Data Analysis and Decision Making – I

Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 13

Welcome back my dear friends, very good morning, good afternoon, good evening to all of you. And I am Raghu Nandan Sengupta from the IME department, IIT Kanpur. And this is the DADM which is the Data Analysis and Decision Making as you can see course number 1 under the NPTEL MOOCs series. And this is a 12 week course, total duration make 30 hours. So, we will have each week 5 lectures, each lecture being for half an hour.

So, this is the 13th lecture which means we have already completed 2 weeks, which can be basically spent lectures and we are in the third week. So, we remember the last day we are discussing, and I was there on one slide for a long time and I was basically tried to dwell into the idea that given any random variables; obviously, you can find its first movement which is the mean and obviously, other information's about the distribution like median, mode all things could be found out.

Then the second moment was the variance. Now, given two random variables; obviously, they would be a interrelationship between these two random variables. So, if there are n such a random variable small n, and this small n has nothing to do with the sample size please remember that; I will be using is interchangeably between the concept of small n and capital n or concerns considering, I may use the value small k or small p, but that will depend on the problem.

So, given n number of random variables being denoted by capital suffix X 1 to capitals suffix X 2, and as you know capital X is are the random variables and their corresponding realized value would be small x. So, they would be a correlation coefficient existing between these two random variables, and as I mentioned it can be proved; I am not going to go into the proof that two random variables have a correlation coefficient, and that value of correlation coefficient basically is between minus 1 to plus 1. And if it is basically in from minus 1 to less than 0, which is negative; then the relationship between two this, two random variables if you are trying to plot, I have drawn this diagram, but I am just repeating it, and you do it in the last slide for the 12th class which was the last lecture.

So, if you are trying to draw the random variables x 1 and x 2 along the x axis and along the y axis, then the relationship if it is negative correlation, it will be in the second and fourth quadrant. Now, if the correlation coefficient is 0 between x 1 and x 2, it will be spread over the first 4 quadrants, because you do not have any relationship. And, if it is positive correlation, obviously it will be the relationship would be in between the first and the third. So, obviously the tan of the angle with respect to the x axis, x axis can be either x 1 and x 2 whatever it is, that would be positive and negative depending on whether the correlation coefficient is positive or negative.

Then I also discussed, what is the covariance's, covariance is basically the expected value if at go to the formula, it is the expected value in the bracket x minus mu x into y minus mu y and I want to find the expected value of that, that has a relationship with correlation coefficient I showed that, then also said that if you have n number of random variable. So, this I was basically initially, I was discussing with the random variable relationship for the correlation coefficient, and the correlation for two random variables 1 and 2, so that 1 and 2 can be any i i, any j between the n number of random variable.

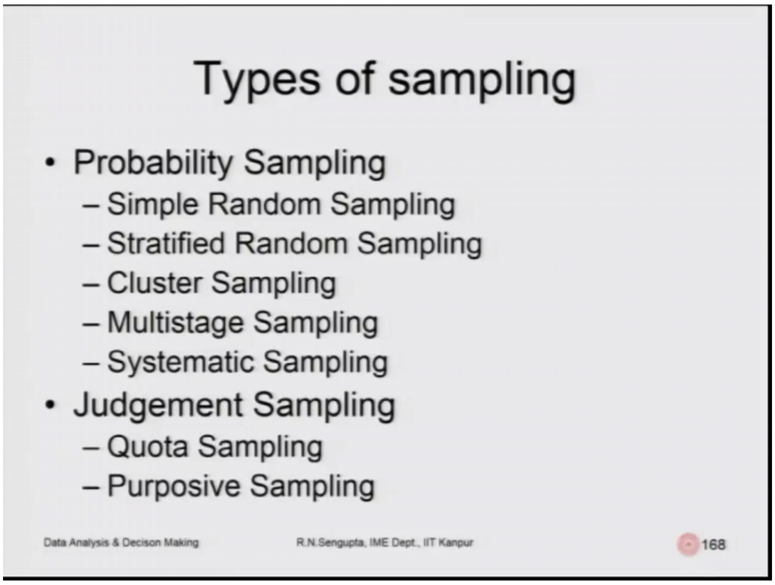
So, if you are n number of a random variable, the covariance co covariance matrix or the variance covariance matrix whichever you want to denote would be n by n, and this n what I did in the last class was basically, I considered the n number of observations for each, it was m as in mango was in or in Mangalore the number of variables. So, whether n or m it does not matter, is only the concept which I am trying to clear to all of you.

So, if you have n cross n matrix. So, the principal diagonal would be the covariance of the first the first element, 1 comma 1 would be the covariance of the first with it itself which is the variance. The 2 cross 2 elements; that means, 2 comma 2 would be the covariance of the second with it second, and similarly if you go to the last one which is n comma n place, it will be the covariance of the nth random variable with itself. And after diagonal element would all be mirror image of each other, they would be the covariance of i to j or j to i, whichever you are looking at.

