Multivariate Procedures with R

Prof. Shalabh

Department of Mathematics and Statistics

IIT Kanpur

Week – 06

Lecture – 26

Chi square, t and F distribution

Hello friends, welcome to the course multivariate procedure with R. So, up to now you can recall that we have discussed the concepts of probability function, probability distribution function, probability mass function etc. And in each of this probability distribution there are certain parameters that are involved. And whenever we are trying to use them one of the basic objective in statistics is to know the value of those parameters. Similar to them we have three more distributions which are not dependent on any such parameter. For example, normal distribution depends on the mean, mu and variance sigma square.

So, similarly we have three such probability density functions chi square T and f which do not depend on any such parameter and they are actually called as sampling distributions. Why they are called as sampling distributions? After we finish this topic we are going to consider some more topics for example a test of hypothesis. So, there you will see that the probability distribution of the statistics which is going to conduct the test of hypothesis comes out to be for example chi square T or f. So, these three distribution play a very important role when we want to draw the statistical inference based on a sample of data set.

Although my objective is not here in this lecture to give you a complete background about these three sampling distribution, but I want to give you that much information which is required for our that is required for us to understand the topics in the lectures in the further course. So, with this objective I will try to take up the concepts of chi square T and f sampling distribution. I will try to give you some basic idea so that you are prepared to understand the basics in the further lecture. So, let us begin our lecture. So, one concept which is very important in statistics to understand the statistical inference is the concept of statistic.

You have to be careful there is a word statistics. So, we are not calling it this statistics, but we are calling only here as a statistic which is without this here s. So, what is this? So, if you assume that x1, x2, xn be a, denote a sample on a random variable capital X and if you try to take any function of such a random variable x1, x2, xn then and if you try to denote it by a capital T then T is called a statistic. So, essentially statistics is a function of random variable. So, obviously it is also a random variable.

So, once a random variable has a probability distribution so its function will also have a sampling distribution. And once you try to draw a random sample and suppose based on the values of the random variable which are suppose denoted by here say small x1, small x2, small xn and if you try to compute here this capital T then capital T is indicated by small t lowercase alphabet and it is called as a realization of capital T. So, it is something like this you are simply trying to substitute x1 is equal to small x1, x2 is equal to small x2 in your capital T and then you are trying to find out its value. So, small t will indicate a numerical value. For example, in case if I take a function of a random variable in the format of sum of xi's means x1 plus x2 plus xn all x1, x2, xn they are the random variable but capital T is a function of random variable.

So, this is called a statistic. Similarly, if I try to take a function x minus mu upon sigma where mu and sigma are known then now this is also a statistics say indicated by capital T. Similarly, if you try to take the example of arithmetic mean, arithmetic mean is defined as 1 upon small n summation i goes from 1 to small n, xi. So, this arithmetic mean is also a function of random variables x1, x2, xn so this is also a statistic. Similarly, if you try to take the sample variance which is defined by here 1 upon n minus 1 summation i goes from 1 to n xi minus x bar whole square this is also a function of x1, x2, xn so this is also a statistic.

$T = \sum_{i=1}^{n} X_i$ **is a statistic.**

$T = \frac{X-\mu}{\sigma}$ **is a statistic only when μ and σ are known.**

$T = \frac{1}{n}\sum_{i=1}^{n} X_i$ **is a statistic.**

$T = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})^2$ **is a statistic.**

$T = max.\,(X_1, X_2, \ldots, X_n)$ **is a statistic.**

$T = min.\,(X_1, X_2, \ldots, X_n)$ **is a statistic.**

Similarly, if you try to take the maximum of x1, x2, xn and minimum of x1, x2, xn they are also a random variable so they are also statistic. So you will see that now we will be using this term statistic very frequently and then you have to simply understand that it is a function of random variable. Now we talk about another concept which is about sampling distributions. So, these sampling distributions are the distribution of certain statistics which do not depend on any parameter. And these sampling distribution they are basically the theoretical distributions and they play a very important role in the construction and development of various statistical tools which are used for drawing statistical inference.

