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

Lecture – 12

Very warm welcome to all my dear students and friends, a very good morning, good afternoon, good evening. Welcome to this DADM-1 course under the NPTEL, MOOC series. And we are for this course, which deals only with statistics multivariate analysis, later on obviously these will come. And this is a 12 week course for 30 hours each week as you know we have you have five lectures, each for half an hour. And this is the 12th one that means, we are in the 3rd week. And I am Raghu Nandan Sengupta, IME Department, IIT, Kanpur.

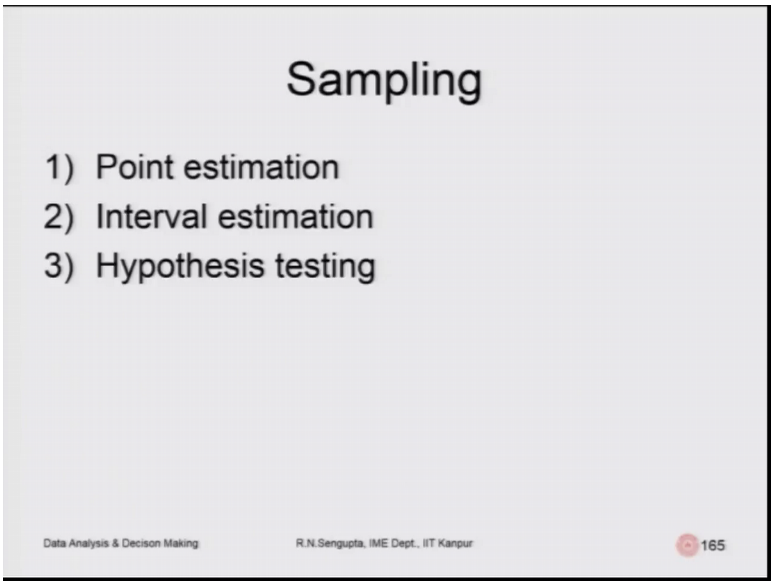
So, if you remember, the last slide or the last concept before we closed the 11th lecture was related to central limit theorem. So, the central limit theorem, basically says that considering any distribution with a certain mean, certain variance. So, we are only dealing with the first moment and the second moment, it can be done for other cases also. And we want to find out what is the sum, the expected value of the sum, and the average sum average of the sum, and in the long run, what is the expected value of that.

So, we found out that for the case if it is normal, then the mean value is mu for the average of the sample. Sample means those numbers of observations, which you are picking up from 1 to small n. And the variance for that sample is sigma square by n, where again small n is the sample size sigma square is basically the population variance.

Now, considering x bar n as a random variable, obviously it has to be because it has got a mean and a variance. Then its corresponding distribution as we told that if the population is normal, the sample distribution for the mean value is normal. Mean value, I am not talking about the distribution of the variance. So, the mean values of the sample distribution is normal. So, when you convert that to a standard normal deviate with the mean value of 0 and variance of 1. Then actually the conversion is like this.

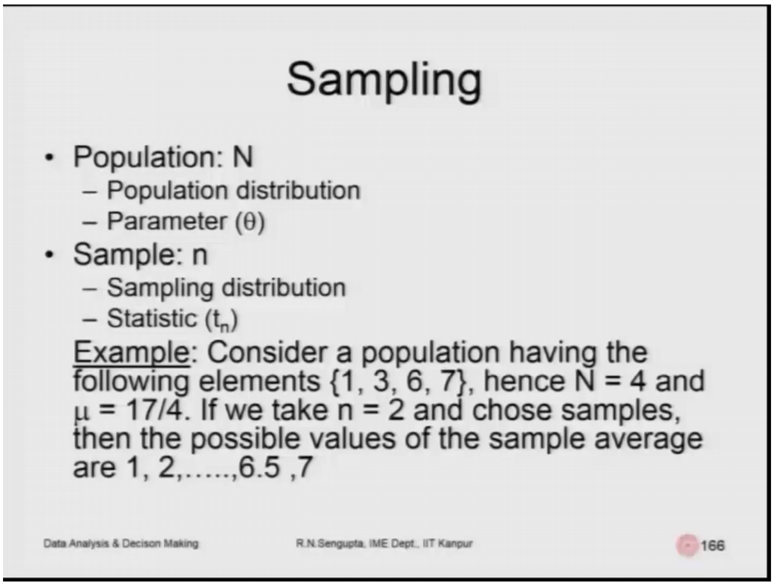
I am just saying that listen to it carefully in the numerator you have x bar n minus mu, because that mu is the expected value of x bar n divided by generally it is basically the standard deviation. So, what is the standard deviation in this case, the variance is sigma square by n. So, the standard deviation would be sigma by square root of n. And that would be normally distributed with 0 mean and 1 standard deviation.

