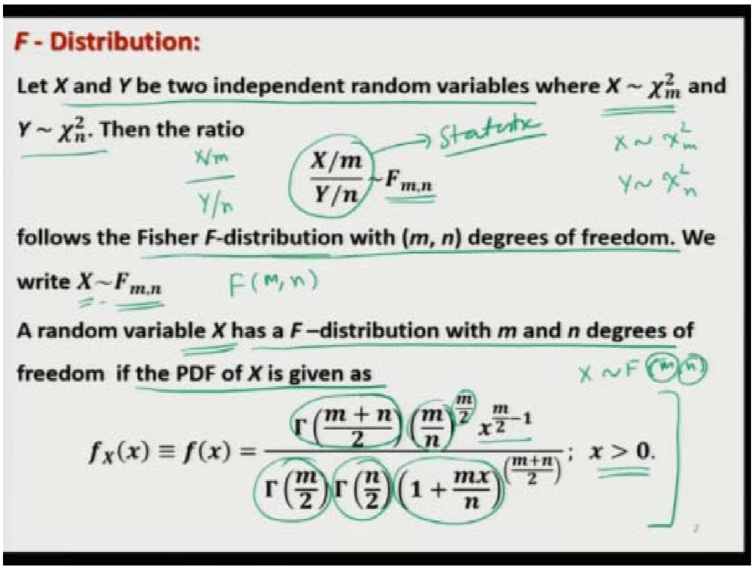
Essentials of Data Science with R Software- 1
Professor Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur
Lecture 52
F - Distribution

Hello friends, welcome to the course Essential of Data Science with R Software and you can recall that in the last two lectures we had talked about the sampling distributions and we have understood the chi square distribution and t distribution. Now, in this lecture we are going to learn about one more sampling distribution which is F distribution, just capital F and now after understanding this chi square and t, I do not think it is a very difficult thing for you to understand this distribution, the only thing what you want to know that what is the statistics whose distribution will be called as F distribution, what will be the form of that statistic.

(Refer Slide Time: 1:14)



And after that we are going to do the similar type of exercise what we have done earlier. So, we begin with this F distribution. So now you see here that in case if you try to consider two random variable X and Y and suppose both are independent and both of them are following a chi square distribution, suppose X is following a chi square with m degrees of freedom that is $X \sim \chi^2_m$ and $Y \sim \chi^2_n$, well m and n can be same also there is no issue.

Now, in case if you try to divide both the random variables by their respective degrees of freedom and then take their ratio, so what are you going to do, you are trying to take the random variable X divided by its degrees of freedom say m, then you try to take another random variable Y try to divide it with the degrees of freedom associated with it and then try to take the ratio of both the quantities.

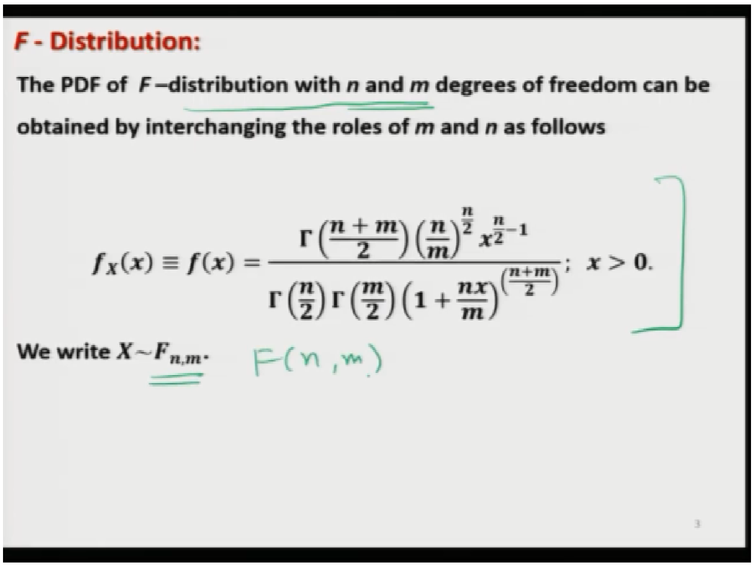
So, this will come out to be here $\frac{X/m}{Y/n}$, and this will follow F distribution and we say that this is a F distribution with m n degrees of freedom or m and n degrees of freedom, and this is also called as Fisher's F distribution and that is indicated by writing here X follows $F_{m.n}$, sometime you will see people also write in the bracket also, in the parenthesis also m comma n, so whatever is convenient to you can use it.

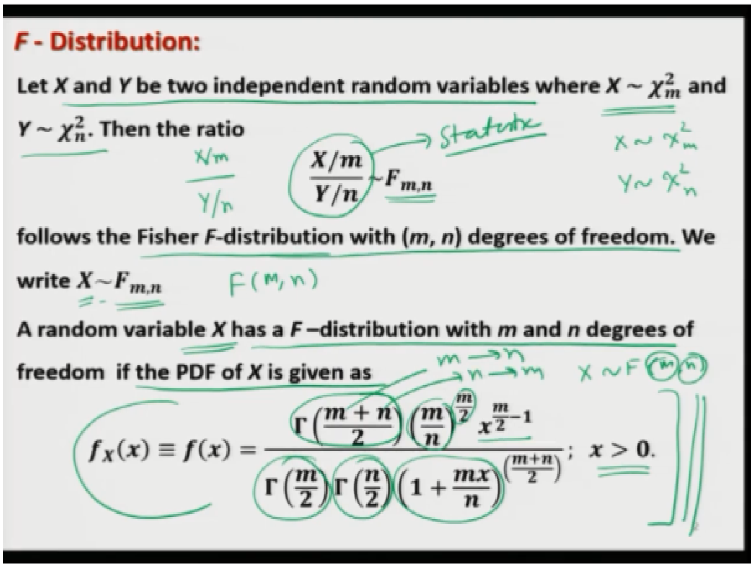
So, now the question is what will be the distribution of this statistic, you can see here this is a statistic, why, because this is a function of random variable. So, if a random variable we have that is suppose X then this X has a F distribution with m and n degrees of freedom if its probability density function is given as like this, you can see once again here this is a little bit complicated expression but as I said earlier do not need to get worried about this complicated form because we are just going to use the information that X follows the F distribution with m and n degrees of freedom and what are the values of m and n that will help us.