And these sampling distributions are the who can say probability functions of some statistics and they are called a sampling distribution. And the probability distribution of a statistics is called a sampling distribution. So, we are going to understand here three popular sampling distribution which are here chi-square which is written as a CHI chi-square distribution, then t-distribution and then capital F-distribution. So first let me try to give you here some basic fundamental about the chi-square distribution. So now you have to understand how we are going to create a statistics whose distribution is going to be called as chi-square distribution.

So let me try to consider here say Z1, Z2, Zn which are say small n independent and identically distributed following the normal 0, 1 distribution that means mean is equal to 0 and variance is equal to 1. So this Z1, Z2, Zn they are my small n random variables from normal 0, 1 that means each of this Zi follows normal 0, 1. Now if I try to consider the sum of their squares that is summation i goes from 1 to small n Zi square then the probability distribution of summation Zi square is chi-square distributed and it is associated with say degrees of freedom. So we say that i goes from 1 to n Zi square is having a chi-square distribution with small n degrees of freedom and it is written like this. So first you write here chi then square and then here n.

So that will indicate a chi-square distribution with n degrees of freedom. And this random variable has the following probability density function which is here indicated here like this. This random variable has the probability distribution function like this which is 1 upon 2 is power of n by 2 gamma n by 2 x is power of n by 2 minus 1 exponential of minus x by 2 when x is greater than 0 and 0 otherwise. One thing I can means explain you here for your understanding that we will not need to remember this form of the probability density function. We simply have to basically remember that this statistics i goes from 1 to n Zi square follows a chi-square with n degrees of freedom where each of the Zi follows normal 0, 1.

So in simple words I can say that sum of squares of the standard normal variate is chi-square distributed and it is indicated by here like this x follows say here chi-square with n degrees of freedom. So this is how it is indicated. And suppose if the random variables do not have normal 0, 1 distribution but they have got normal mu sigma square. So if I say that let x1, x2, xn be independent and identically distributed random variables which follows normal mu sigma square in that case what we have to do? Somehow we have to change this variable in such a way whose distribution is normal 0, 1. So that if I try to take here this xi minus its mean divided by its standard deviation then these are going to be normal 0, 1.

So what we try to do here then can we try to find out the sum of squares of this xi minus mu upon sigma. So it is here like this summation i goes from 1 to n xi minus mu upon sigma whole square will then follow a chi-square distribution with n degrees of freedom and it is indicated again by chi-square n. And this is actually more precisely called as central chi-square distribution. So the next question comes what is then non-central? So if I try to say that if suppose mean of any of the random variable x1, x2, xn is not equal to 0 then whatsoever be the distribution of summation xi minus mu upon sigma whole square that will be chi-square but that will have one more parameter that is called a non-centrality parameter and that distribution will be non-central chi-square distribution. So briefly I can tell you if any of the random variable has got a non-zero mean then the sum of squares of those random variables will follow a non-central chi-square distribution.

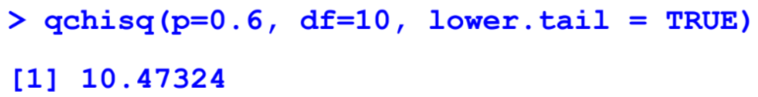
And just for your information we are not going to find out the mean of this chi-square random variable with n degrees of freedom is n and its variance is twice of n. And this chi-square distribution is not symmetric like as normal and it can realize only those values which are greater than or equal to 0. And this concept this degrees of freedom it is a well-defined concept we are not going to discuss here about it but I will simply show you the use and application of this concept or degrees of freedom. So in the chi-square random variable the choice of degrees of freedom will specify the shape of the curve of the distribution. For example, if you try to seek here n equal to 2 then you will see here this curve here is like this.