Now, if you look at the correlation coefficient again in the similar way, this is a matrix of n cross n, where the principal diagonals would be 1, because that is the correlation coefficient existing between the first with itself. In the 2 cross 2 element based would be the correlation coefficient of the second with itself, similarly the n comma n would be the correlation coefficient existing between the nth variable with itself. And the off the diagonal element would be the correlations existing between i to j or j to i wherever; however, you basically trying to imply, and that can be positive or negative depending on whether the correlation coefficient is positive or negative.

And this correlation coefficient of the or the this covariance's, as I discuss they are only basically for the population. So, obviously if we have a sample, the corresponding correlation coefficient in the covariance's would be denoted like this. So, we where I did and I would not use the word digression, I have did basically go into the trying to clear the concept of correlation coefficient and covariance's, so having said that I will type start with the types of sampling and then slowly going to the sampling distribution.

(Refer Slide Time: 06:16)



Types of sampling

- Probability Sampling
  - Simple Random Sampling
  - Stratified Random Sampling
  - Cluster Sampling
  - Multistage Sampling
  - Systematic Sampling
- Judgement Sampling
  - Quota Sampling
  - Purposive Sampling

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 168

So, there a different types of probability sampling. So, one is the simple random sampling; one is the stratified random sampling. So, in the stratified random sampling, you divide the total sample in starters and basically you take proportional number of observations such that you can understand what is the overall property of the distribution considering that that holds set of status, constitute the population.

Consider the concept of a cluster and sampling, will take cluster in different areas and basically try to find out, the characteristics of the clusters that we are going to take from the super sub-sample; now the word sample, I am using in a sense that you are thinking sample, but you will take the observation, as part of the observation for the sample in such a way that the basic economic or portray the characteristics of the starters or the clusters you are going to take.

Consider this very simple example that you want to find out the overall average income of the set of people who has staying in say for example a city, like Bombay or Calcutta or Bhopal or Kanpur whatever. Now, if you consider, if you take randomly a sample from any section of the society or any locality of the city; obviously, it may mean that you are not able to portray the characteristics of the set of people, who are staying in the city in the best possible manner, because they may be very rich people in one area; they may be middle class people in one area; they may be very poor people in one area, where the overall set of observations you going to take from the rich section of the society or the middle class set of people from the society or the very poor set of people from the society, may not give you the exact data pertaining to the overall sample which you are trying to portrait.

So, say for example, if you want to take a sample of size n, you will divide the overall proportions of the sample in such a way, that you take proportional number of observations from the very rich, middle class, poor and so on and so forth from the set of the society that they give the picture, for the overall sample such that you are able to portray, the characteristics of the population to the in the best possible manner.

In the in the case of multi sampling procedure, consider this very simple example, I will I will give you an example and that will make things clear to you. Say for example, you are doing you are a marketing firm and you want to do a marketing survey, but the marketing survey enters a lot of cost. So, if you are total overall budget for the marketing is unlimited; so obviously, you can get as much samples as you take, depending on how what is the accuracy of your experiments.

So, the experiment or the test you want to do may be, say for example, you want to find out that what would be the response of a set of people from the society, who are going to buy your products; maybe it is refrigerator, maybe it is a washing machine, maybe it your cream, maybe it is for example, some product you are going to electronic product, we are going to float in the market.

Now, considered your budget for the marketing services limited; so, what you will try to do is that, you will basically start with a small set of observations, take the sample and then try to basically test the observation from the sample or take the characteristics from the sample; in such a way that you are able to get as closely as possible to the overall characteristics which you want from the population.

So, what can be the overall characteristics from the population you want to test, it can be say for example, what is the average life of the product which is going to survive; or what maybe it what is the hazard life; or maybe you would like to basically find out what is the reliability of the product. So, when you are trying to do that, you will try to basically find out the set of observation in such a way, that you are able to take the minimum number of the observation considering the overall cost, but you are able to basically minimize your as you are tried able to minimize, you will also able to basically find out that what is the best possible combination of the observation which you can take, such that you are able to predict the overall population characteristics to the maximum possible degree.