But you must know what is this form so this form here

$$f(x) = \frac{\Gamma\left(\frac{m+n}{2}\right)\left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1}}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)\left(1+\frac{mx}{n}\right)^{\left(\frac{m+n}{2}\right)}};\ \ x > 0.$$

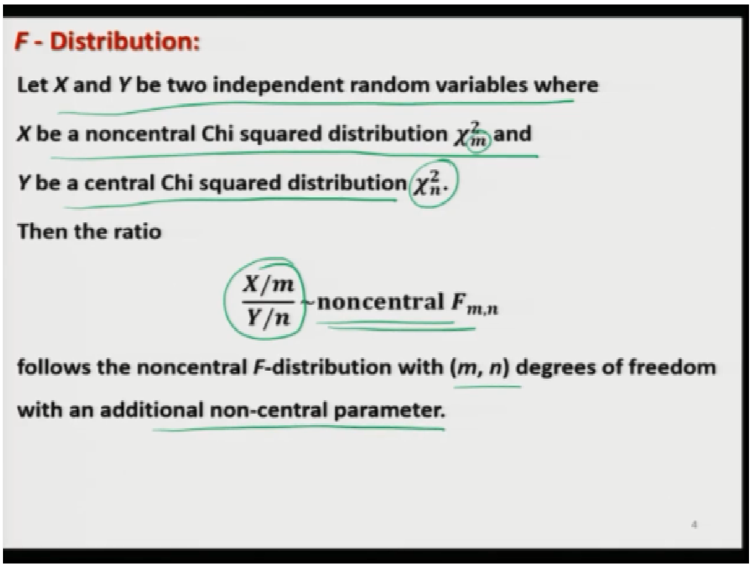
(Refer Slide Time: 4:08)

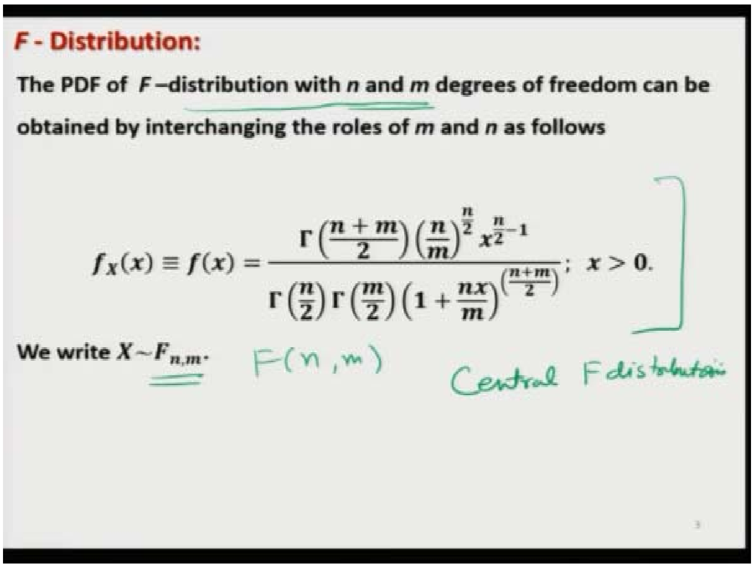




And in case if you try to interchange the degrees of freedom then the PDF can be obtained just by changing the degrees of freedom in this form which you have, so for example, the PDF of F distribution with n and m degrees of freedom can be obtained from this PDF which is for m and n degrees of freedom just by interchanging the roles of m and n, that means wherever you have here m try to replace by n and wherever you have here n try to replace by here m, that is all, so once you try to do it here you will get here this PDF and this will be indicated here X follows F n, m or F n, m like this, whatever you want.

(Refer Slide Time: 5:02)



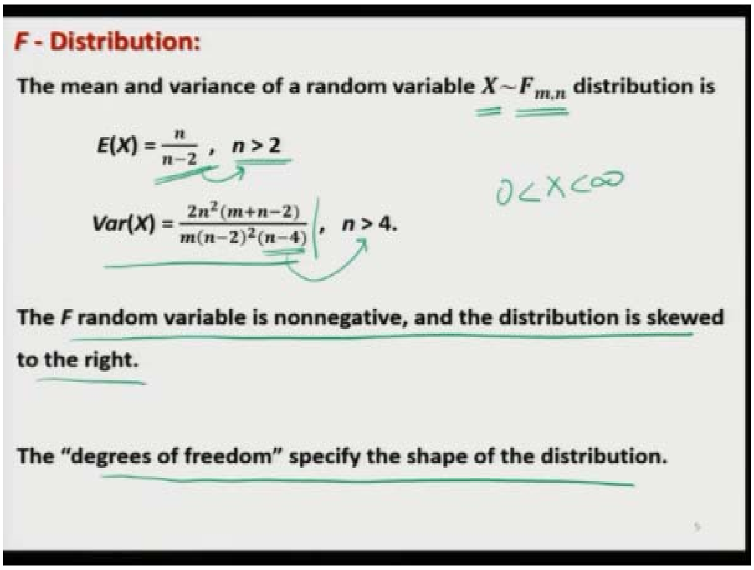


So, and this is actually called as what you have seen here or here, they are actually central F distributions, so once you have central F distribution then there will also be a non central F distribution and you have seen that in the case of chi square and t also you have a non central chi square and non central t distribution.

