 3. And this is 1 this is 5, I have 2 and some intermediate value 4. So, let say I have one 1s, 2 2s, 3 3s, 2 4s and 1 5; so 1, 2, 3, 4, 5, 6, 7, 8, 9: 9 values. So, let a is you know let us calculate the mean. So,  $\bar{y}$  is going to be 1 plus 4 plus 9 plus 8 plus 5, so 9 numbers which is 5 plus 5; 10, 9, 10 and 17; 2. So,  $\bar{y}$  is nothing but 3. So,  $y - \bar{y}$  whole cube is going to be give me a value of minus 2 whole cubed plus minus 1 whole cubed into 2 plus 1 whole cubed into 2 plus 2 cubed. So, I have minus 8 here I have minus 1 minus 2 here plus 2 here plus 8 here. So, 2 cubed is 8, 8, 8. So, these exactly balance each other out and in for this particular distribution I get  $y - \bar{y}$  whole cube to be 0 summation of  $y - \bar{y}$  whole cube to be 0.

So, clearly what you see is your you know your data is slowly shifting to the right from a value so, but if we have if you bias the data even to the right side even more then we will slowly get a value which is much positive than  $\bar{y}$  and than otherwise there is another metric of you know of characterizing a distribution which we call as kurtosis.

(Refer Slide Time: 08:08)

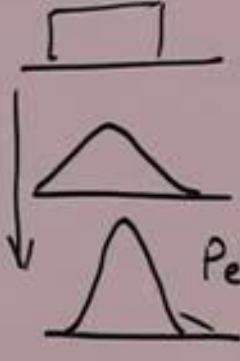
**Kurtosis**

$$a_4 = \frac{\sum (y - \bar{y})^4}{[\sum (y - \bar{y})^2]^2} = \frac{m_4}{(m_2)^2}$$



There is another metric of you know of characterizing a distribution which we call as kurtosis. So, the kurtosis is a way of measuring the peakedness of a curve or how flat or how sharp is the curve.

(Refer Slide Time: 08:21)

3


$$a_4 = \frac{\sum (y - \bar{y})^4}{[\sum (y - \bar{y})^2]^2} = \frac{m_4}{(m_2)^2}$$

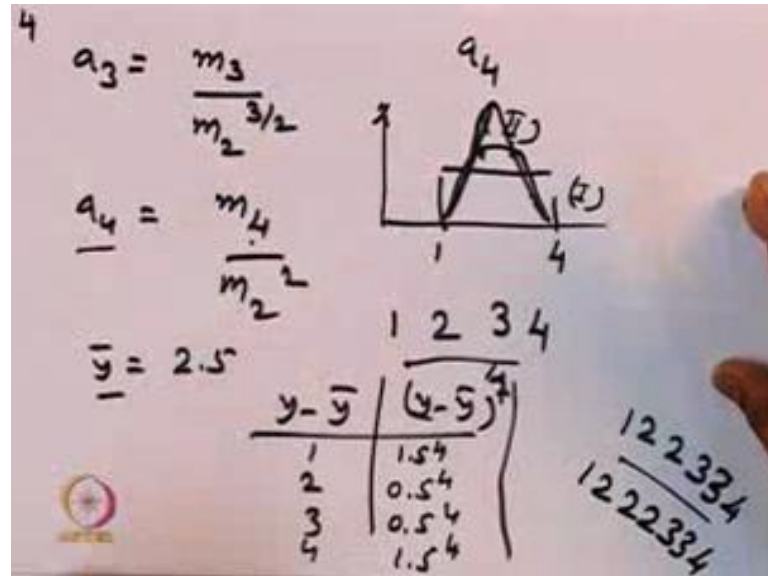
Peakedness  
Flatness



So, I can have 2 curves let us say this is one situation this is another situation this is another situation. So, what I see is the peakedness of the curves this is increasing right this is called it measures the peakedness or flatness of a curve and it is given by this metric of a 4 which is defined by summation  $y$  minus  $\bar{y}$  whole to the power fourth by

summation  $y$  minus  $\bar{y}$  whole square whole square. So, this is nothing but  $m_4$  by  $m_2$  square. So, if we compare our definition of  $a_3$  and  $a_4$ .