(Refer Slide Time: 02:42)



Now, in the case of sampling; so we are going to the real of sampling. We will consider three different ideas. These are very broad ideas, but I will just stick to the bare minimum as you quiet. One is the point estimation, concept of point estimation: one in the concept of interval estimation. And one is the concept of hypothesis testing. So, what we mean by point estimation, interval estimation, and point estimation. I will go through that slowly.

(Refer Slide Time: 03:15)



Consider for the sampling, actually what why you do sampling is that you have a population of capital N or infinite. And what you have about the information on the population in the population has a distribution. Like it can be the geometric distribution, Poisson distribution, exponential distribution, (Refer Time: 03:34) geometric distribution, negative manual distribution, normal distribution, uniform discrete distribution, uniform general district uniform distribution, continuous one.

Now, you have the parameters. So, the parameters for the normal one more mu and sigma square, for the exponential one was a, or lambda or a or theta whatever you imply portion was lambda, for the binomial it was n and p, and all these parameters are there. So, we will denote the parameters in general by the theta. Theta can be there a vector or a scalar. So, if it is only one parameter, so consider like the Poisson distortion, theta is lambda. And if it is normal distribution, theta is a vector consisting of two elements mu and sigma square.

Now, if the parameters are known, we know the distribution in totality, and we can do our work. But, the question arrives that two thing, number one if we basically pick up a Poisson set of samples or observations of size small n, which means basically we are picking out observations x 1 with capital X 1 remember that those are random variables. So, hence as I mentioned time and again in like.

In the beginning, so they would be given by capital X or capital Y or capital Z. So, they would be random variables capital X 1 till capital X n. And we want to find out the distribution of the characteristics of the sample. So, it can be characteristics can be mean of the sample, characteristics can be mode of the sample, characters characteristics can be median of the sample, characteristics can be variance of the sample. And there are many characteristic, if you want to find out or some functional form also. So, given the parameters, we will try to basically do the study whether the characteristics sampling distribution is available to us point one.

Point number two is that given the parameters are unknown from the population. How we can estimate them, estimate means, fine (Refer Time: 05:43) them as close as possible using the sample sampling information. That means, the sampling information means from capital X 1 to capital X n, such that we are as close as possible to the parameter value provided we meet few characters criterias. So obviously, what are those criterias, I will come to that also. So, given a population of capital N and denoting size, population distribution would be given with the parameters known. If they are not known, you have to find out as I said in the in just few seconds back.

Now, the corresponding sample, which you pick up as I said capital X 1 to a capital X n is the sample of set of observation, until we have a sampling distributions. And the corresponding statistic, which you see here on the on the slide, which is small t suffix n, that is the characteristics corresponding to the population parameter, which you want to find out using the sample size that the sample statistics t n is as close as possible to the population parameter.

So, we want to find out that what is the characteristics t n from the sample size that is close as possible to the mean value of the distribution or else we want to find out that what is the sample characteristics from the, obviously from the sample, which is as close as possible to the median of the population or you want to find out that what is the sample characteristics t n. So, t n is obviously changing, which is as close as possible to the variance of the population.

So, we will try to study not in details just as relevant to this concept or DADM. So, consider an example, very simple example. There are lot of like rigidity in how you are trying to consider the example, but it will give you a feel that what we are actually looking at (Refer Time: 07:36) by the word rigidity means that we are not assuming all the things to be true, but in a very real (Refer Time: 07:46) in a very theoretical sense. This is the situation, and how we can attempt considering, what I discussed about the statistic the sampling population considering the parameters being known, not known whatever it is.