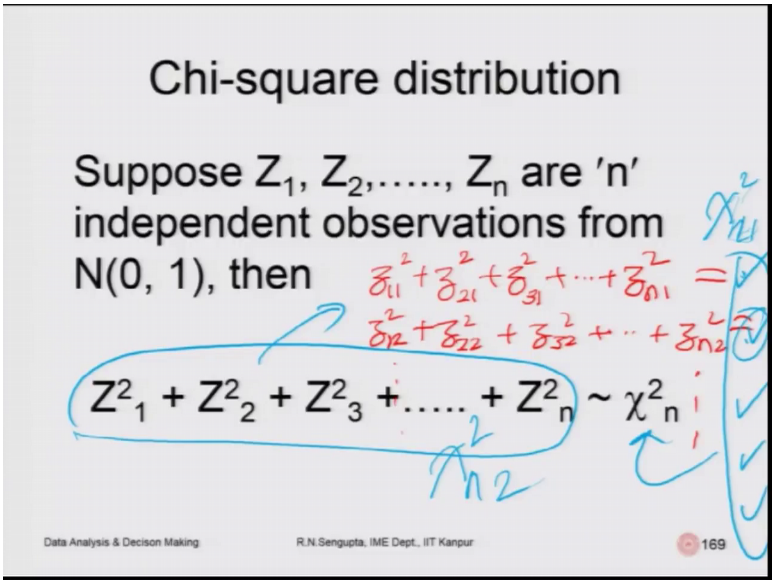
So, what you do is that you take say for example, 20 observation test the what is your stopping criteria; I am using the word, what is stopping criteria in the scientific manner, in the sense that you have stopping criteria based on which you will basically find out, what is the observed set of observation in the sample size for the first set. Then predict, how close you are to the overall population characteristics; take a decision whether you want to go forward or you want to end your experiment their.

And the multistage concept would imply that once you basically design your stopping criteria, you will continue taking set of observation in whatever set you will take, it can be one observation at a time; it can be say for example, 10 number of observation at in a at a time, it can be say for example, in a way that you are trying to increase of set of observations as you go stage by stage; or it may be possible you are trying to decrease your set of observations as you go from step to step. And you can continue taking that such that you get the minimum number optimum, I would not use the word minimum, but you take the optimum number of set of observation such that you are able to minimize your cost, but at the same time you are able to find out the characteristics of the population to the best possible manner. And what is that manner, I will come to that later on.

So, what you want to basically achieve. This judgmental sampling and which was basically will consist of quota sampling and purposive sampling would basically have some idea based on which will try to basically you take the set of observations. And the quotas would be fixed in such a way that you are able to meet your criteria, so that cause can be coming from the reset of efficiency that that cost effect can be coming from trying to minimize your overall lost whatever it is, those words of efficiency, words of laws are used being used in the in the in the statistical sense, but they can be converted into very practical sense also; and then basically you do your observations accordingly; and then plan your, your sampling distributions accordingly.

So, as I said that if you when I was discussing about the continuous distribution, I did not mention three of the distributions; though I mention them very fleetingly, they were the Chi-square; the F distribution; the t distribution would be used quite or frequently considering that I want to study the normal distribution, considering the normal distribution to be true; such that we can find out the characteristics of the normal distribution and then try to use those characteristics from the normal distribution to find about, the Chi-square, the f and the t such that we can find out some information from the sample corresponding to its mean and variance such that we are able to predict or give some a set of information's about the population parameter.

(Refer Slide Time: 12:59)



Chi-square distribution

Suppose $Z_1, Z_2, \ldots, Z_n$ are 'n' independent observations from N(0, 1), then

$z_{11}^2 + z_{21}^2 + z_{31}^2 + \cdots + z_{n1}^2 = \chi^2_{n1}$

$z_{12}^2 + z_{22}^2 + z_{32}^2 + \cdots + z_{n2}^2 = \chi^2_{n2}$

$$Z^2_1 + Z^2_2 + Z^2_3 + \ldots + Z^2_n \sim \chi^2_n$$

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 169

So, we are only interested to find out technically the population parameter being the mean or the variance. So, let us consider the Chi-square, suppose there are Z 1 to Z n and this capital Z 1 are the standard normal deviate with a normal mean of 0 and variance of 1. So, there Z 1 to Z n are the n number of independent observations which are kept, so consider this. There are n number of boxes and each box has are inside them, they have the standard normal deviate that is the whole populations is there in that box, but we denote the boxes as 1 to n.

So, if you pick up one observation from first box, that would basically denoted by initially before you know that value, it will be donated by Z capital Z 1, once you open the chit, so consider you are picking up the chit from normal distribution that value actually can be any value from minus infinity to plus infinity, when you open that chit the value you it is known to you, it becomes a realize value and it is denoted by small z 1 comma 1; that means, the first one is basically for the box number and the second one is basically the reading number.

So, if you pick say for example, that the third chit from the fifth box, you so it will be small z 5 comma 3; 5 being for the box, 3 being in the observation number. Now, what you do is that, you continue picking up chit one at a time from the box 1 to n, and you go in that sequence. So, 1 to n you pick up note down find out the value. Second time you pick up, note the values from 1 to n box or 1 to n observation note down and continue doing it.