(Refer Slide Time: 09:10)



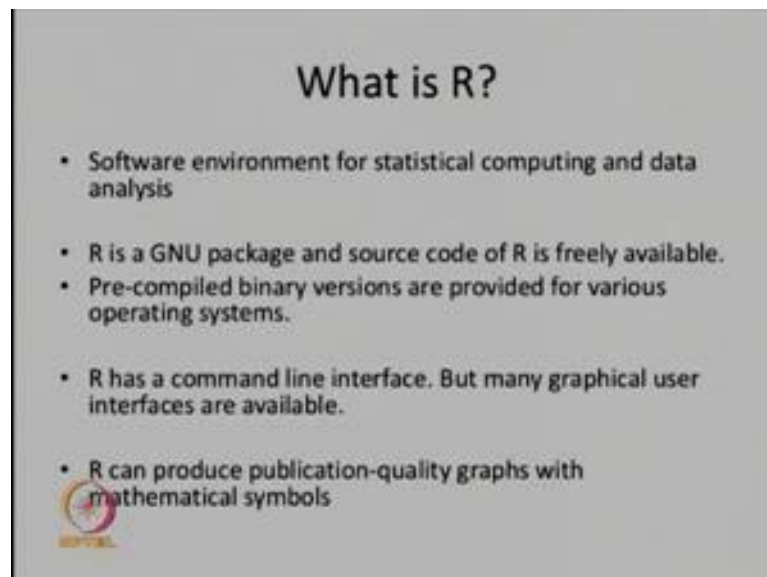
So,  $a_3$  I will; I write  $a_3$  was defined as  $m_3$  by  $m_2$  whole to the power  $3/2$   $a_4$  is defined by  $m_4$  by  $m_2$  whole square and as I said. So, the aim here is to define these matrixes in such a way that you come up with a non dimensional term and that is exactly how you have defined  $m_4$  because  $m_4$  has powers of  $y$  minus  $\bar{y}$  whole to the power 4. So, the definition of  $m_4$  has whole power 4 this has whole power 2. So, you have to you know square it to generate something which has the same units of  $m_4$  and that is why you are  $a_4$  is defined by  $m_4$  by  $m_2$  square.

So let us calculate a simple example where we calculate  $a_4$ . So, we want to calculate  $a_4$ , let us say our data is. So, let us take a very flat distribution. So, let us say our data is 1, 2, 3, 4. So, each of these values only appear once in this case. So, your  $\bar{y}$  is equal to 2.5. So, I can have  $y$  minus  $\bar{y}$  1, 2, 3, 4,  $y$  minus  $\bar{y}$  whole square equal to the power fourth. So, this is 2.5. So, 1.5 whole to the power 4 this is 0.5 whole to the power 4. So, you can go through the calculation and see what value of  $a_4$ , you get for this distribution versus for another distribution where let us say, 1, 2, 2, 3, 4, 3, 3, 4. So, we have made the distribution. So, the other point let us say slightly higher.

So, we have generated another distribution. So, let us say this is distribution one which is completely flat and this is distribution 2 which is slightly more peaked because you have these 2 points which are occurring at a slightly higher frequency please go through this calculation and see what value of a 4 you get you can generate one more distribution where you arbitrarily let us say you make it 1, 2, 2, 2, 3, 3, 4. So, it is a symmetric. So, let us say the next distribution is a symmetric, but 2 has a higher value you can make a you know keep a keep on making it more and more peaked and see what kind of value you get.

