Data Analysis and Decision Making – I
Prof. Raghu Nandan Sengupta
Department of Industrial & Management Engineering
Indian Institute of Technology, Kanpur

Lecture – 14

Welcome back my dear friends, very good morning, good afternoon, good evening to all of you. I am Raghu Nandan Sengupta from the IME department IIT Kanpur. And this is the DADM-1 course which is the Data Analysis and Decision Making 1 course under the NPTEL, MOOCs series. And we are going to start the 14 lecture which means we are in the second last class for the third week, because each week, if you remember there are 5 classes, each class being for 30 minutes, total durations for the program is 12 weeks which is 30 hours.

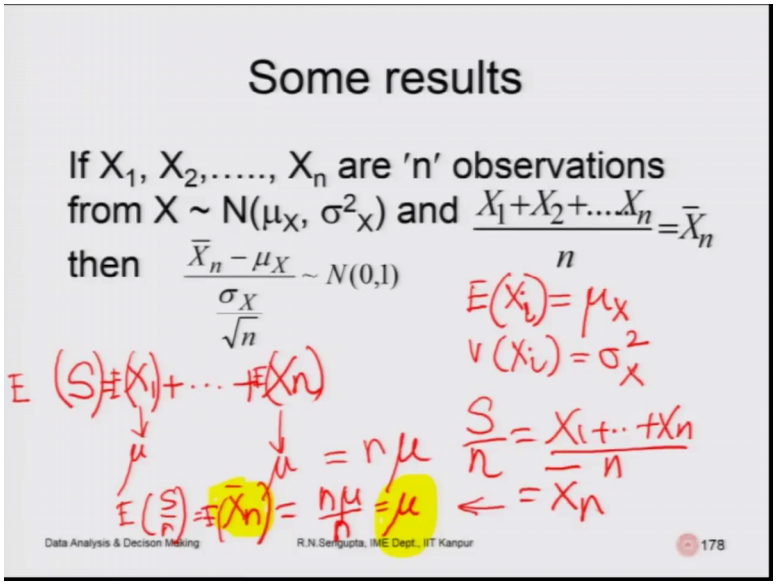
So, if you remember that in during the last 2 lectures, we were discussing about different type of continuous distribution. And based on that I told you that there were three important distributions which was the F distribution, chi square distribution and t distribution or in this order I should say chi square, t and F. And we basically showed that what was the PDF of this distribution, how they are formed. In the chi square they are basically formed by the summation of the squares of n number of standard normal. And in the t distribution, it was the ratio of a z distribution divided by the square root of a chi square with divided with the n degrees of freedom.

And in the case of the F distribution, you had two different like two different sets of observations one with m degrees of freedom, another with n degrees of freedom, chi square both them of them. And when chi square divided by the degrees of freedom in the numerator the chi square degrees of divided by the degrees of freedom in the denominator, you had the F distribution with the corresponding degrees of freedom where m and n, where m would be in the numerator, n would be in the denominator. And obviously, you can reverse it and you can get the corresponding F distribution with n and m, where m is say for example, I am trying to denote by mangoes and n as in Nagpur.

Now, we will give you some important results. The results will be helpful in trying to utilize the tables for the z distribution, if you have seen earlier then the tables for the chi square table for the t distribution, table for the F distribution. And I also mentioned that

time and again, when we are doing the t distribution that the t distribution as the sample size increases, it becomes exactly equal to the standard normal. So, it basically mimics the standard normal. And in place of t distribution table, you can use the chi square for the standard normal deviate table also with they would be error, but obviously, that can be minimal as the sample size increases. So, with this I will start the few results discussion.