Now, remember that the boxes have infinite sets of observation, because they are the population. So, whether you pick up a chit note it down, and again replace in the box or do not replace in the box does not matter; because the concept of simple random sampling with replacement without replacement does not matter here, technically it does not matter. So, obviously, if somebody saying that what if you are trying to pick up from the from a sample, will add sample not a population, and what if you continue picking a one observation at a time without replacement; then what will happen to the observation the long run. So, in that case my answer would be you pick up, observation note it down and again replace into the box, so that it is simple random sampling with replacement such that the overall set of observation which is there in the box, if you keep you can observations wonder a time, is actually infinite in the long run.

So, considering the observations you pick up, you note it down so it is basically a small z 1 1, small z 2 1, small z 3 1 till small z n 1 which is the first set of observations being picked up from box 1 to n, note them down and you basically square them up, square those values; so that is basically small z 1 1 whole square then small z 2 1 whole square so and so forth.

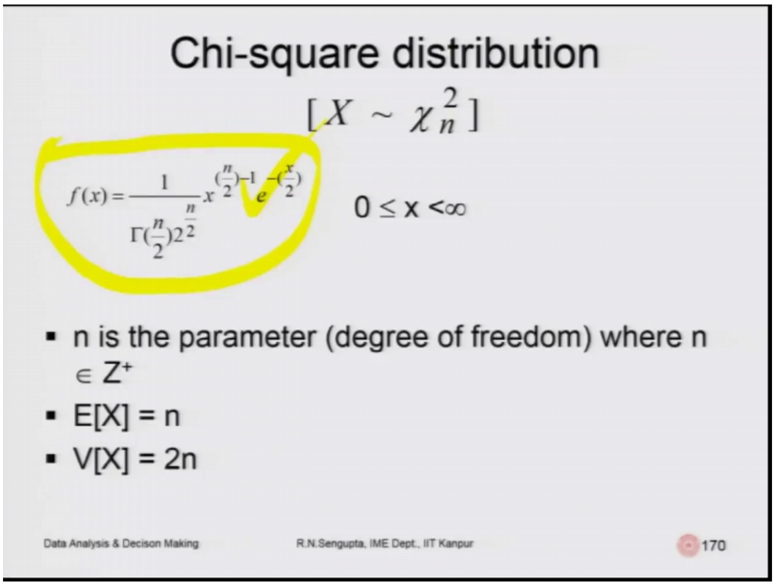
Then you pick up the second observation set of observation. So, they are denoted by small z 1 comma 2, small z 2 comma 2, small z 3 comma 2 till the last one. Then you pick up the third one, it will be small z 1 comma 3, small z 2 comma 3, small z 3 comma 3 and so on and so forth. So, for each observation you pick up you square them, and you basically add them up. So, when you are adding up; it basically if you have noted down, let me so if you a[re] found out the first observation, so you square them up; the first observation from the second box, the first observation from the third box, you add them up, you note that down value.

You continue for say for example, the second picking up, so sorry you note that down then the last value, and you continue doing it. So, you add them up, add them up. So, the values which you get on the right hand side; I will just mark you to with that different color. So, this would be. So, they are what, they would technically be or realize value from an unknown distribution which your picking up inform infinite number of times.

So, if I basically plot these values, all these which you have then that becomes actually a Chi-square distribution with n degrees of freedom. So, technically the first value would be the Chi-square 1 comma 1 or 1 1 which is the Chi-square distribution, let me denote not 1 comma 1, it will be n would be best. So, Chi-square with n degrees of freedom and the second and the second number which is 1, is basically the observations we are going to get. So, obviously these are the realize value.

So, if you basically go to the second one. So, it will be Chi-square n and 2, where the number 2 denotes the second observations. If we continue taking this set of observations as denoted as I noted that down, they would basically with a Chi-square with n degrees of freedom. And I will come to that discussion later on.

(Refer Slide Time: 18:09)



Chi-square distribution

$[X \sim \chi^2_n]$

$$f(x) = \frac{1}{\Gamma(\frac{n}{2}) 2^{\frac{n}{2}}} x^{(\frac{n}{2})-1} e^{-(\frac{x}{2})}$$