So, these exercises will help you get an idea of how to go about generating or coming up with important matrix of quantifying the statistics of the data. So, in the later part of this class today I wanted to discuss about an analytical tool or statistical how can you use a software to do this statistical analysis of course, we very well saw that even for these 4 points, if you have to start to do the calculation by hand beyond a point we are not able to do. So, we need a tool which would enable us to do these calculations if you have clear data sets or where the data is to be actually read from a file where you have let us say you know observations from 10 different experiments so on and so forth.

(Refer Slide Time: 12:44)

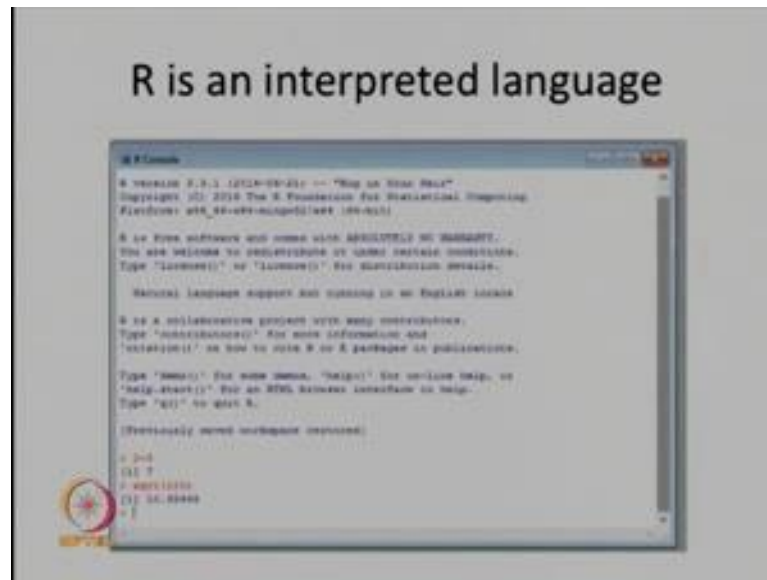


In that case, I wanted to you know introduce you to this language called r. So, this so what exactly is R? R is a software environment which is used for data analysis specifically it is a GNU package and the source code of r is freely available. So, that is



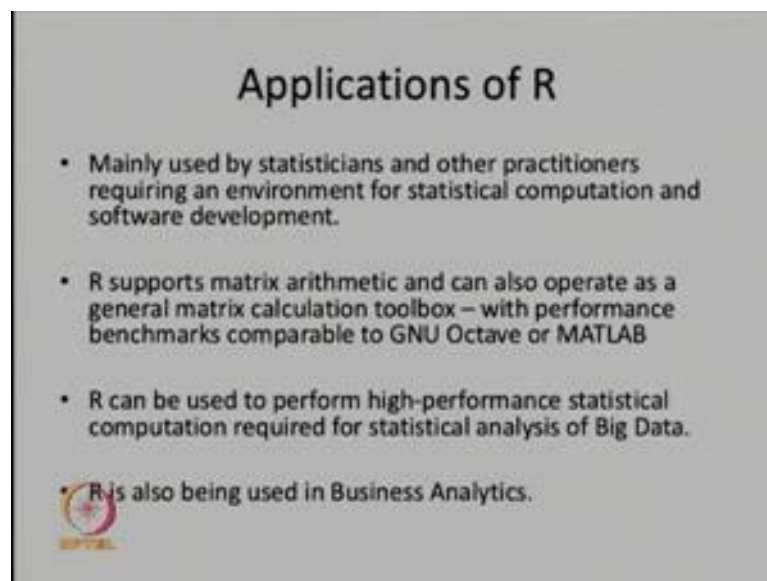
the best part of it you can and it has a command line interface and it has other interfaces also. So, and more importantly it can produce publication quality graphs with mathematical symbols.