So, the distribution of non central F is very simple. Suppose there are X and Y are two independent random variables where X is a non-central chi-square distribution with m degrees of freedom and Y be a central chi square distribution with n degrees of freedom, that means any off between X and Y one of them has to be a non central chi square, then the same ratio that $\frac{X/m}{Y/n}$ that will follow a non central F distribution with m and n degrees of freedom.

And obviously there will be one more parameter here, that will be the non centrality parameter and if you try to choose this non centrality parameter to be 0 then the distribution will become simply central, form of this non central is also little bit complicated but then anyway we are not going to consider it and if you want to use it then in the R software if you try to look into the different options from the help you can see there is a parameter like ncp and that is also true for non central chi square or t distribution also.

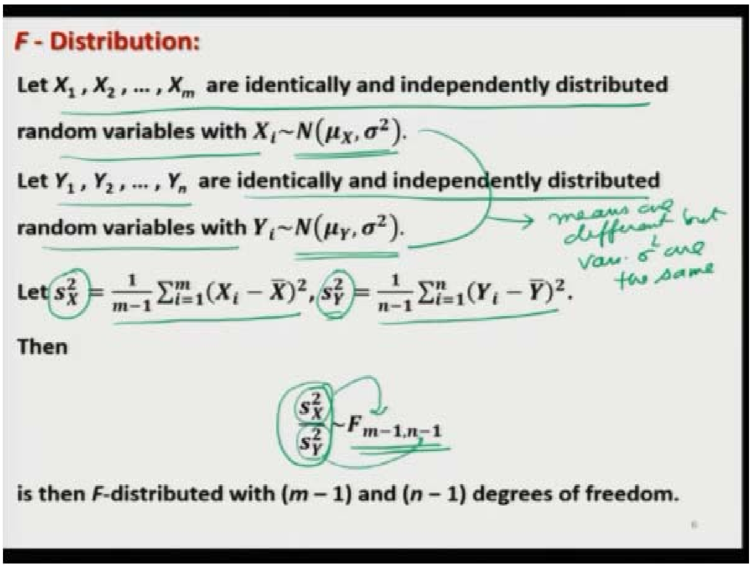
(Refer Slide Time: 6:43)



So, if you try to find out the mean and variance of this F distributed random variable then suppose, X is now following a F distribution with m and n degrees of freedom then the expected value of X will come out to be n/(n-2) and variance of X will come out to be here like this, so here you have here n minus 2 so we have to assume here n is greater than 2 and the variance of X is given by here like this, so here you have a quantity n minus 4 then you have to assume here n greater than 4, why? Because X is taking the values between 0 and infinity, so that is why the mean of our random variable which is always taking the value between 0 and infinity, the mean cannot go to in negative.

So, the F random variable is non-negative but the distribution of F is skewed to the right. Now you know what is skewed, you already have done coefficient of skewness, cortexes etcetera, and you have here the degrees of freedom and they specify the shape of the distribution. The difference between chi square t and F distribution with respect to the degrees of freedom, is that in the case of f there are two degrees of freedom whereas in the case of chi square and t there is only one degrees of freedom, that is characterizing the distribution.

(Refer Slide Time: 8:15)



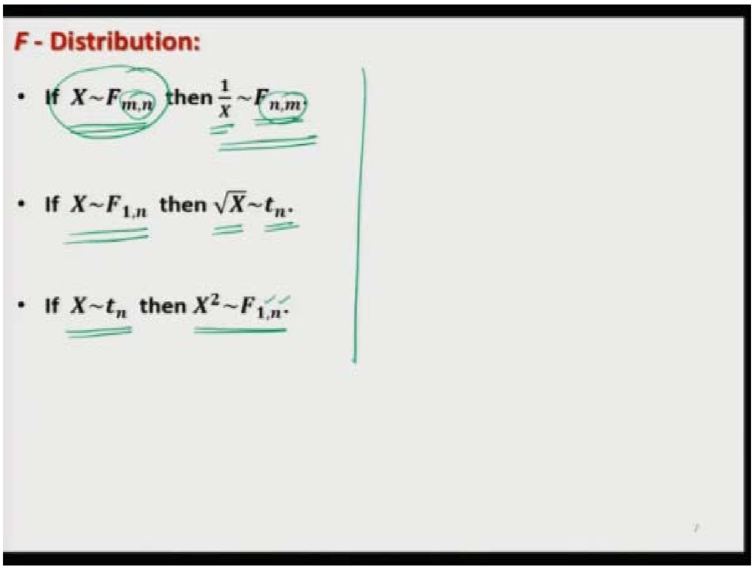
So, now let us try to consider some important results because of which F has become very popular in statistical inference area and it helps us a lot in taking different types of statistical conclusions, statistical inferences based on test of hypothesis and other things. So, suppose, $X_1, X_2, \ldots, X_m$ they are identically and independently distributed random variables and following $N(\mu_X, \sigma^2)$ and let there be another set of variables $Y_1, Y_2, \ldots, Y_n$ which are also identically and independently distributed random variable following a $N(\mu_Y, \sigma^2)$.