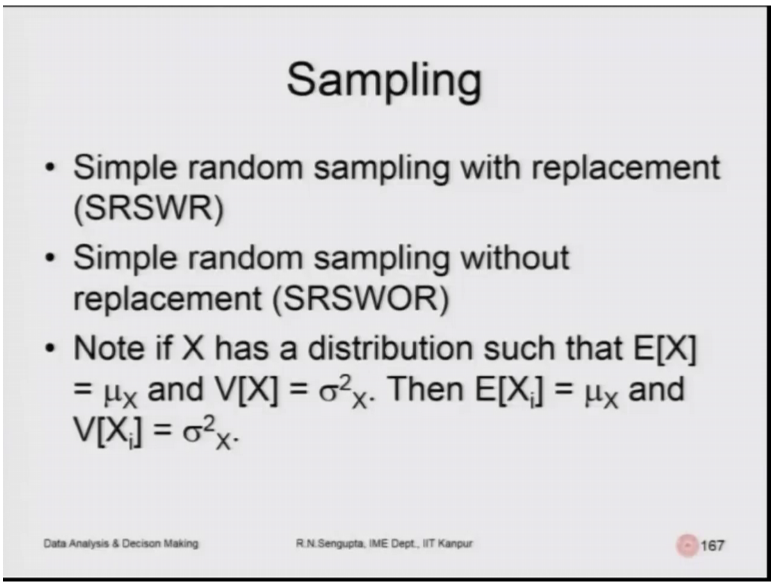
Consider the population having a following elements 1, 3, 6, 7. And so obviously rigidity being that capital N is 4, which is not possible, but we are considering that. So, hence capital N is 4. And the mu value would obviously be, because if they are 4, the corresponding probability will consider the corresponding probability to be uniform. So, the probabilities are 1 4th for 1 3, 1 4th for 3, 1 4th for 6, 1 4th for 7, hence the mean value is 17 by 4.

Now, what we do is that we take a sample size of n is equal to 2. So, n is equal to 2. So, we choose the samples in this way. So, we could have obviously we will consider that we are choosing in the theoretical sense the sampling are done in such a way that if you pick up number 1, you still have number one from the population to pick up in the second trial. Actually that is not true, because if we pick up 1, the chit marked 1 is not there.

So, we will consider this the sampling being done with replacements. So, the total number types of samples, we can have with 2 observation the sample can be in one extreme it is 1, 1 because you can pick up 1, and then again you pick up 1. And another extreme you have basically the sampling sample or the observations as 7, 7. So, so this is what it says, let me continue reading it.

And choose the samples, then hence the possible values of the samples averages, because if it is 1, 1, so it will mean 1 plus 1 divided by 2, which is 1. And if it is 7, 7, it will be 7 plus 7 divided by 2, which is 7; so whole set of average values for a sample of size 2 would be b starting from 2 to 7. Now, obviously they are the actual realize value for the sampling distributions average value, which you are talking about and if it is a distribution, you would like to find out what is the mean, what is the distribution parameter so on and so forth.

(Refer Slide Time: 10:11)



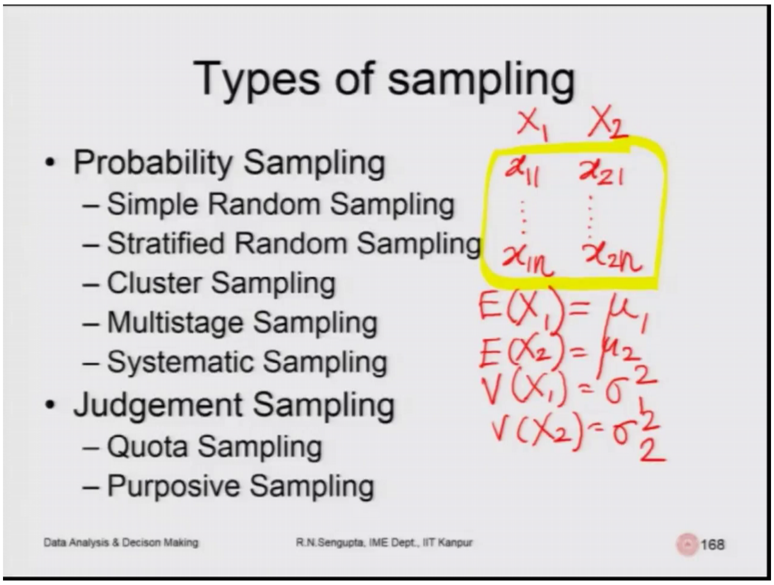
Now, simple and sampling can be of two types, as I said one is the sampling, sorry my apologies it is just went here and there the slides. So, sampling can be of two types. One is simple random sampling with replacement, which is SRSWR. And simple random sampling without replacement with these SRSWOR that means, we do with replacement, without replacement.