(Refer Slide Time: 13:09)



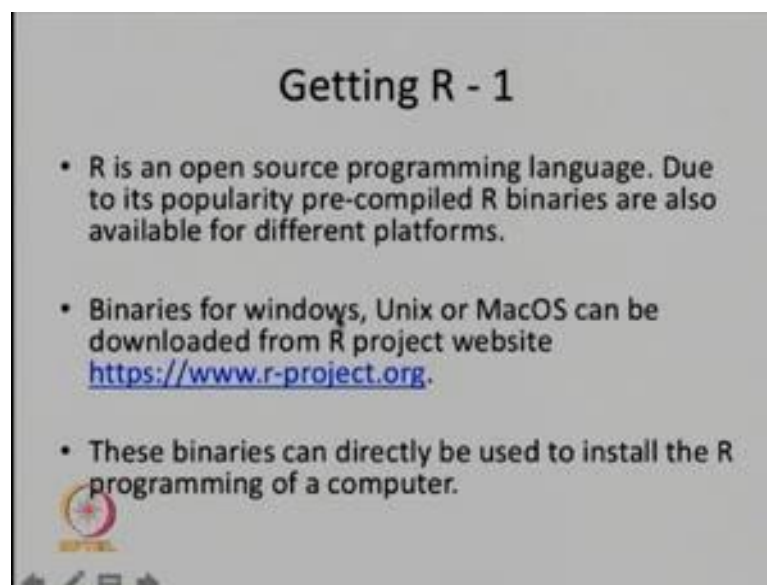
So, R is essentially an interpreted language this is a sample example of a console of R. So, you have you know this is as it is written clearly here r is free software and comes with absolutely no warranty. So, you can believe me you can download it from the net and I will come to the details of how you can download it and how you can use it.

(Refer Slide Time: 13:29)



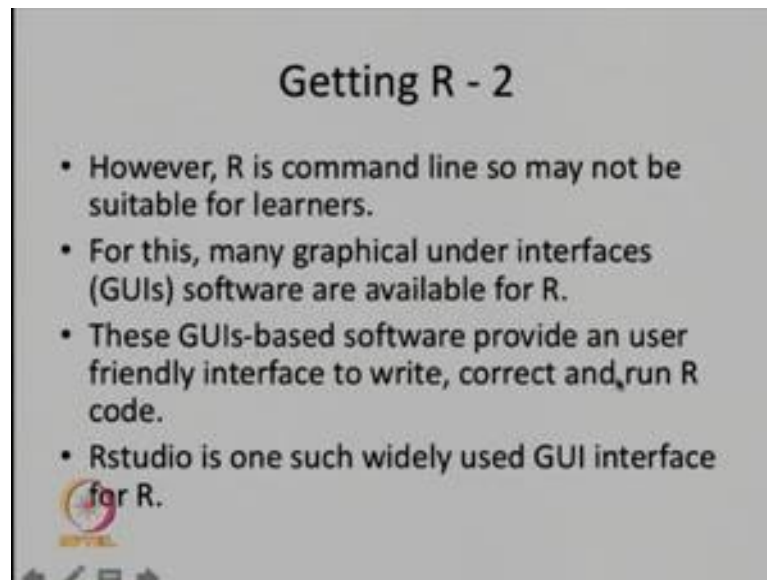
So, what are the applications of this R language it is used by statisticians it was established in the University of Auckland and it is now widely used to the extent that there are group of researchers who contribute to the further development of this language. So, it is requiring it is used by statisticians for statistical computation and software development R supports matrix arithmetic and its performance is comparable to that of you know expensive softwares widely used expensive softwares like MATLAB for which you need to purchase a license and R can be used to perform high performance statistical computation and to the extent that it is also used by the business fraternity.

(Refer Slide Time: 14:16)



So, it brings us to the first question how do you get R. So, R is an open source programming language. So, you can download it from this. So, there is a website called r-project.org and the best part about it is available in all the different formats, all the different operating systems. So, you can download it for windows, you can download it for Linux or you can download it for Mac.