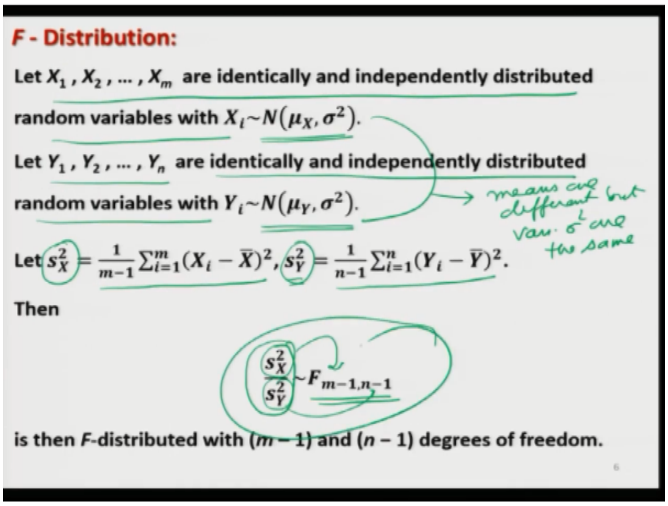
Now you have to observe here what is happening, here the means are different but variances that is sigma square are the same, so that is a very important condition, that we are assuming that the variance of both the populations of X and Y they are the same, mean may be different.

So, then based on these X's and Y's we try to compute the sample variances, so sample variances based on X is indicated by $s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$and the sample variance of Y is indicated by $s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$, then here comes the final result that the distribution of $\frac{s_X^2}{s_Y^2}$, that is the ratio of sample variances of X and Y, this follows a $F_{m-1,\, n-1}$, that is the very important result which is going to help us when we want to draw different types of statistical conclusion based on the data.

So, what you have to remember the ratio of the sample variance follows a F distribution and whatever are the degrees of freedom of the random variable whose sample variance is in the numerator that will come first and the random variable whose sample variance is in the denominator that will come later.

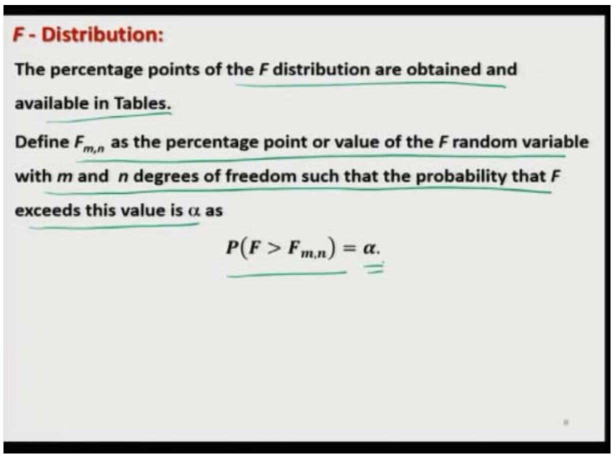
(Refer Slide Time: 10:47)





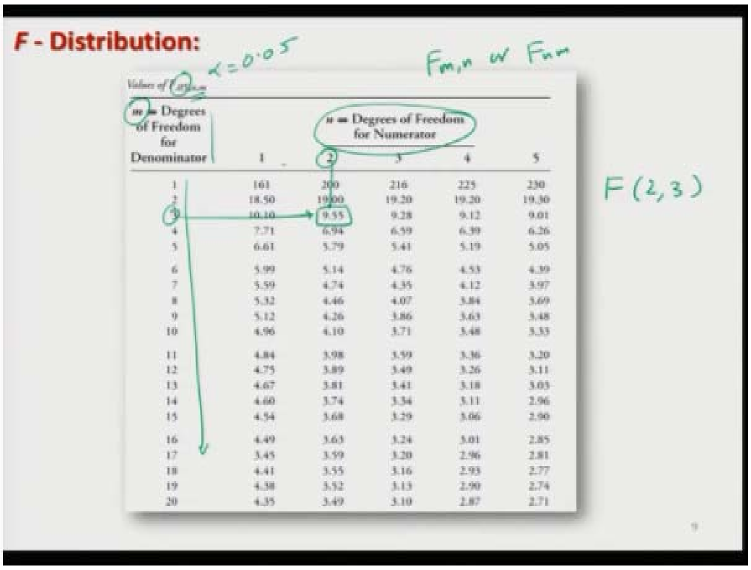
Now, some important result that if X follows a F(m, n) degrees of freedom then 1 upon X will also follow a F distribution, only their degrees of freedom will be interchanged. So earlier it was m n, now it will become n, m, be careful about my voice, that in the first case $X \sim F_{m,n}$ and in the second case $\frac{1}{X} \sim F_{n,m}$, and in case if $X \sim F_{1,n}$ and n remain as such n then the distribution of $\sqrt{X} \sim t_n$. And if $X \sim t_n$, then $X^2 \sim F_{1,n}$. So these are some results which sometime you may need, well all this results including the distribution of the random variable of F distribution like as here X/m and Y/n, like this one or their mean variance etc. or even this result also or this result, they are based on strong mathematical proofs and they are available in the statistics book, but definitely I am not considering them here.

(Refer Slide Time: 12:22)



And just like the case of chi square or t distribution the probabilities in the case of F distribution they are also obtained and they are available in the tables, and actually earlier before the use of software people use to use this tables for getting different types of probabilities because the computation of integral is quite complicated. So, if I define that F m, n as the percentage point or value of F random variable with m and n degrees of freedom such that the probability that F exceeds this value is α, that means F is greater than F m, n and then the probability of such an event is α.