(Refer Slide Time: 03:04)



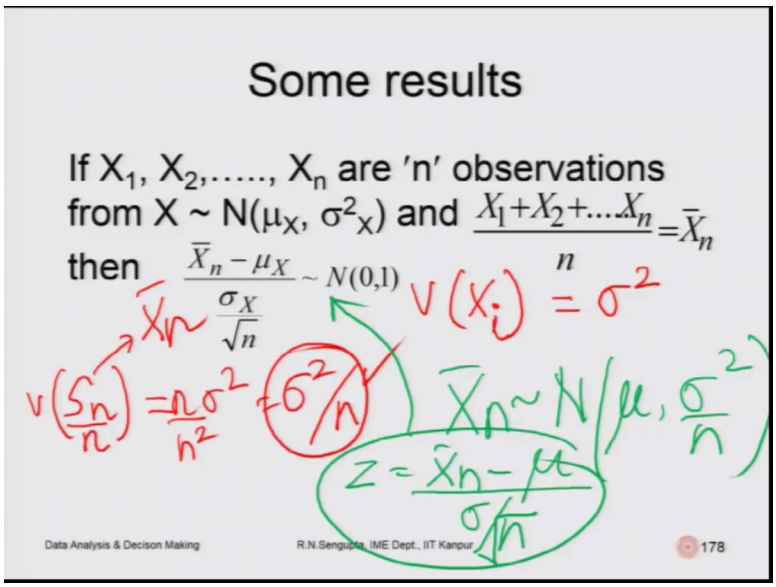
The first result, considered that you have n number of observations which are coming from must normal distribution with mean as mu and variances sigma square. So, you pick up observation the first one, second one, third one and there are infinite set of observations. So, it would be basically similar to the fact that you do it simple random sampling which with the placement order you do simple remnant sampling without replacement that would not matter.

So, you pick up X 1, X 2, X 3, X 4 till X n. And when you pick up these observations what you actually have is this is true. So, let me again reiterate because these we have done, but still I would like to make it very clear. So, let me use the red color. If you pick up the first observation, the expected value of the first observation would always be, because whatever the first observation you pick up that in the long run can be any of the observation from the normal distribution.

So, the mean value would be this with a suffix X being X being the random variable. So, if I put out i where i means any one of the n number of observation. So, this is true. Similarly, variance this is true with X is the random variable denoted. Now, what we actually need is initially in the sum. So, sum is basically X 1. Now, if I want to find out the expected value of the sum, so what I need to do I need to find out the expected value of X 1 till dot dot till the expected value X n. So, each of them is mu, I am not writing the suffix now, this is also mu. So, it will be some of them would be n times mu, so it will be n mu.

Now, sum divided by n would basically be the sample mean. So, if it is sample mean I want to find out I am again coming here. So, I want to find out the expected value that comes out to be n mu by n which is mu. So, this is what we have been highlighting. So, if I consider this one, this X n bar as a random variable, so it will be normally distributed with what would be the mean, the mean would be this. So, the first part which is related to finding out, this is just a repetition, please bear with me. So, the mean value would be mu only ok.

(Refer Slide Time: 06:09)



Now, let me come to the variance. Now, when I am trying to find out the variances, so let me consider the first one and then generalize it for the ith one. So, first let me write down one. So, the variance of this, because I can you pick up any observations. So, in then actually it is equal to this I am not writing the suffix. So, if I take the ith one, this is the

variance. Now, I had told you these, these absorptions which are picking up are all are technically independent of each other they are known as IID independent and identically distributed.

So, if they are independent, and if you remember we had discussed the concept of independence then the corresponding covariances between the ith and the jth, i can be 1 to n, j can be 1 to n provided i is not equal to j. So, all the covariances would be 0, which means, if you remember I had drawn that n by n matrix or m by n matrix, whatever it is which denotes the number of variables, in that case of the diagonal element are all 0, the principal diagonal would basically be the covalent the first 1, 1 comma one element would be the covariance of the first with itself, so 2 comma 2 would be the covariance of the second with itself and so on and so forth, so they would always be the variances.