(Refer Slide Time: 14:39)



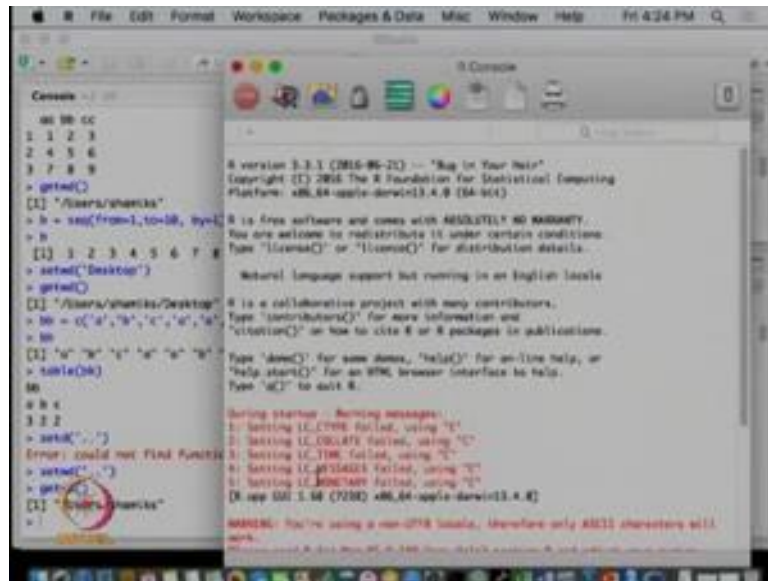
So, now R itself is a command line interface. So, sometimes people want graphical user interfaces for easy use of the facility and even to understand how it is used. So, R there are various GUI softwares which you know run our code. So, RStudio is one such then.

(Refer Slide Time: 15:03)



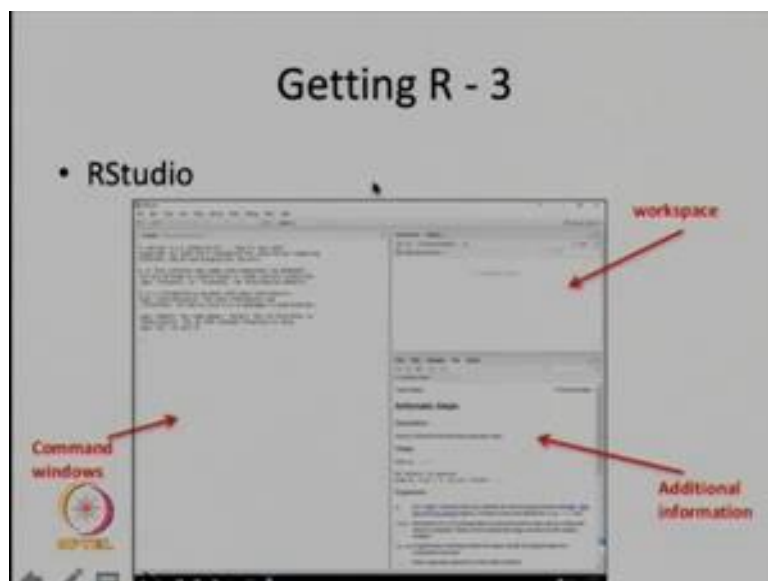
So, let me give you an example of how these R and RStudio work. So, this is a console of R, this is an R console.

(Refer Slide Time: 15:10)