So, what I mentioned few minutes back, when you have that population of 4, if you picking up observations, and if you are doing with replacement. Hence the corresponding probability distributions or the probabilities for picking up any observations does not change. As for the case if you pick up 1, and if you consider that the number 1 is also present in the population the next time. And the next time and so on and so forth. Hence the probability always remains 1 4th. So, it is simple random sampling with replacement.

For without replacement, the probability of picking up 1 twice is 1 4th for the first time and the second time, it is 0 by 3, because there is no 1. And then total number of such observations the population has decreased by 1, hence is 3. Note that if x has a distribution says that the expected value of the distribution is mu suffix x, mu is again nothing to do with the normal case. Variance is sigma square suffix x, again nothing to do with the normal case.

Then as I said in the last class, which is the in lecture number 11. If you pick up any observations at random x i, x i is i-th place you are going to pick up. So, i-th place can be filled up by any one of the observations from the population corresponding the probability it has. So, if you want to find out the expected value of that i-th place, which is E X i would exactly it will be equal to mu. And the corresponding variance of x i would be sigma square suffix x. Now, remember one thing if you I want to find out something to do with the co variances, which I said so that would basically be taken into consideration, when you find on the covariances.

(Refer Slide Time: 12:05)

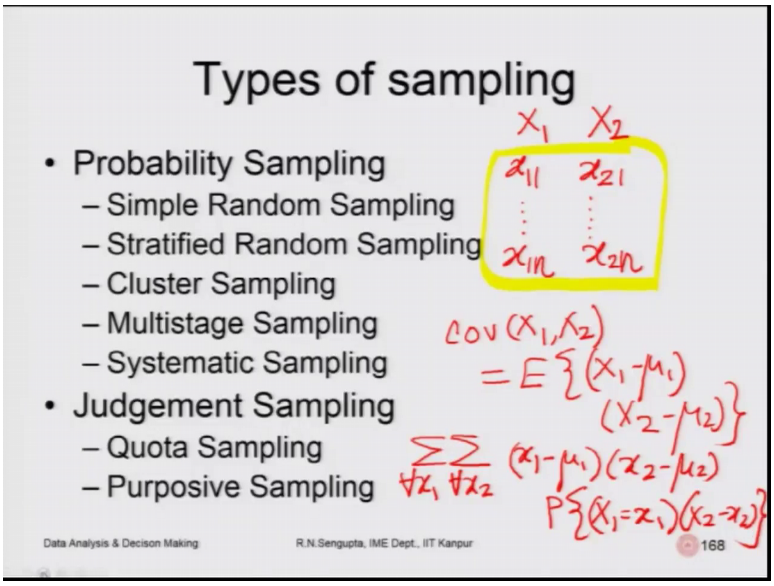


So, before I go into trying to basically describe very briefly, though the probability sampling the simple random sampling on all this things. So, let me come to the concept of the co variances, which I did not go into details, but I will just mention it here. So, consider let me use the red color yes. So, what you have is basically this X 1 and X 2 are the random variables. So, the random variables I will denote by small x 1. The first one is basically corresponding to the random variable number. The next one is the reading number. So, this will be 2, 1. And it consists quantities 1, n.

So, this is the number of observations, which you are picking up. This is the n is the number of observations you are picking up. Now if I want to find out and consider, this is a large set of values. So, if I want to find out the expected value, so the expected value of E X 1. We will consider it is given by. So, these are just the samples, whichever it picked up. So, let me highlight it with some other colors. So, these are the off set of samples, which I picked up.

So, let me go back to the discussion. So, this is given by 1 mu suffix 1. This is given by mu suffix 2. Variance of X 1 is given by sigma square 1. Variance of X 2 is given by sigma square 2. So, note this down. And then, I will proceed. So, this is the background. And I will erase it, but I will basically now come to the concept of co variances. And then again come back to whatever it is typed on the slide. So, I will erase it. So, this part which is highlighted can (Refer Time: 14:16) continue remaining, because whatever I have written about mu 1, mu 2, sigma 1, sigma square suffix 1, sigma squares of suffix 2 are for the population.