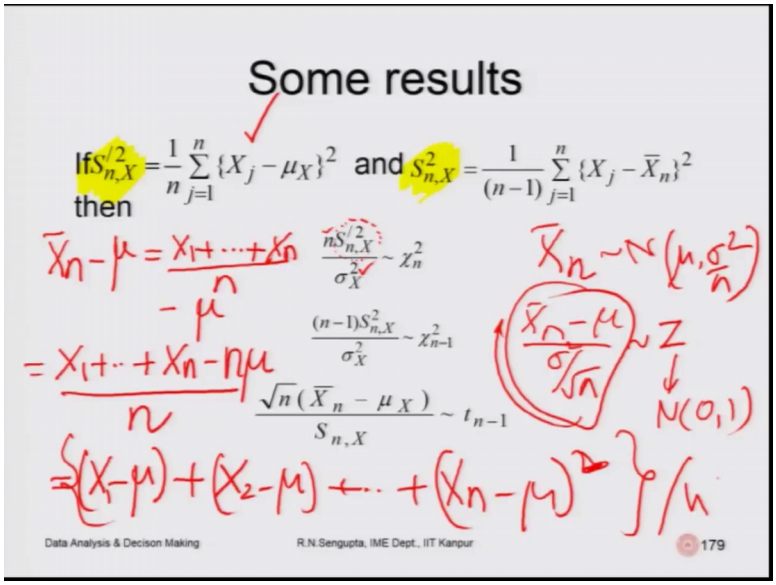
So, when I have the covariances or I want to find out corresponding covariances for this m n number of observations, I can do it accordingly. So, let me come back. So, I want to find out the variances of this, so this will be sigma square. So, each variance would becoming, the principal diagonal part. So, how many such values will be there it will be n, because the (Refer Time: 07:53) of the diagonals are covariances are 0. So, you have basically sigma square being added up n number of times.

Now, I want to find out the variance of S n by n which technically means I want to find out the variance of the sample mean. Now, this n which is in the denominator when it goes outside it, obviously becomes the square. So, this becomes sigma square by n as we have already mentioned. Now, technically means, so this one and the initial set which we said then the mean value of X n bar is mu actually proves that the distribution, which we have. And, we use another color for the time being.

So, distribution of X bar n is normal with mean mu, so mu without the suffix I am utilizing and the variances are sigma square by. So, if it is normal I can convert it into a standard normal. So, again I use the concept, this mean value of X n bar which is this and the standard deviation that, that was what the formula was the random variable minus the mean value divided by the standard deviation. So, standard deviation would be now sigma square root basically so this becomes square root of n. So, if I go take this. So, actually this is standard normal with zero mean and one standard deviation are one

variance as it is a standard normal deviate. So, remember this, this would be utilized time and again.

(Refer Slide Time: 09:31)



Now, let us come back to the actual results slowly. So, I will spend some time in this slide. So, please bear with me I will be going not slowly, but I will be explaining in a little bit detail. Now, consider there are initially they consider and let me build up the story. Consider there are two different samples, consider there are taken for two different factories or two different sets of people or two different sets of students or two different sets of objects, whatever it is. Consider there in sample one from one population, which is a normal distribution with certain mean, certain variance I am considering the normal distribution.

And in the other case, consider sample two with again a normal distribution with a certain mean, certain standard deviation. And remember the mean of both of them are different, standard deviations of the variance of both the normal distributions are different; and the sample size in both the cases from the population you are taking one or two are different. So, let us consider the sample size, which you are taking for population 1 from population 1 is m, m as in Mangalore or Mangoes. And the sample size of the set of observations you are taking from, from the population 2 is basically n or n as in nose as an in n as in Nagpur. Now, I will be using the m and n interchangeably, but when it is

very specific to the problem specifically for the F distribution and make it very clear, where we are taking m, where you are taking n.

Now, consider for the time being, we are taking n n as in Nagpur. Now, we define two standard error square or the variance of the sample one we will denote by. Let me use the highlighter now one would be S dash square and one be the S square without a dash. The suffix basically denotes, that you are taking the sample size of n. If it is m it will be replaced by m. And the next symbol which is x basically denotes what is the random variable how we are denoting, it can be x, x 1 or it can be y, y 1, y 2, y 3 whatever it is.

Now, S dash square with the suffix dash being there, we denote it by the, the, so called standard error square where we divide by the sample size n only. So, it is X bar minus mu whole square divided by n. And, if I denote S without the dash, it is basically in this bracket we sum it up, obviously, the squares we sum it up in both the cases we sum up the squares and the of the differences. And what is the difference it is X i's each observation in the first case also with the dash, we said that without the dash we are saying it is X i's minus now it is not the mean. The mean has been replaced by the sample mean which is X bar n.