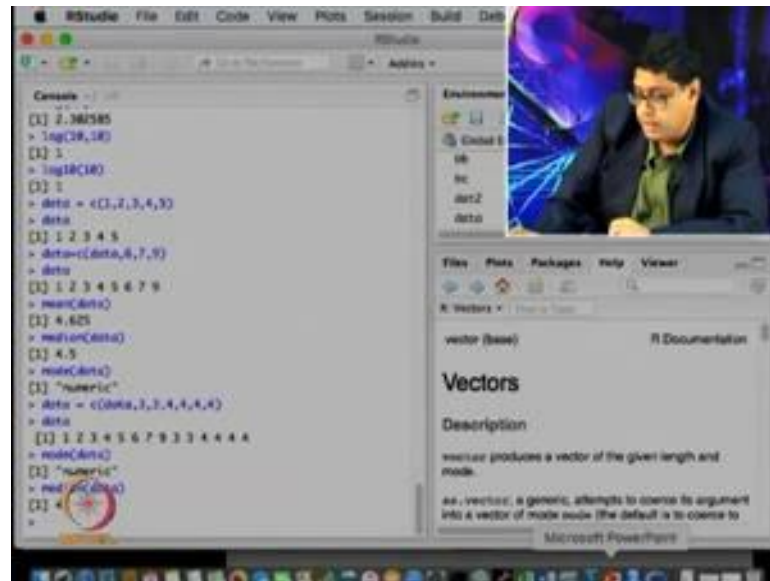
And this is an RStudio console. So, when we say console. So, you can download R and you can write down. So, in the console of course, you saw the difference between RStudio and R here everything you have to write down and then get to your point here you have a way of you know browsing through different aspects seeing what are the tools viewing and also the help file is much more easy accessible. So, you can download E you know any of these 2 softwares I recommend that you download RStudio for your use.

(Refer Slide Time: 15:57)



So that brings us to our studio. So, we can clearly see. So, this is for example the console of R studio. So, you have 3, if you look at the RStudio command, there are 3 different windows this is called the workspace. So, let us see if you have generated some variables you can you will see them being recorded here with the full information and this is the main command window where you will actually enter various things to do.

(Refer Slide Time: 16:26)



So, let us just do some simple computation in RStudio. So, I can open RStudio. So, I can let us say if I want to do simple arithmetic I can define a as a variable and I can assign it the value of one I can write a equal to 1. Now the value of a is not displaced, but what you see is in this section; you can see the value of a being generated and its value is; obviously, written here. So, in order to know exactly what is the value of a if you write a and press enter then you get a value of one. So, similarly I can do b b is equal to 2. So, note that in statistical language if you irrespective of whether you give a space or not it will still work it will not crimp it will still work. So, in both these cases b b is also stored at 2 and b b, b c is also stored too even though you gave spaces before the equal to, but for your own clarity it is better that when you write there is a space in between an equal to or any other symbol.

So, in the RStudio console itself we can do basic calculations. So, for example, I can write a plus b b enter. So, I get the value of 3 I can do simple arithmetic. So, I can write a a power b b. So, that is x to the power y right and I can enter I can evaluate the value of

one. So,  $a$  is 1;  $1$  to the power  $2$  is  $1$  I can write  $b$  to the power  $b$ . So, I can do. So,  $2$  square is  $4$ , I can do you know simple calculation. So, if I do sine of  $30$  degrees. Remember that for you know it is calculated in radians. So, sine of  $30$  degrees we always think it is half, but this guy is giving a value of minus  $0.98$  this is because it is calculating in radians.

So in order to calculate the value of sine of  $30$  in radians you have to write  $\frac{\pi}{6}$  if you write and you know good part is it gives you know, what is the way in which you write? So, sine of  $x$  is how you have to enter this value and within this you can do anything if I put enter here now I get a value of  $0.5$  which was not negative. So, when you are doing trigonometry calculations you have to enter these values in terms of radians. Similarly I can do the same thing  $\cos$  of  $\frac{\pi}{2}$  return me a value of this see. So, one thing is these things these values are calculated numerically. So, this is the reason why you see that when its value of  $\cos$   $\frac{\pi}{2}$  it is not  $0$ , but it is coming as  $6.12$  whatever into  $10$  to  $e$  minus  $17$  means  $6.17 \times 10^{-7}$  which is as good as  $0$ , but it is not exactly  $0$  and this is because these values are internally computed by a code. So, it is an approximate.