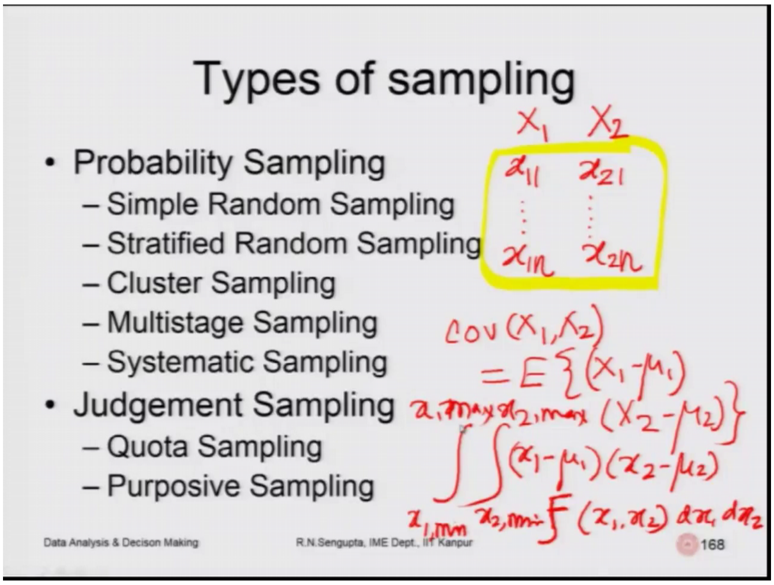
(Refer Slide Time: 14:26)



Now, co variance between X 1 and X 2 would be given the relationship, which is between X 1 and X 2. So, I want to find out the covariance, it will be X 1 minus mu 1. So, the E is the expected value. So, multiplied, so there is no space, I will continue in the next line X 2 minus mu 2. So, actually what I am doing, when I am trying to find out the corresponding co variances.

So, expected value would mean integration on this or the summation. So, if there are two variables, it will basically double summation. So, what I am doing is double summation for the discrete case, for all X i value X 1 values, for all X 2 values. For the case when I have X 1 minus mu 1 into X 2 minus mu 2 multiplied by the case, when I have the probabilities. So, these are the joint probabilities.

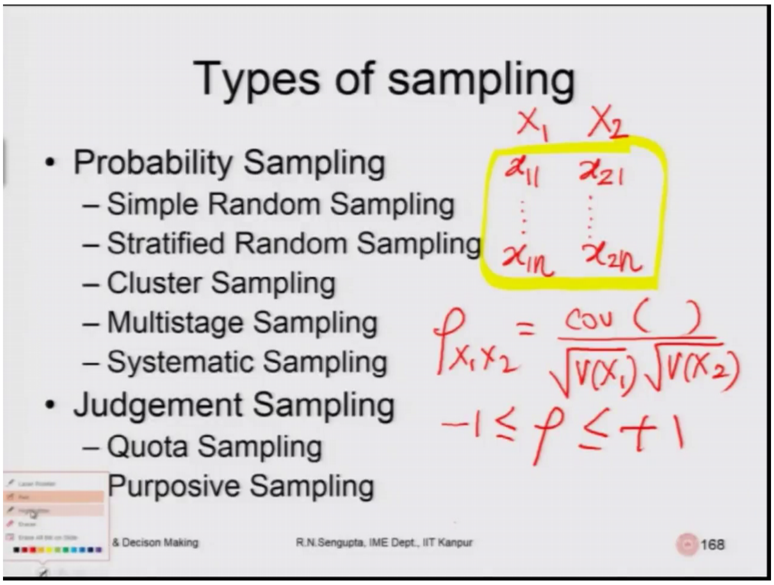
(Refer Slide Time: 16:02)



I am going to come to that details later on do not be too much bother about it. And the case when you have to the continuous case, it is double integration. And in this case probability gets replaced. So, this will become F this capital f or let me put it as small f. So, capital F would be difficult for us to understand, immediately I will come to that small f f f f. But, now it is a joint distribution of X 1 and X 2. So, it will be X 1 comma X 2, dx 1 dx 2.

So, this is integrated for all X 1 comma minimum, X 2 comma minimum. So, this is the minimum the maximum value: X 2 comma maximum, X 1 comma maximum. So, you can find it. Now, if you want to find out the relationship, co variance obviously is important. So when I basically do the co variance part, I will erase it again, I will come to those all details later on also.