And in case if you try to choose here n equal to 5 here then you will see here that its curve is here like this. And in case if you try to choose here n equal to 10 then its shape is here like this. So you can see here depending on the degrees of freedom the shape of the curve changes and now you can imagine that once the shape of the curve is changing so the different probabilities will also change depending on the value of small n which is the degrees of freedom. And in statistics there are some tables which are popularly called as chi-square tables. So the percentage points on this chi-square distribution have already been obtained and they are available in so called chi-square tables.

But anyway now we are going to find them only on the R software also. But it is important for you to understand what these values are going to indicate. So in case if I define this chi-square n as the percentage point or the value of the chi-square random variable with n degrees of freedom such that the probability that chi-square exceeds this value is alpha like this. Probability that chi-square exceeds the value of chi-square n and this probability is here alpha. So, this is going to be simply here the integral of this probability density function of this chi-square and that is integrated over the range chi-square n to infinity.

But you do not need to compute it, the tables are available and they can be obtained directly into the chi-square command in the R software. So, for example, if you want to find out such quantiles in the R software then the command here is q ch i s q and then inside the parenthesis you have to specify the vector of probabilities degrees of freedom as df and then lower.tail is equal to f will give you the probabilities on the lower tail side. So now if you want to compute this value on the R software then this function qchisq() it gives the quantile function and it calculates the quantile which is defined as the smallest value of small x such that f x is greater than or equal to p and f x is the pdf, cumulative distribution function defined as the probability x less than equal to x at any given point say small x and it is computed over the probability density function of chi-square distribution with say df degrees of freedom.

Say for example if you want to know the 60 percent quantile q which is explained by probability x greater than or equal to q is greater than or equal to 0.6 and with it can degrees of freedom that is df is equal to 10 then this can be obtained here by qchisq p is equal to 0.6 df is equal to 10 and if you try to execute it on the R software it will come out to be here like this 10.47324 or equivalently if you want to use here the command that lower.tail is equal to true that is the default option and it will also give you the same value here and you can see here this is the screenshot. So let me try to show you this thing on the R console so that you get here more confident for example if I try to see here this value. So you can see here this will come out to be here like this and if you even if you remove this option here lower.tail then you can see here this is you are going to get the same value. So this is here the default value.



```
> qchisq(p=0.6, df=10, lower.tail = TRUE)
[1] 10.47324
```

So I think this much of information is enough for me to explain you the concept of say this chi square and it and this is enough for us to understand the concepts in the further lecture. Now after this I come to another sampling distribution which is called as t distribution t is written as a lower case t. So now you will see that I am going to use here two probability density function which is a normal 0, 1 and say another is chi square and based on that I will try to define here a statistics and if you try to find out its probability function then that will be the t distribution. For example now if I say suppose I have here two random variables say x and y and x is following normal 0, 1 like this and y is following this here chi square with n degrees of freedom and both of them are independent.

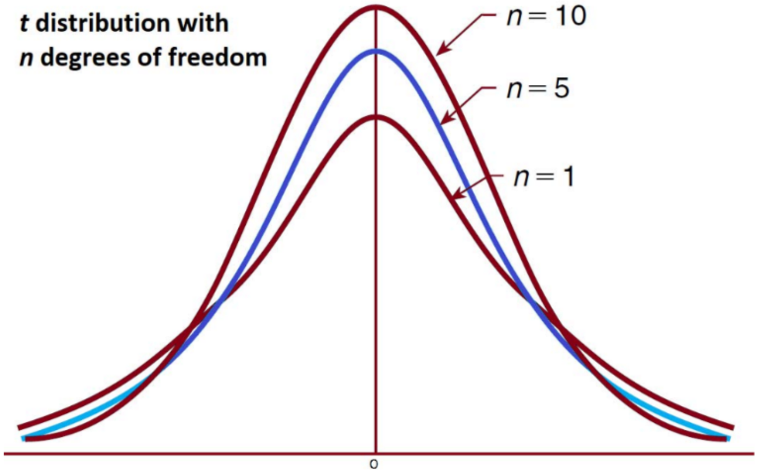
So these are independent. Then if I try to take the ratio of say x upon square root of y upon n that is here like this then this will follow a t distribution with small n degrees of freedom and this is called as central t distribution and its PDF is given by here like this. So suppose if I say that there is a random variable x which has got the t distribution if the probability density function of x is given by here like this. So you can see here this is little bit complicated structure like as gamma function of n plus 1 by 2 square root of then n pi in the denominator and gamma n by 2 in the denominator then 1 plus x square upon n raise to power of minus n plus 1 by 2 and here in this case x lies between minus infinity and plus infinity and this is written as say x follows t n like this x follows t distribution with n degrees of freedom. So similar to the concept of here central t distribution we also have a concept of non central t distribution and the same concept carries on that if any of the random variable here if this has got a mean which is non-zero then we will have a non central t distribution with some non central t parameter but since we are not going to use it so I am not explaining you here. Okay, so the mean and variance of this random variable which is following t distribution with n degrees of freedom is here like this mean is 0 and variance here is n upon n minus 2.