(Refer Slide Time: 13:06)



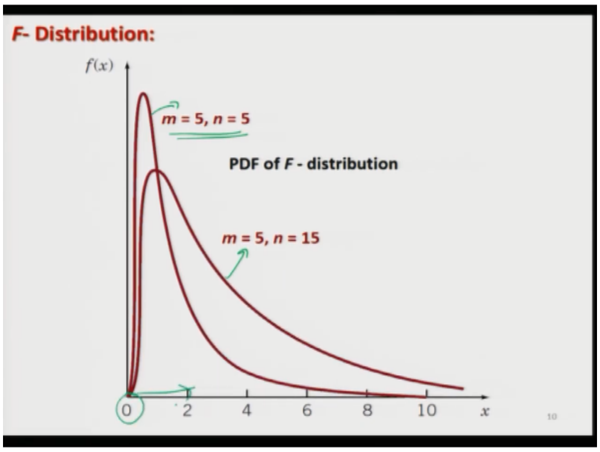
Values of $F_{0.05,n,m}$

m = Degrees of Freedom for Denominator	n = Degrees of Freedom for Numerator 1	2	3	4	5
1	161	200	216	225	230
2	18.50	19.00	19.20	19.20	19.30
3	10.10	9.55	9.28	9.12	9.01
4	7.71	6.94	6.59	6.39	6.26
5	6.61	5.79	5.41	5.19	5.05
6	5.99	5.14	4.76	4.53	4.39
7	5.59	4.74	4.35	4.12	3.97
8	5.32	4.46	4.07	3.84	3.69
9	5.12	4.26	3.86	3.63	3.48
10	4.96	4.10	3.71	3.48	3.33
11	4.84	3.98	3.59	3.36	3.20
12	4.75	3.89	3.49	3.26	3.11
13	4.67	3.81	3.41	3.18	3.03
14	4.60	3.74	3.34	3.11	2.96
15	4.54	3.68	3.29	3.06	2.90
16	4.49	3.63	3.24	3.01	2.85
17	3.45	3.59	3.20	2.96	2.81
18	4.41	3.55	3.16	2.93	2.77
19	4.38	3.52	3.13	2.90	2.74
20	4.35	3.49	3.10	2.87	2.71

So, these values are tabulated in the form of different tables for different values of α. For example, you can see here this is here the value of α is equal to 0.05 and then it is giving here the degrees of freedom here n and m, well in case if you are trying to use this a table you have to see whether the tables are in the form of your so called $F_{m,n}$ or $F_{n,m}$, what is the degrees of freedom in the numerator and denominator, so you have to use it accordingly.

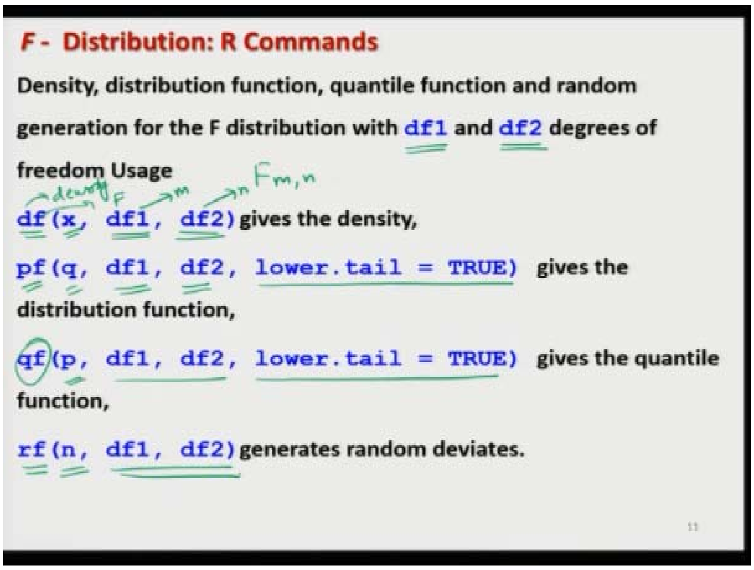
So you can see here that here in this case they are trying to indicate by m the degrees of freedom for denominator and n for the degrees of freedom for the numerator, so they are given as 1, 2, 3, 4, 5 etc. and they are also given here like this and if I want to find out here the value of F(2 , 3), so that means I have to look here 2 and here 3 and this value comes out to be here like this 9.55. So this is how we used to use the table but anyway we are going to use only the R software now.

(Refer Slide Time: 14:14)



And if you try to plot this F distribution, so for example, I have given you here two example, this curve is for m is equal to 5 and n is equal to 5 and this curve is for m equal to 5 and n equal to 15, so you can see that as the degrees of freedoms are changing the shape of the curve is also changing, and all the value they are starting from 0 that means all the values are lying between 0 and infinity. So this is how the curve will look like and you can see that this is not the symmetric curve.

(Refer Slide Time: 14:49)

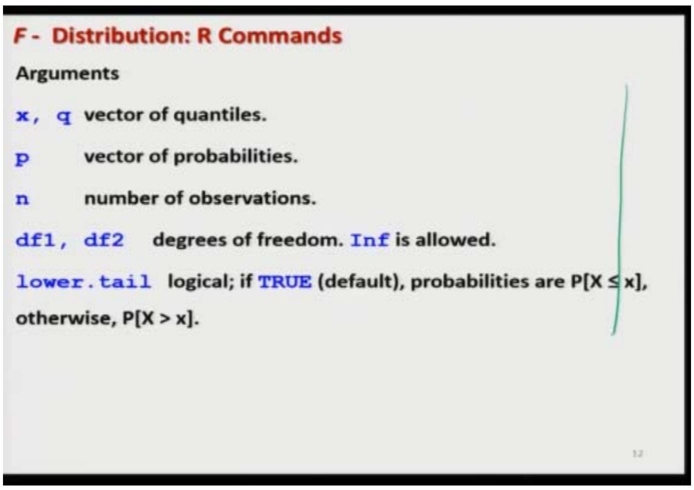


Now, in case if you try to use the R software for computation of different probabilities, quantiles, CDF, etc., from the F distribution then there are two parameters which are need to be specified and those parameters m and n which are the degrees of freedom they are indicated by here df1 and df2. So, in case if you want to find out the density of F, then you have to give it the command here df.