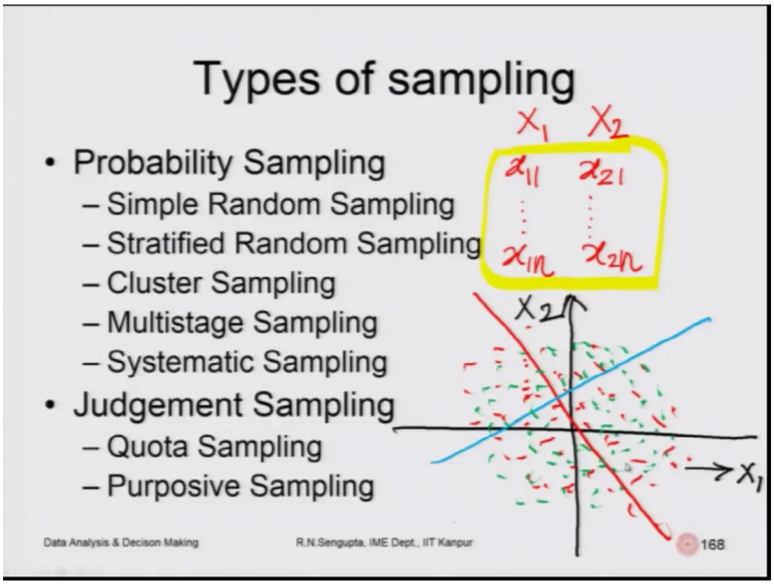
(Refer Slide Time: 17:09)



So, what I am trying to find out is basically the correlation that is the relationship between X 1 and X 2, so that is given by covariance. So, I am just writing the bracket this co variance between X 1 and X 2 divided by the square root of variance of X 1, which is the standard deviation of X 1 divided by standard deviation of 2.

Now, if I look at the co-variances between two random variables, so how would they look like, so and obviously remember. It can be proved that the covariance I am not putting the suffix, it can be between any two random variables is always this. So, inclusive of minus 1 and plus 1 it is always this.

(Refer Slide Time: 18:08)

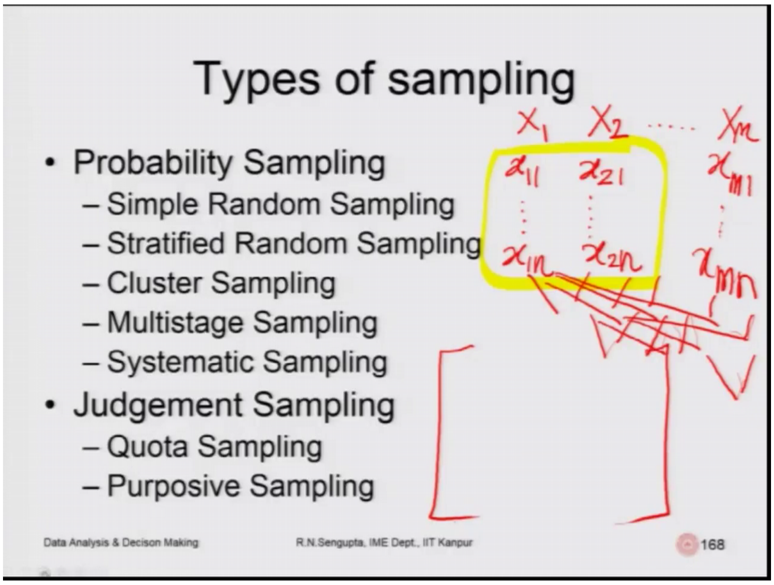


So, let me draw simply the diagram to make you understand. So, and use the red one, I am so black one to draw the coordinates. So, the coordinates are I am basically plotting X 1 along the X axis, X 2 along the Y axis. And now I will try to denote the correlation coefficient the rho value corresponding to the case, where it is negative it is 0, and it is positive.

So, I will use three different colors. The first let be for me for the negative one. So, if you find out the negative one, generally they would not plot it like this. So, the general trend line would be in this direction. If I plot the positive one, I will only draw the trend line for the positive one. So, this dot should be there. So, the positive trend line would be something like this. And if I do for the 0 case, that means correlation not there. Let me use the color green. So, they are everywhere. So, obviously they are spread over along the four quadrants such that you cannot say- what is the correlation value. It is almost equal to 0.