So this is greater than 2 so that the variance is non-negative. So this t distribution is strict the t density has a thicker tails than the curve of the normal density function so it indicates that t density has greater variability than the density of normal distribution. Right and the concept of degrees of freedom is specified the shape of the distribution and we will see that when the degrees of freedom are more than 30 when the shape of t and the normal distributions are almost the same. So then if there are more than 30 degrees of freedom then either you are using the normal distribution or t it will give you the same probability for a given region. And then in case if you suppose assume that the variance of normally distributed random variable is not known then in that case we define the t distribution in this particular way.

For example if I say that let X1, X2, Xn are identically and independently random variables following and each of this Xi is following say this here normal with mu and sigma square and suppose sigma square is unknown which is estimated by this small s square. So sigma square is here unknown and it is estimated by that is sigma square hat. Well you will come to the concept of that how do you estimate it but at the moment I am simply trying to say suppose it is unknown then on the basis of the given sample of data somehow we try to know the value of this sigma square. So it is here 1 upon small n minus 1 summation Xi minus X bar whole square. Now in case if you try to define here a statistics here like this square root of n X bar minus mu upon s then this will follow a t distribution with n minus 1 degrees of freedom.

So sometimes people ask that how this minus 1 is coming why not tn. So I would say here one simple concept is that because the sigma square is unknown so we are trying to estimate it from the given sample of data. So that is why 1 degrees of freedom is lost. And in case if you want to find out the percentage points of the t distribution then they are also available in the tables and those tables are popularly called as t tables. So in case if you define this tn as the percentage point or the value of the t random variable with n degrees of freedom such that the probability that t exceeds this value is alpha like this here probability t greater than tn is equal to alpha then the tables are available for the given value of alpha.

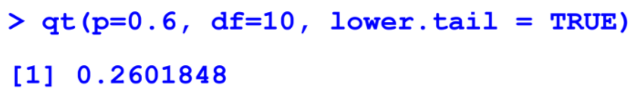
But we also try to compute it on the R software what I will try to show you. But before we try to show you that how you can compute it let me try to show you that how the curve of the t distribution with n degrees of freedom changes with the when the degrees of freedom change. For example, so degrees of freedom are suppose n equal to 1 then the curve will here look like this one you can see here. And similarly if you try to take n equal to here 5 then you will see here this is here the curve. And if you try to take here see here n equal to here 10 then you will try to see here it will be here like this.



And in case if you try to take here n equal to 30 here then you will see that here the curve of t distribution and the curve of normal distribution they will remain the same. They will become almost the same and you can see here these curves are mainly differing with respect to the hump of the curve. So as and then all of them are symmetric around you can see this here line you can see here in this line and this line the curves are symmetric. And here I am trying to show you that if you try to compare the t distribution with normal so if you see here with this line here t I am trying to show here t distribution and if I try to show you here the normal 0, 1 that it is indicated by this broken lines. So you can see here they are more or less similar and there is some difference in this part here and this difference becomes almost negligible and both the curves overlap when the degrees of freedom of the t distribution they become greater than or equal to 30.