So, I can do the same thing I can do let us say  $\log$  of  $10$ . So, if you see the syntax it is  $\log$   $x$  comma base. So, I can write you know, this is another way of writing is. So, let us say if I do  $\log$  of  $10$ , I get the value  $2.3$ ; that means, that it is actually calculating the natural logarithm and not the  $\log$  base  $10$ . So, let us say if I do  $\log$   $10$  comma  $10$ . So now, it is giving a value of  $1$ .

If I entered the base and this is my number this is the base with which I am calculating my logarithm it is giving me the value of  $1$ , but says if you just write  $\log$  it will give you a value with respect to  $x$  the natural  $\log$ . So, I can also do  $\log$   $10$  of  $10$  then also you get a value of one. So, you can easily go through the list of these kind of inbuilt functions which do the basic calculations now let us say I have ten values right I have ten values and I want to calculate the you know let us say standard deviation mean or median of a distribution how do I do it.

So, what you do here is you enter let us say data is, because I did these you know I generated this data before it is already showing up as there, but I can write I can rewrite data, I would use this expression  $c$  of  $1$  comma  $2$  comma  $3$ . So, let us say I enter as a

vector. So, when the syntax is `c` and within that you have you put numbers by default you want to put numbers. So, when I do `enter` and then I write `data`. So, what you see here `data` got initialized to a row vector which has 5 entries one 2 3 4 5 in order to type `data` I should just write `data` and then I get back what it is and because it is a row vector it is showing as one of 1, 2, 3, 4, 5.

So now, I can change you know let us say I have 5 new entries I can write `data` is equal to, I can write `data` is equal to `c` of so I had my original data and I am overrating adding 3 more numbers say 6, 7, 9. So, I write `data` is equal to `c` of `data` comma 6, 7, 9. Now if I type `data`, so what you see here? `data` has now become a 8 column entry where in additional to 1, 2, 3, 4, 5, you have 3 more numbers which have been added.

So I can just get the value of `data` here by writing `data` and `enter` and then I get this value now calculating these basic matrix in you know in R is super simple. So, in order to calculate the mean of `data` I will just write `mean` of `data` and I will `enter` and I get the exact value which is 4.625. I can calculate a median of `data` `median` is 4.5. So, I have 1, 2, 3, 4, 5, 6, 7, and 8. So, my median is at position between 4 and 5 and which is nothing but 4.5 by 2 which is what it is giving 4.5. Now what is the mode of this distribution? We can clearly see that these are different values which do not have any, so there is no particular value which has you know which is maximal in frequency. So, if I write `mode` of `data` let us see what it gives.

It says `numeric`, which means you know, so this is it is not giving an exact value because it does I do not have any particular value which is repeating. So, let us again change the you know expression for `data` by writing `data` is equal to `c` of `data` comma 3; 3 comma 3 comma 4 comma 4 comma 4 comma 4. This is how I modify `data` I can type it here, but I can clearly see here, now a `data` is showing up as a fourteen column vector. Now once it is bigger than `a`; you know certain size of course, it is difficult to see here, but what you can do is you can write `data` and enquire its value. So, you have this entire distribution of `data`. So, I can just write.

Now, if I do `mode` of `data` is giving me `numeric` value. So, we have to see because now. So, let us see `median` `data` it is giving the value 4. So now, I guess if we arrange them in terms of ascending order then 4 will be there in the center and that is why this meeting is giving you a value of 4.