So, as earlier we understood d means density and f is the F distribution and then you have to give here the value x and then you have to specify both the degrees of freedoms as df1 and df2. So, if you are trying to write down here F(m, n) then df1 is going to be m and df2 is going to be n, and if you want to compute the CDF at some point q then you have the command pf and then you have to give here the value of q, df1, df2 and then if you want to use the option lower dot tail is equal to TRUE you can use it according to your need.

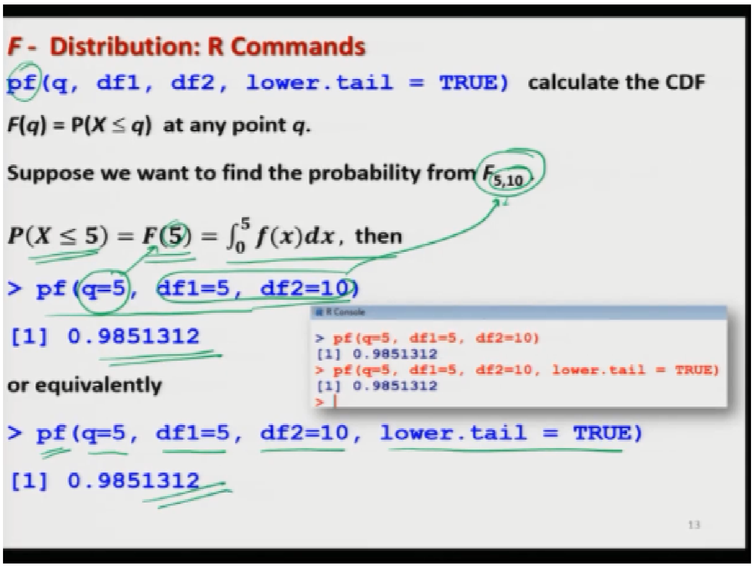
And similarly, if you want to find out the quantile then the command here is qf, so you try to give the quantile that you want to find as p and then df1, df2 and then according to the requirement you can use the lower dot tail to be TRUE or FALSE. Similarly, if you want to generate the random numbers from F distribution then the command here is rf and then you have to specify the number of observations that you want to generate by n and df1 and df2 they are the degrees of freedom.

(Refer Slide Time: 16:23)



And, yes, these are the same abbreviation which I just explained you.

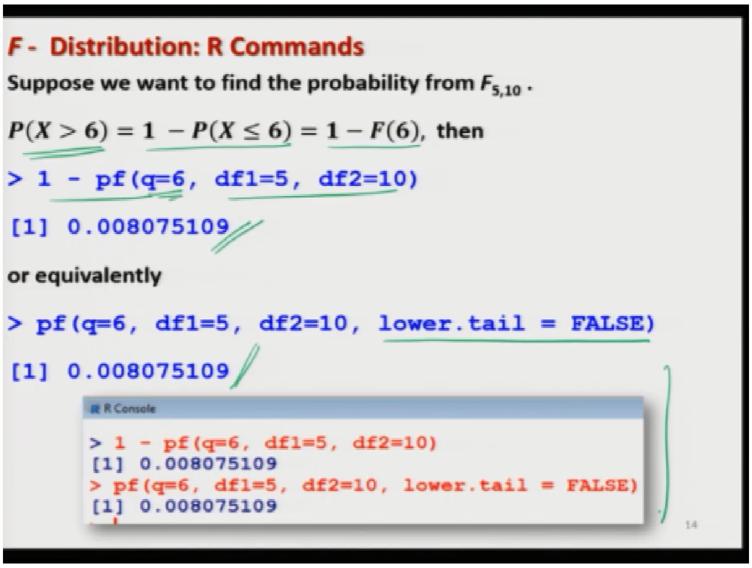
(Refer Slide Time: 16:26)



So, now we try to take some examples and we try to compute different types of probabilities, quantile, function and random numbers we try to generate. So, suppose, we want to compute the probability that X is less than or equal to 5, where X is following the F distribution with 5 and 10 degrees of freedom which is given here. So now if you want to compute this 5 you have to write it as in the form of CDF that is the most simple option and this is simply integral 0 to 5 fx dx and if you want to compute this CDF you have the command here pf inside parenthesis q equal to 5 and degrees of freedom 1 is equal to 5 and df2 is equal to 10.