$0 \leq x < \infty$

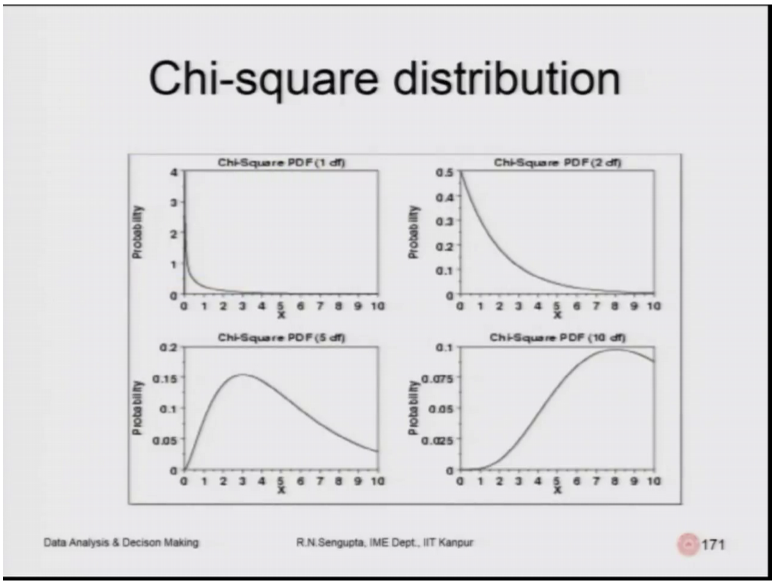
- n is the parameter (degree of freedom) where n $\in Z^+$
- $E[X] = n$
- $V[X] = 2n$

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 170

So, this is a Chi-square distribution as you can see. So, Chi-square distribution random variable X and the degrees of freedom is n. Now, the f of x which is the PDF of Chi-square is denoted by this function, so let me highlight it. So, this is the Chi-square distribution which I have, corresponding to the fact that we have the basically the Chi-square distribution. And the x values which is the random variable is basically from 0 to infinity; obviously, because if z, z squares are there so obviously, negative z's are not possible, negative values of Chi-square or not possible. So, they would be from 0 to infinity.

n is the parameter or the degree of freedom, where n is obviously, basically any integers from 1 to infinity. And the expected value of the Chi-square distribution is n, and the variance of the Chi-square distribution is 2; and it can be proved, but I am not going to go into the detail proves as required.

(Refer Slide Time: 19:09)

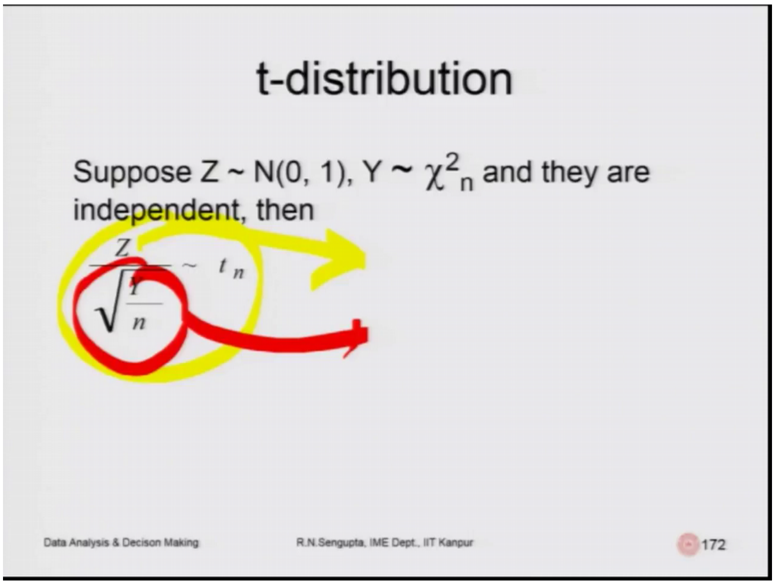


Chi-square distribution

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 171

So, these are the Chi-square distributions being plotted for the degrees of freedom being 1. So, this is the where I am when I am pointing out my finger. So, this is the Chi-square distribution of PDF, and along the X axis you have the X values; along the Y axis you have the PDF. So, this is for degrees freedom of 1; the second one is basically for degrees freedom of 2; the third one is for basically for the degrees of freedom of 5; the fourth one mean that degrees of freedom for 10 and so on and so forth, they can be plotted.

(Refer Slide Time: 19:43)



t-distribution

Suppose Z ~ N(0, 1), Y ~ χ^2_n and they are independent, then

$$\frac{Z}{\sqrt{\frac{Y}{n}}} \sim t_n$$

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 172

So, these I will repeat time and again. So, first let me basically give you a feel about this distribution. The second distribution which we can talk about is the t-distribution, which is now also known as a student t-distribution or the Gaussian distribution. Suppose you have, now consider the picture like this, before I use the word suppose, consider the picture like this. So, in one room we have those n number of boxes which where each box are standard normal deviate with mean value of 0 standard division 1 and each box is a population.