So that is why you will see in the tables of the t probability you will not find the values of the probability beyond n equal to 30. Anyway if you want to compute these values in the R software then our command here is qt and then we try to give here the p which is here the vector of probabilities then it is that it is here degrees of freedom here df and then I try to write down here the lower.tail is equal to true so that you can find the probabilities on the lower tail otherwise you have to find out. If you want to find out the probability on the upper tail then you have to write down lower.tail is equal to false but this is here the mean default and if you want to use here dF is equal to inf that means infinity that is allowed. So now if you want to compute this value on the R software then this qt is going to provide us the value of the quantile function actually this quantile is defined as the smallest value of small x such that capital F x is greater than or equal to P where this capital F is the cumulative distribution function that is F x probability x greater than or equal to x at any point small x on the probability density function of t with df degrees of freedom. For example suppose you want to determine the 60 percent quantile q which describes probability x greater than or equal to q is greater than or equal to 0.6.

Say on the probability distribution curve with 10 degrees of freedom so df is equal to here 10. So this I can write down here dF is equal to 10 and p is equal to 0.6 and if you want to execute it on the R software you will get here the value 0.26. And similarly if you want to use here the option here lower.tail is equal to true then once again you will get here the same value and you can see here this is indicated in the screenshot also but look at me try to show you these things on the R console also.



```
> qt(p=0.6, df=10, lower.tail = TRUE)
[1] 0.2601848
```

So let me highlight this command and copy it. So you can see here this is here like this and then if you want to remove this option lower.tail is equal to true you will see here like and on the other hand if you want to if you make it here actually here false then you will see here this value will here come out to be here like this. It has no meaning actually because this dF is an increasing function. So let me clear the screen and come back to our slide and try to come to our next sampling distribution F. So this is indicated by here capital F.

So suppose now there are two random variables X and Y. X is following a chi-square distribution with m degrees of freedom, y is following the chi-square distribution with small n degrees of freedom and suppose both of them are independent. Then the ratio that is x divided by its degrees of freedom is small n and the y divided by the degrees of freedom is small n. If I try to consider this ratio then this follows a F distribution with m and n degrees of freedom. So this is also called as Fisher F distribution with m and n degrees of freedom and we write it like this here x follows F m,n. So a random variable x has a F distribution with m and n degrees of freedom if the pdf of x is given here like this.

$$\frac{X/m}{Y/n} \sim F_{m,n}$$

So once again you can see here because this is a complicated expression but as I said earlier in the case of chi-square and t distribution we are not really going to use this function. But anyway this is given here like this. This is here gamma m plus n by 2 m upon n raised to the power of m by 2 x raised to the power of here m by 2 minus 1 and in the denominator gamma m by 2 gamma n by 2 1 plus mx upon n raised to the power of m plus n by 2. This is also defined only for x greater than or equal to 0 and you know that this function is a gamma function which I have used in other places also. And the thing is and here there is one advantage that if you try to interchange the degrees of freedom.

For example, you can see here I am writing here say here F distribution with m and n degrees of freedom. But if you want to have the pdf of here F of n, m that means the degrees of freedom are interchanged then you do not have to do much but you simply try to interchange the values of n and m in the given pdf which is here and you will obtain the pdf here like this. And we indicated by here x follows F distribution with n and m degrees of freedom. Now if you try to see for this random variable x which is following F, m and n degrees of freedom the mean is mean of x is obtained by here n upon n minus 2 and greater than equal to 2 and the variance of x is obtained by here this quantity. And this F random variable is non-negative and it is skewed to the right.

And the degrees of freedom they specify the shape of the distribution but here you can see that as in the case of chi square and t there was only one degree of freedom that is n but now there are here two degrees of freedom which are here m and n. So on the other hand if you try to consider the case when the variance of the random variable is not known to us so in that case how are you going to define this F distribution and this is we are going to use in the further lecture. So if I try to say here that let x1, x2, xm they are identically and independently distributed random variables where each of this xi is following a normal distribution with mean mu x and variance my square. And similarly this y1, y2, yn that is another set of identically and independently distributed random variables with yi following normal mu y and sigma square. So you can see here they have got different means but same with the same variance.