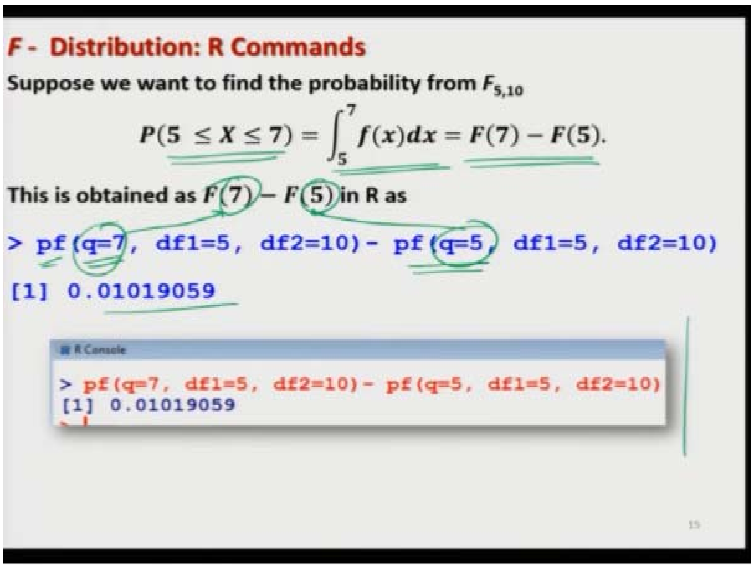
So, you have to understand how these values are coming so this q equal to 5 this is coming from here this 5 and this df1 and df2, 5 and 10 they are coming from these degrees of freedom 5 and 10. Once again you will get the same value after execution of pf q equal to 5, df1 equal to 5 and df2 equal to 10 like this, so you can see here it is not difficult to compute such probabilities.

(Refer Slide Time: 17:38)



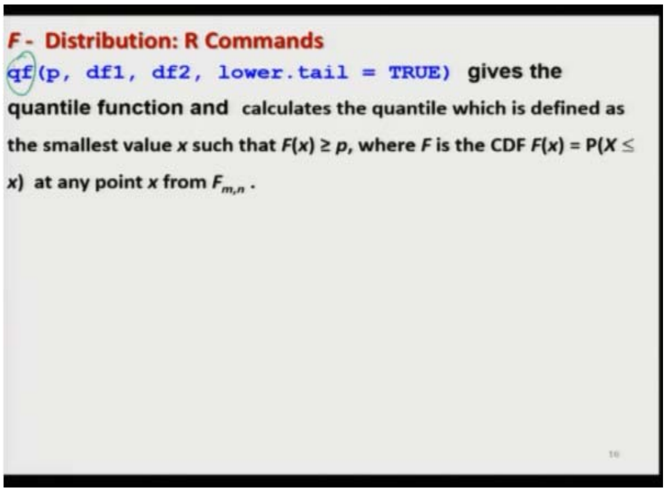
And similarly, if you want to compute a probability of this type probability X greater than 6, so that can be computed by 1 minus probability of X less than equal to 6 which is nothing but your 1 minus F(6), this can be computed by 1 minus pf q equal to 6 and you have to specify the df1 and df2 and you will get here this value and if you want to use here the option lower dot tail is equal to FALSE then you get here the same probability here and this is here the screenshot I will try to show you it on the R console also.

(Refer Slide Time: 18:08)



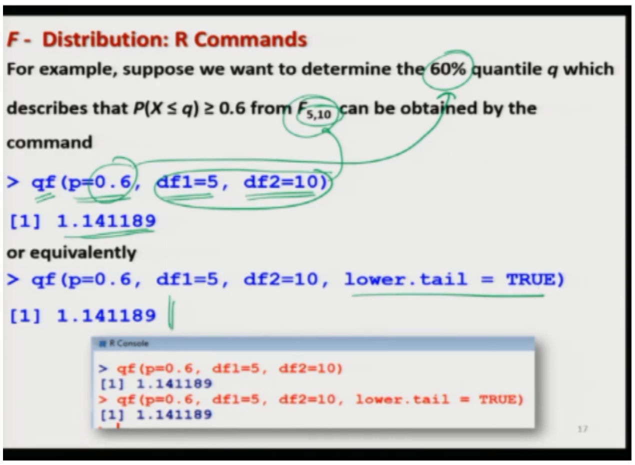
And if you want to compute the probability that X is lying between 5 and 7 so then this can be obtained by the $\int_5^7 f(x)dx$ and this can be written as F(7) minus F(5) and now both this CDF's can be computed by the command pf with q equal to 7 and pf with q equal 5, df1, df2 they remain the same. So, what is this q equal to 7 this is coming from here, what is this q equal to 5 this is coming from this 5, and if you try to execute it you will get here this probability. So, you can see here it is not difficult at all in R to compute and this is here the screenshot of the same operation.

(Refer Slide Time: 18:49)



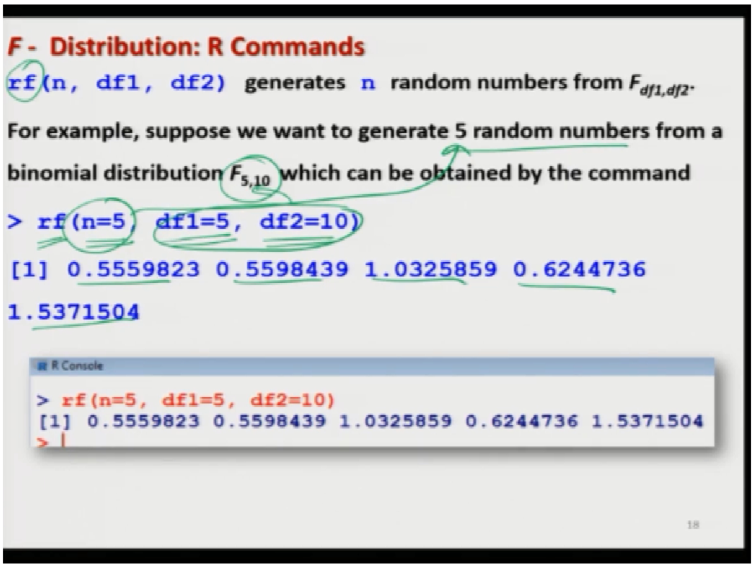
In case if you want to compute the quantiles in the F distribution then the command here is qf.