Now, the moment you replace the sample mean, what happens is that rather than dividing by n you divide by n minus 1, because you lose one degree of freedom; that means, your mobility in a way reduces. So, let me explain what we mean by losing one degree of freedom. Now, what you do is that you pick up the observations 1, 2, 3, 4 till n. So, the first time when you pick up, you check whether the first moment on the mean is known, if the answer is yes, then obviously, you will use the mean value mu and continue with the you are your studies. But, if the answer is no, that means, the mean value of the population is not known; obviously, you has to be replaceable. So, mu is being unknown, so you have to basically replace that with the corresponding characteristics from the sample.

So, what is that is basically the mean value which is X bar n. So, the moment you use the set of observation the one time you are losing some degrees of freedom or some set of information is being lost from the set of observation, hence it is divided by not by n by n minus 1. So, this is the simple way how I can explain. So, considering the standard error

squares one with S dash another with S without the dash, you will basically have two basically simple chi square distribution.

Now, step back. So, what we said is that chi square distributions are basically the sum of the squares of n number of standard normal deviate. So, if you consider the standard normal deviate, in the case when you have S or S dash, the corresponding standard normal deviate and dividing by the corresponding values of the standard deviation of the population would basically be like this.

So, I will basically only consider the first case. So, the first case is and that would make things clear to you. So, when you are considering this, so actually in that case X n is normally distributed with mu sigma square by n, you convert it into the standard normal. If you convert it to the standard normal, what you have is this X n bar minus mu by sigma by square root of n is equal to Z; so, technically Z N 0 1. So, you agree with this?

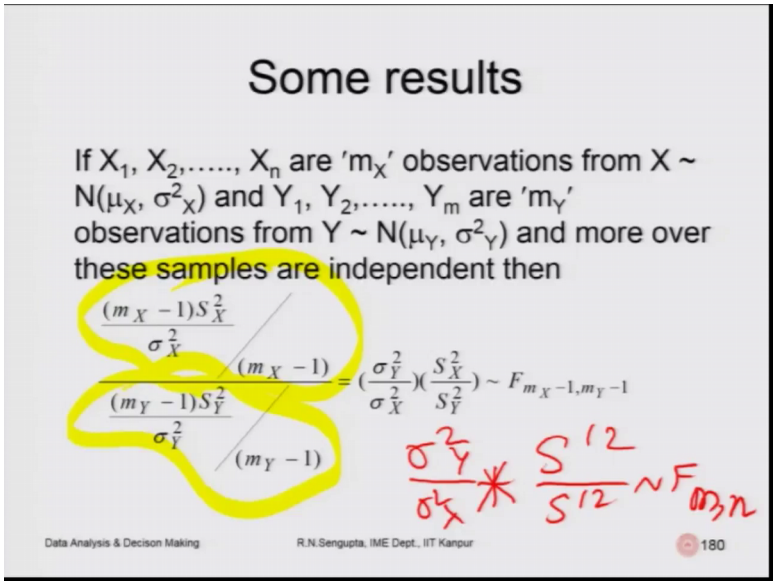
Now, what do you do for the chi square, you pick up Z, square them up add them up. So, let us do that. So, I am picking up observations I am squaring them up. So, technically when I use this squaring would basically make this square and this will go in the numerator. So, this part and this part are taking care, what about this, I am just going to dot it. Now, X n are what X 1 so minus mu. So, I have X n minus plus sorry and n into mu, sorry I did not write it properly my apologies let me, let me write it properly, so you will understand. I will first write it and explain it to you. This is X bar n minus mu this whole thing divided by n no this is not square.

Now, look here. So, these are what each observation minus is mean value. So, if you go back mean way, and if you go back to the actual case you have squared them. So, they are squared. If you go back to S dash is exactly the same formula which we are trying to imply. So, when you basically have this part divided by its corresponding square of the standard deviation square not the square root square of the standard deviation, it actually becomes the sum of Z 1 square plus Z 2 squared plus Z 3 square plus dot, dot till Z square suffix n which is actually a chi square with n degrees of freedom replace that S dash with S.

You have the corresponding chi square distribution, but you are, now losing one degrees of freedom that is why it is given a chi square n minus 1. And n minus 1 in that case you

write n minus 1 S without the dash square divided by sigma square gives the chi square with n minus 1 degrees of freedom. Now, when I come to the case of trying to find out the t distribution, similarly I am not going to go to the proof, you can prove it that X bar n minus mu here divided by the standard deviation of the standard error, which is S S n without the dash divided by square root of n is basically a t distribution with n n minus degrees of freedom.