Suppose if sigma square is unknown and which is here unknown. So this sigma square that can be estimated for each of this x and y set of random variables based on the observation available in the respective sample. So suppose I say here let s6 square be equal to 1 upon small m minus 1 summation i goes from 1 to small m xi minus x bar whole square and let sy square be equal to 1 upon small n minus 1 summation i goes from 1 to n yi minus y bar whole square. They are going to estimate the variances sigma square in the respective population. Then if I try to take here the ratio of here sx square upon sy square then it will follow a F distribution with m minus 1 and n minus 1 degrees of freedom.

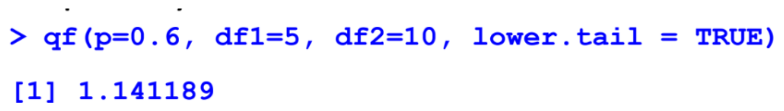
Let $s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2,\ s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$

Then

$$\frac{s_X^2}{s_Y^2} \sim F_{m-1,n-1}$$

And it is indicated here like this. So once again you can see here that this 1 degrees of freedom in m and n has been lost because we are trying to estimate the sigma square in both the populations. And similar to chi square and t tables we also have the f tables where the percentage points of the F distribution have been obtained and that is actually indicated here as say f mn. So if I try to define f mn as the percentage point or the value of the f random variable with m and n degrees of freedom such that the probability that f exceeds this value is alpha that is here like this. Then these values have been tabulated for different values of m and n alpha.And here I can show you that how the curve of probability distribution function of f will look like if I try to take here m is equal to 5 and n equal to 15 it will here like this.

And if I try to change here this degrees of freedom for example m equal to 5 and n equal to 5 then it will here look here like this you can see. You can follow my here pen in blue colour. The quantile function for the F distribution with df1 and df2 degrees of freedom like as here df1 is equal to equivalent to m and df2 is equivalent to here n. They can be obtained from in the R software by the command qf which is here small p. We have to define as the vector of probabilities df1 and df2 they are the degrees of freedom something like here small m and small n and then we have to use here the command lower.tail is equal to true which is the default command and in this df1 and df2 this inf that is infinity is also allowed.



```
> qf(p=0.6, df1=5, df2=10, lower.tail = TRUE)
[1] 1.141189
```

So you can see here these are the thing of this command. Now I try to show you that for example if you want to find out the 60 percent quantile q which describes that the probability that x less than equal to q is greater than equal to 0.6 from this F distribution with 5 and 10 degrees of freedom then this can be obtained by the command here qf then p is equal to 0.6, df1 is equal to 5 and df2 is equal to 10 and you can obtain here this value 1.141189 on the R software. And similarly if you try to add here this lower.tail is equal to 2 then again you will get here the same value which is the default value, right. And this is here the screenshot but let me try to show you this thing on the R console also, right. So you can see here this is here like this and if I try to remove here this option lower.tail is equal to 2 again you get the same value here, right. So now we come to an end to this lecture. So now you can see here in this lecture we have covered the aspect of sampling distributions and yeah we are going to use them in different context in the lectures in this course and that is how they are going to be useful.

One good part is that you can see that they are not depending on any such parameter like as normal used to depend on mu and sigma square similarly binomial depends on N and P etc. So sometime they are also called as nonparametric distributions because they are they do not involve any parameter. So at the moment my request to you all is that although I have covered this topic quite briefly but it will be good if you can have a look into the book and try to understand that how they have been obtained and what are their different properties. The more you learn better you will understand this subject and then how to generate the random numbers, how to generate different types of quantities in the R software using chi square T and F distribution if you try to understand they will help you. So you try to practice it and I will see you in the next lecture till then goodbye. Thank you.