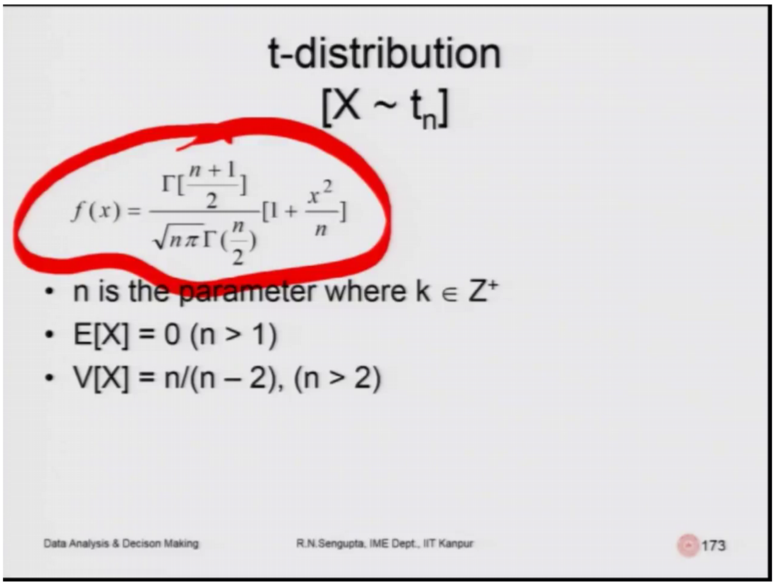
In a second room there is only one box, again it is standard normal distribution. So, what you do in the first room you basically generate the Chi-square distribution, and in the first room basically you take the z value or the z distribution as it is. Now, what you do is that as you find out all the realized value. So, for the first round of the experiment which you do, you find out the first set of observation in room one; note it down that is the Chi-square, n degrees of freedom and that is a first reading.

As you complete that you also take the z value from the second room, which is the normal standard normal deviate and this is the first observation. What do you do is that, you divide the standard normal deviate then value, divided by the Chi-square, divided by degrees of freedom, so that denominator is square root and note down that value and keep it.

Then you do the second time, pickup n number of z values from the first room, square them up, add them up that is the Chi-square the second reading as you complete. So, you take also second reading from the second room which is the z distribution. You find out the ratio that is the ratio being in the numerator being the realized value from the second rooms.

So, the second time divided by the realized value from the first room Chi-square distribution with n degrees of freedom divided by the degrees of freedom that square root of that in the denominator, note it down. Continue doing it, as you continue doing it, you will have different the ratio would give you different values. If you plot them, that distribution is actually a t-distribution with the degrees of freedom of n, as denoted now denoted here. So, you take in the numerator the Z values and in the denominator you take the ratio which is basically the Chi-square divided by the degrees of freedom, so that the ratio would basically be given by the t-distribution.

(Refer Slide Time: 22:16)



t-distribution

$[X \sim t_n]$

$$f(x) = \frac{\Gamma[\frac{n+1}{2}]}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}[1 + \frac{x^2}{n}]$$

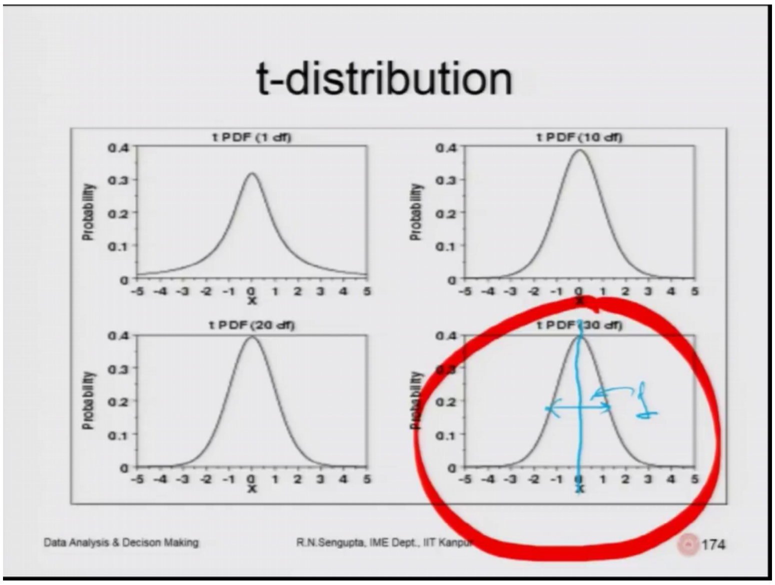
- n is the parameter where k $\in Z^+$
- E[X] = 0 (n > 1)
- V[X] = n/(n – 2), (n > 2)

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 173

So, the t-distribution x as a random distribution PDF as t-distribution with n degrees of freedom, and if you find out, so obviously the f of x would basically be this value PDF which is given, and obviously you can find out if it is a ratio of z divided by square root of Chi-square, so obviously the values would be from minus infinity to plus infinity.