So, if I basically consider the correlation coefficient, it will give me what is the relationship. So, now the correlation coefficient is 0.5, it would mean the rate of change of X 1 and X 2 is of 1 unit to a half a unit. And if it is minus, it is 1 unit. It is minus 0.5, it is 1 unit to minus point half, it means if it increases for the negative case, if it increases the other decreases.

(Refer Slide Time: 20:36)



Now, consider that correlation coefficient values and the co variances. So, first let me consider the correlation coefficient values for n number of random variables. So, this n has nothing to do with the sample observations small n. So, again let me erase it. So, I am spending a little bit time here. So, I am sure you will you will bear with me. So, let me draw here let one; one, it I am containing it.

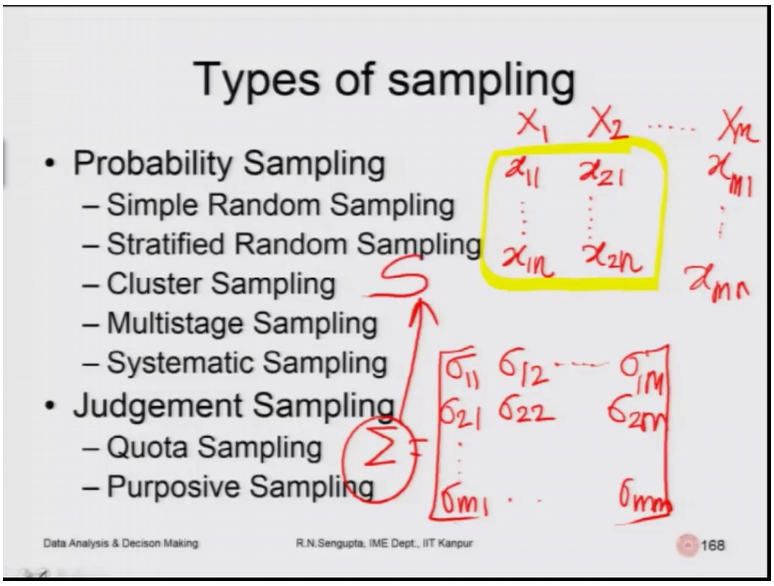
So, now come back to the sampling, which I drawn. So, consider there are random variables still X n. So, these are given by x oh my mistake. Let me consider this n as m, let me consider is that m. So, it is m 1 till x m. n So, m number of random variables each be are being set of n number of observations. m as in mango, n as in Nagpur.

Now, I want to find out the variance, covariance or the co variances between them. So, if I use that formula, which I which I just stated, how you find out the covariance between two random variables. And when I write them, it is basically an m by n by m by m matrix, because there are m observations. So, I try to find out the co variances between the 1st, the 2nd; the 1st, the 3rd; the 1st, the 4th; the 1st and the last. Then I find out between the 2nd and the 3rd, 2nd and the 4th, 2nd and the 5th, and 2nd and the last.

Similarly, as I continue doing it: I find out the m minus 1 and the m 1. And I basically find out the all the co variances. Now, when I am doing that co variances, obviously you will ask the question that what, if I am trying to find out the co variances between two different random variables, would not the covariance also exist between themselves. Answer is yes, they will exist. But, we have already considered that the co variances of a random variable with itself is the variance.

If you put on the formula, what you have actually is when you are trying to find out it is basically x minus mu whole square. So, if in the case for the co variances, if it is the formula is x minus mu whole square. So, in the case when it is a covariance in between two different random variables, it is X minus mu 1 into Y minus mu 2. Hence, when both X and Y are same. Hence, mu 1 and mu 2 would be same. Then we get the variances corresponding it to the co variances, where the co variances are between the same random variable.