(Refer Slide Time: 18:57)



And suppose if you want to compute the 60 percent quantile on the same distribution F, so we have to write down here qf then p is equal to 0.76, then df1 is equal to 5 and df2 equal to 10, so this 0.6 is coming from this here 60 percent and df1 and df2 they are coming from these degrees of freedom 5 and 10, and this value will come out to be here 1.14 and similarly, if you want to use here the command here lower dot tail is equal to TRUE the wave is the default option once again you will get the same value here.

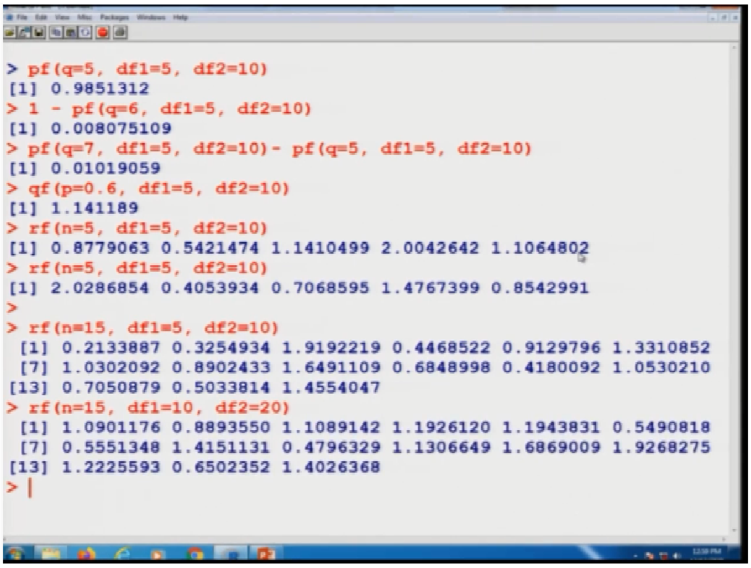
(Refer Slide Time: 19:34)



And similarly, if you want to generate the random numbers from F distribution then my command here is rf and suppose, if you want to generate 5 random numbers from the same F distribution 5 and 10 degrees of freedom, so I can write down here rf n equal to 5, that means 5 random numbers df1 equal to 5, df2 equal to 10, so this n equal to 5 is coming from the requirement that we want 5 random number df1 and df2 they are coming from the specification of the F distribution and you can see here you are getting here 1, 2, 3, 4, 5 values and all are greater than 0 so that is the property of the F distribution.

So, now you can see here that it is not difficult to execute these things on the R console but now I will try to show you these things on the R console which is a very simple thing whatever I have computed here I already have reported it here but I just want to make sure before you that these commands are working.

(Refer Slide Time: 20:38)



So, now if you want to compute the probability X less than equal to 5, then you can see here it will come out to be like this without any problem and you do not need any extra package for the computation of F probabilities. Similarly, if you want to compute the probability X greater than 6 you can see here this comes out to be like this, this is the same value which I have reported here.

Similarly, if you want to compute the probability that X lies between 5 and 7, so it is the same expression and if it will give you the value 0.01 which is reported here and similarly if you want to compute the 60 percent quantile then the command here is qf and this will give you the same value which I reported.

And if you want to generate here the random numbers from this distribution you can see here the suppose you want to generate five numbers from df1 equal to 5 and df2 equal to 10, then you are getting here the same values and definitely these values are going to be different from the values which I have reported in the slides because these are random numbers and even if you want to change it you can see here you will get here different set of random numbers, and even if you want to change the value of here n suppose, I want 15 observation, so it will give you 15 observations and if you try to change the degrees of freedom. For example, I can make it here 10 df1 and df2 can we make here 20, you can see here this is giving you the random numbers.

So, now we come to an end to this lecture and to this topic also. Now you have understood what are the sampling distributions chi square, t and F and for F you have to simply remember which is the most important part that if you have two random variables they are independent, both are following chi square distribution, so just try to define a new random variable by the random variable divided by its degrees of freedom and then just try to take the ratio of these two transform random variables, that means random variable one divided by its degrees of freedom divided by random variable two divided by its degrees of freedom and this will follow F distribution with the respective degrees of freedom.

So and second result is about the sample variance, that if you have two samples from two different population which are differing with respect to mean, the variance will remain the same, well that is the statistical need that when we try to find out this result then we have to assume that the population variances of both the population are the same, if they are not then we have to do something else. So, but in this case you have to assume that both the variances are going to be the same and if not something else will happen but in this case their ratio will follow a F distribution.

So, this result we are going to use heavily when we are trying to conduct the test of hypothesis. So, now I have completed the chi square, t and F distributions now it is your turn, try to have a quick look, try to revise them, try to solve different types of problems from the book, from the assignment and try to learn how to execute them in the R software.

Well we are now approaching towards the statistical incurrence part and now at every point you can see that whatever we have learnt in the past that we are going to use here, how to compute the probabilities and the different types of conditions and different types of concept they are trying to characterize, different types of properties which are hidden inside the data, all those things are coming, now do you think that if you do not understand these things can you really understand the data science?

Answer is absolute no, and you will see that now we are venturing into more details and whatever results we have used here they are going to help us in getting the correct statistical tool for finding out the correct statistical inference that is our basic objective. So, you try to revise this thing try to understand them and I will see you in the next lecture till then good bye.