So, n is a parameter where the k the value of n, would basically be all the standard normal the integer values, expected value is 0 when is greater than 1 and the variance is given by n by n minus 2. Now, you noticed very interesting one thing the accepted value is 0 and the variance as n tends to infinity; obviously, that ratio would slowly tend to 1, which means that the t-distribution in the long run will have a expected when a variance which is exactly equal to the standard normal deviate; that means, the standard normal deviate will mimic the sorry; the t-distribution will mimic the standard normal deviate as the sample size increase. We will come to that and show that later also.

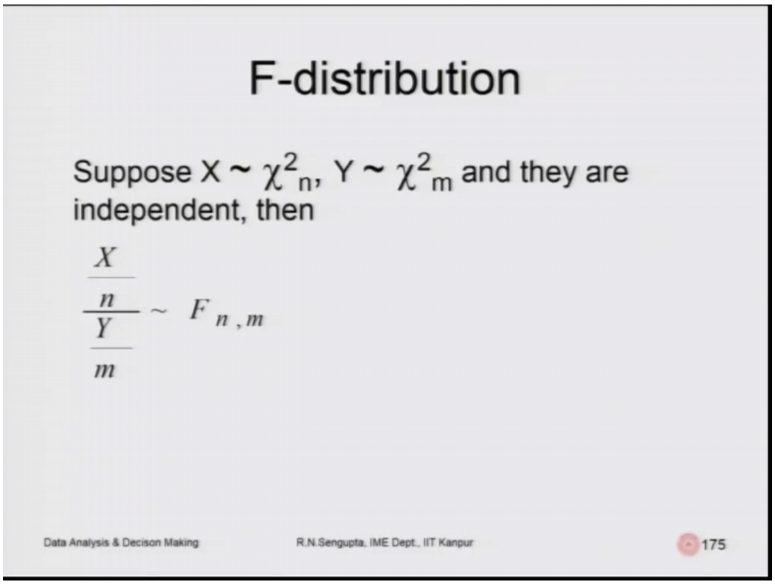
(Refer Slide Time: 23:41)



So, this is the t-distribution for on the first value where I am hovering my finger, this is the PDF, PDF for the t-distribution for degrees period of 1; this is the PDF for the t-distribution for degrees freedom of 10; the third one is for degrees of freedom of 20; the fourth one is degree of freedom 30. If you look here, so you what you will can find out is what I just said few minutes back. So, the mean values in almost exactly this 0 and variance in the long run would be equal to 1. So, this value the variance would be the square of the of the standard deviations of plus minus (Refer Time: 24:20) will be the standard division should which is exactly equal to the standard normal deviate, and the distribution which you see which, I have marked in red circle would almost mimic the standard normal distribution in the long run.

So, if you are trying to use some tables, why we are going to use the tables; I am going to come to that later. If you are trying to utilize the tables for the t-distribution, you will able to find out that that distribution almost mimics the standard normal deviate.

(Refer Slide Time: 24:51)



F-distribution

Suppose X ~ χ^2_n, Y ~ χ^2_m and they are independent, then

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

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 175

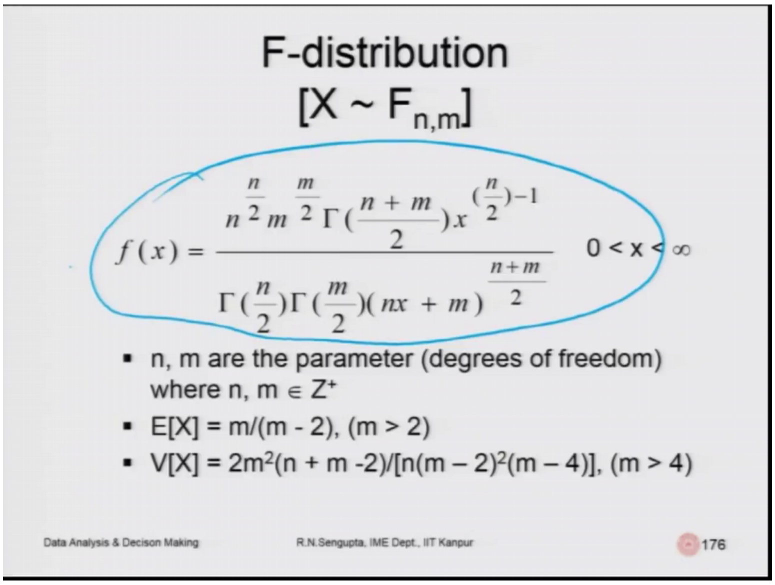
The third distribution which we will continue is basically the F-distribution. Now, the in F-distribution the scenario, I will basically give the background of the scenario. So, this one room which has basically a n number of standard normal deviate; that means, there are a boxes, each being box a standard normal population. In the second room, you will basically have m number of observation, which means there are m number of boxes from each box again you can pick up the standard normal deviate.