(Refer Slide Time: 18:57)



Now, if you remember I did mention that we are going to take m and n number of observations from two sources and consider they are given. So, you have X 1 to X n and this should be sorry they should be n. So, they are basically m suffix x observations from normal with mu suffix X and sigma square suffix X number of this distribution and their mean value, and standard deviation square of the variance is given. And Y 1 to Y m, so that means, consider there are two rooms you take which are actually the two populations, you are picking up observations.

Y 1 to Y m, so they are m suffix y number of observations. So, there are normal distribute again Y with mean mu and the suffix I would ignore, but I will still say that for the first time, it is nu suffix y and sigma square suffix y being the mean and the variances. And moreover consider these are IIDs and they are independent. So, obviously, if you have chi square divided by its corresponding degrees of freedom in the numerator, and again as chi square divided by the degrees of freedom in the denominator,

you have Fdistribution with the corresponding degrees, where the first value is the degrees of freedom for all sample size in the numerator. And the second degree of the F distribution being the sample size or the degrees of freedom for in the denominator; so, this is what is writing I am writing.

So, I know and I am basically going to explain that with the equation which is there and give you all the information. Now, consider what is not written here, I will mention that. So, it will become clear to you. Consider that, you have the mean values for both the population of 1 and population 2 which is in two different rooms which is basically the distribution of X and distribution of Y both being normal, the mean value is known. So, in that case, you will have m minus not the one, because minus 1 would not come. So, m into s dash remember dash means the population mean is known m into S dash square divided by sigma square for the first one divided by m, because now m is the corresponding sample size of the set of observation for the first room, so that will come in, in the denominator.

So, if it is the denominator has m where I am hovering my pen and in the numerator I am talking only, the upper part upper part means this only. And in the numerator of that upper part, you have a chi square distribution. So, the chi square distribution by its degrees of freedom would give you the case that we are slowly trying to give some information of the F distribution.

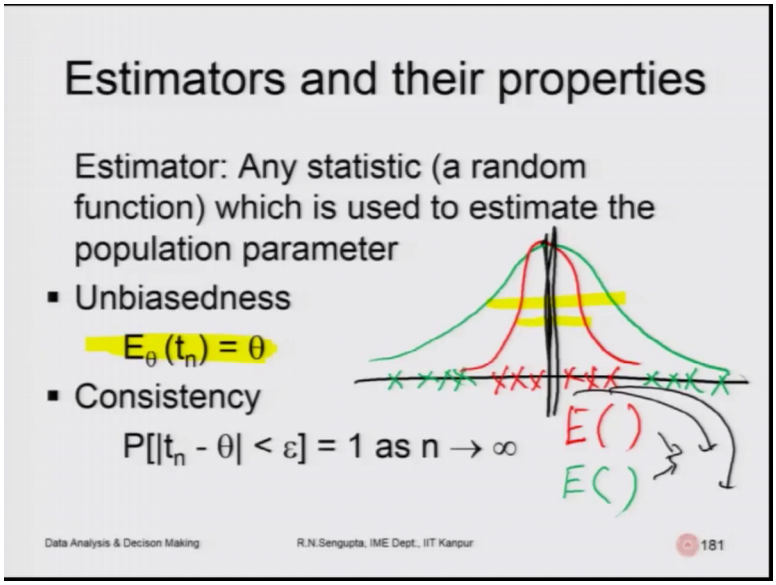
Similarly, if you go into the lower part which is the denominator here, there you have n into S dash, because here even though it is not written dash I am considering dash means the population mean for the second a set of observations or second population is also known. So, again, if you see it is m into s dash squared divided by sigma square for that second population divided by the total thing divided by the degrees of freedom for the second sample would give you the actual information saying that I am trying to basically formulated an f distribution.

So, the actual formula which we have is basically for F distribution with degrees of freedom as m comma n would be given by m s square divided by sigma square divided by m. And that whole thing in the numerator in the denominator it would be n s square divided by sigma square divided by n. So, that would actually, become in the case, when the degrees of freedom are not being lost. So, this is what I am this is a multiplication

this will become S square this is also become a square you will basically have m m comma n.