(Refer Slide Time: 24:21)

### Creating vectors in R

<pre>&gt; a = c(1,2,3,4,5,6) &gt; a [1] 1 2 3 4 5 6</pre>	<b>Custom vector</b>
<pre>&gt; b = seq(from=1,to=7,by=1) &gt; b [1] 1 2 3 4 5 6 7</pre>	<b>Sequence</b>
<pre>&gt; c = rep(1, times=10) &gt; c [1] 1 1 1 1 1 1 1 1 1 1</pre>	<b>Repeat</b>
<pre>&gt; d = rep(1:3, times=5) &gt; d [1] 1 2 3 1 2 3 1 2 3 1 2 3</pre>	<b>Repeat of range</b>
<pre>&gt; e = rep(seq(from=1,to=7,by=1), times=3) &gt; e [1] 1 2 3 4 5 6 7 1 2 3 4 5 6 7 1 2 3 4 5 6 7</pre>	<b>Repeat of sequence</b>

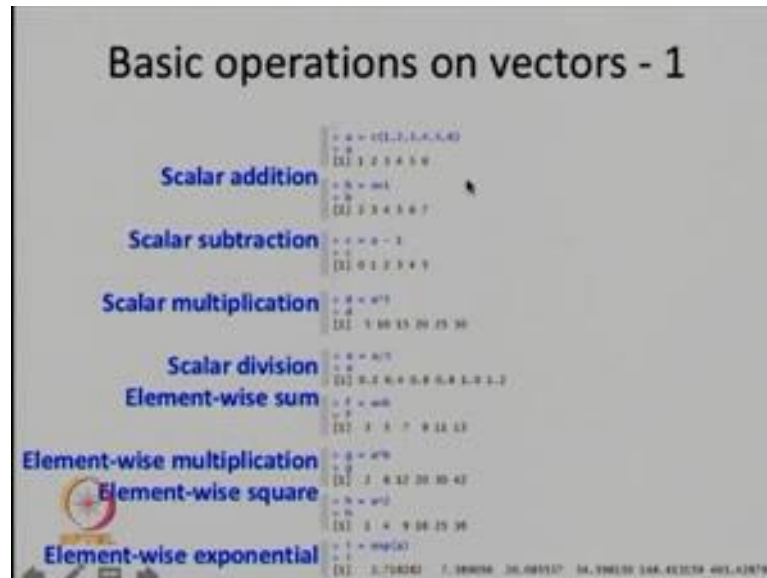
**All variables are vectors. Variables are case sensitive.**

So, if I go back to the presentation. So, let me just briefly you know say what we have done. So, we you can create a custom vector. So, using this c open bracket and then you have various entries 1, 2, 3, 4, 5, 6, in that way you will get these values you can enter them you can you know right introduce this vector as a sequence. So, I can write from 1 to 7 by 1 which means I want a sequence which is increasing in units of 1 I can generate this vector you can repeat it.

So, you have one which you want to repeat ten times you can generate this vector you can do this repeating sequence repeating of a range you can generate this victor similarly for sequence you can be. So of course, you need to remember that these are case specific. So, in one line you write b in the next line; you write capital B, you will be shown an error.



(Refer Slide Time: 25:10)



So, and we briefly discussed about you know all these you know basic manipulations like addition subtraction multiplication division. So, please note that if you have a vector if you have a vector then when you do these sums. So, let us say a is this particular vector when I write b is a plus 1; 1 is getting added element wise. So, that is why you are having you know 2, 3, 4, 5, 6, 7 from 1, 2, 3, 4, 5, and 6. So, these operations operate on element wise basis. So, which is why if you do c is equal to a by 5 you will get this particular values or a star b.

Now if you do a star b again it is an element wise operation first 1 will be 1 into 2. So, we had modified b somewhere else oh to b is a plus 1 c is a minus 1. So, a star b you can see that this will accordingly change. So, this is a plus b this is a star p c 1 into 2 is 2 in the last case 6 into 7 is 42. So, you have these particular elements or you can also do a power to exponential whatever. So, this gives you an idea of the usability of this particular language R which is very easy to learn you can download it you can use it for analyzing your data with that I stop here.

In the next class, we will do 1 more session with the language R to see; how we can import data from you know. So, of course, it is good enough to write 6 values, 7 values, but you have a trench of data then you need a way to import this data into this art software and operate on them and we will also briefly discussed how to do plotting.

With that I thank you for your attention today. And I look forward to having our next lecture.

Thank you.