Now, you pick keep picking the observations and the first round, second round, third round continue, as you are doing it from the first room; that means, if you pick up the observation from the standard normal deviate, as you will square up each values add them up. So, that would be Chi-square. Similarly, as you continue you will pick up the first set of observations from the second room from each box square them, add them up that is the Chi-square of m degrees of freedom; in the first room it is n degree of freedom.

Continuing the second time, you will have the second realize value from the first room; similarly second realize value from the second room; then the third realize value from first and second; fourth realize value from first and second fifth sixth so and so forth. So, all of all the distribute values which you are taking from room number 1 is Chi-square with n, and the second room would be Chi-square with m.

Now, if you find out the ratio of the Chi-square from the first room divides by its degrees of freedom, divided by the Chi-square of second room divided by degrees of freedom, then you will get an F-distribution which has a degrees of freedom as n comma m. So, if you reverse the ratio; that means, you take the ratio for the second room to the first room; obviously, it would be a F-distribution of m comma n that is Mango comma Nagpur. So, I will come to those concept later on in more details.

(Refer Slide Time: 26:42)



F-distribution

$[X \sim F_{n,m}]$

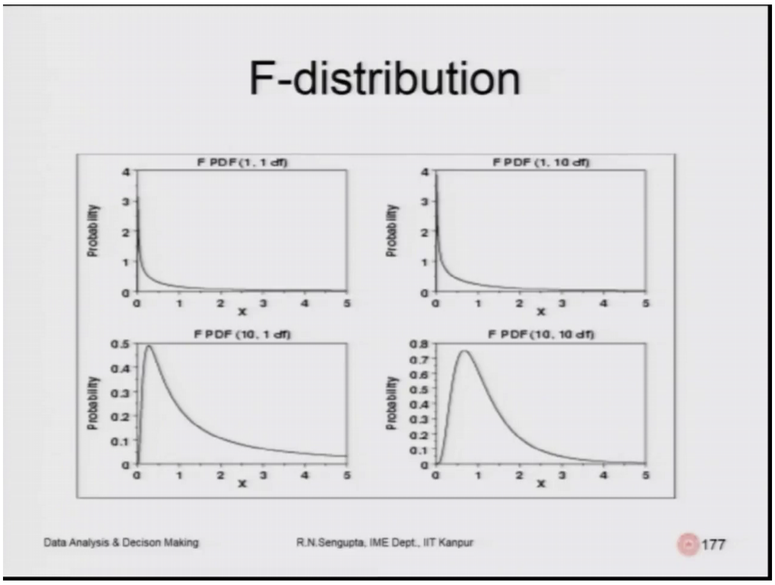
$$f(x) = \frac{n^{\frac{n}{2}} m^{\frac{m}{2}} \Gamma(\frac{n+m}{2}) x^{(\frac{n}{2})-1}}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})(nx+m)^{\frac{n+m}{2}}} \qquad 0 < x < \infty$$

- n, m are the parameter (degrees of freedom) where n, m $\in Z^+$
- $E[X] = m/(m - 2)$, $(m > 2)$
- $V[X] = 2m^2(n + m - 2)/[n(m - 2)^2(m - 4)]$, $(m > 4)$

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 176

So, F-distribution is given by F n comma m, so that would basically n and m being the degrees of freedom the f of x which is the PDF is given by this, and obviously, as Chi-square is a is square; that means, square of the zs; obviously, the ratio would also be a positive value. So, x value for the Chi-square would be from 0 to infinity; n m n n m m are the parameters, n and m are obviously, integer values. The expected values equal to m by n minus 2. So, depending on which ratio you are taking and the variance would be given accordingly.

(Refer Slide Time: 27:14)



F-distribution

Data Analysis & Decison Making R.N.Sengupta, IME Dept., IIT Kanpur 177

So, if you look at the F-distribution. So, the PDF of the F-distribution for 1 comma 1, the because there are two degrees of freedom 1 comma 1; that means, n or m or n or m whichever you look like I will come to that later on, how you look at the tables. In the second cases, the PDF for 1 comma 10; in the third case we have the PDF for the case for 10 comma 1.

So, you are reversing the degrees of freedom, and the forth value we have basically the PDF for the 10 comma 10. So, with this I have discussed very briefly I am telling you before end, that three distribution Chi-square, t and F will be coming back to the time and again has we utilized for the hypothesis testing, the interval testing and for the multivariate distribution also. So, with this I will end the 13th lecture.

And thank you very much for your attention. Have a nice day. Bye.