The moment you replace S dash with a corresponding S dash with S without the you have basically the corresponding ratios being for the chi square which in both the cases are losing one degrees of freedom in, in, in population 1, 1 degrees of freedom in population 2. Hence, it becomes an F distribution with m minus 1 and n minus 1 degrees of freedom. So, we will be using this quite often later on.

(Refer Slide Time: 23:44)



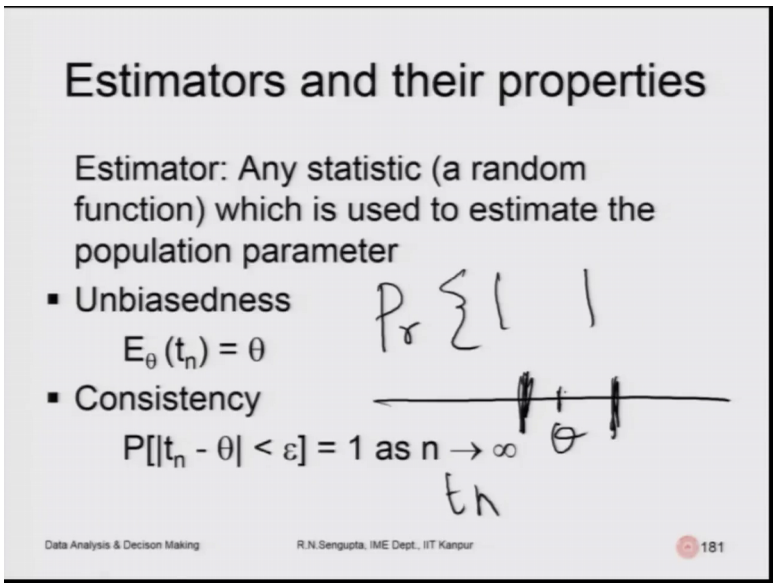
Say estimators and their properties. So, any statistic a random function which is used to estimate the population parameter would basically have actually will try to see that the two properties meet one is the unbiasedness and one is the consistency. Unbiasedness basically means as I keep picking up observations in the long run the expected value of the estimate from the sample would basically be exactly equal to the population parameter which you are trying to find out. So, it means that expected value match. And consistency means that as the sample size increases in the long run the, so called variance should basically decrease. I will give this give you not an example I try to draw two in order to make things easier for us.

So, I will draw three or four different type of diagrams in this slide and then come to them. So, and explain this consistency and, and, and an unbiasedness corresponding. So,

let me draw. So, consider that I have the graph paper y-axis, x-axis I am plotting two different distributions and they would be denoted and taking some observations. And I am taking this set of observations, and marking them with two different colors. So, let me take the first one with green. So, the observations, if I plot along the x-axis, they are like this. So, if I plot them consider the dispersion is normal for the time being, it basically it is like this. Now, I take another distribution again normal I use different colors and the observations are like this.

Now, what do they give, in both the cases expected value of the red and the expected value of the green both are of sorry, both are equal to the line which we have this one. So, these are the case for both of them, but check what is the variance for the green one I am going to use the highlight. And, now let me use the yellow color for the green one the variance is, so large for the red one the variance is low. So, the first property of unbiasedness is being met by both of them, consistency would be that the red one would be more consistent as the sample size increases with respect to the green one. Now, the corresponding consistency property, I will try to again use a simple diagram.

(Refer Slide Time: 26:54)



So, what I am meaning, I am use Pr means the probability, probability of the difference this is the mode which I am taking. So, consider the straight line this is the value of theta and these are the limits. So, what I am saying the t n value would lie between this limit in this range and the distance between theta and t n will start shrinking and basically go into

0 that means, the length actually decreases, and becomes 0, in the long run as the sample size increases which means the sample size increasing means, it is trying to basically mimic the population parameter as closely as possible.

So, with this I will end the 14th class and continue more discussion about consistency and unbiasness later on. And use these three distribution of chi f and t more and more in trying to basically come up with the properties in the hypothesis testing and the interval estimation case.

Have a nice day. And thank you very much.