(Refer Slide Time: 23:15)

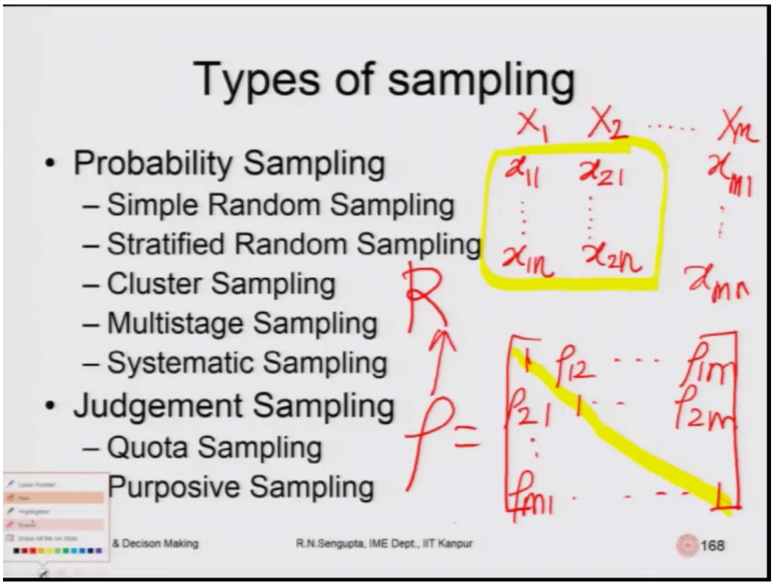


So, now when we write the m by m matrix that is mango m is mango not n as in Nagpur. The m by m matrix actually this denotes the following, let me erase the part else it gets no this should be there. So, the first variable, which is 1 comma 1 cell is sigma 1, which is the covariance of the first with itself. The second one along the row is a sigma 1 2, which is the covariance of the 1st to the 2nd, and I continue doing it till the last one, which is the covariance or the 1st with the m-th one.

Now, I come to the 2nd row. The 2nd one is sigma 2 1, so which means they are a mirror image of 1 2 and 2 1, because trying to find out the expected value of X minus mu 1, multiplied by Y minus mu 2 is equal to the same thing X finding the expected value of Y minus mu 2 into X minus mu 1. 2nd element on the 2nd row is sigma 2 2, which is basically the variance of the 2nd or as you can say put in another word is the covariance of the 2nd with itself, continue going doing it.

So, if I go to the last element it is m comma m 1, and the last one is m comma m. So, this is the variance covariance matrix. For the population, you will have a variance-covariance matrix for the sample also; similarly, when I come to the correlation coefficient. So, variance-covariance matrix, I will denote by this for the population. And this gets replaced by the by the variance-covariance matrix corresponding to the sample as s capital S, these are capital S.

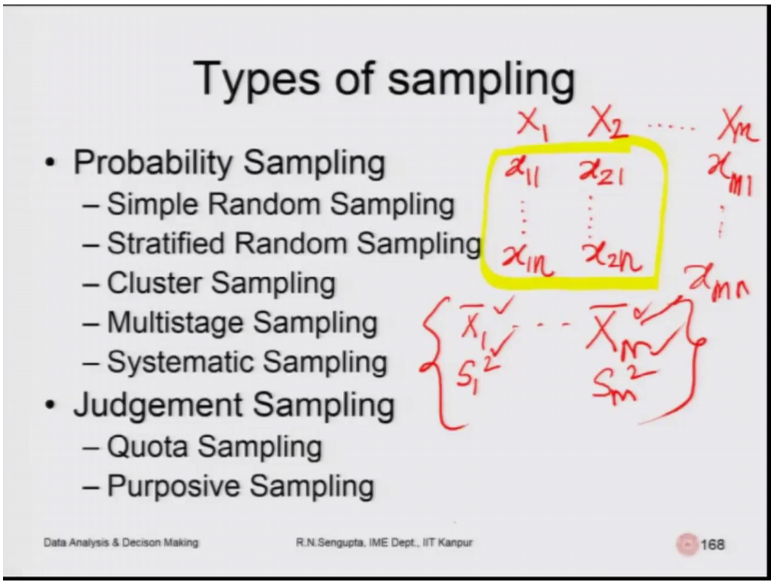
(Refer Slide Time: 25:05)



When I replace and try to find out the correlation. So, let me remove this whole part, everything let me remove. I will again draw it. So, it will become easy for you. So, now I am trying to draw the correlation matrix for m cross m. So, first element obviously, correlation is basically of the first random variable with itself, so that relationships 1. The 2nd one is basically the correlation of the 1st 2 it the 2nd. And similarly the last one is this.

The second rows are accordingly. So, so correlation 1 2, 2 1 are symmetric, so they are same. The 2nd element is 2, because the correlation of the 2nd with itself is 1. Similarly this, last element is this. So, the principal diagonal let me highlight it, yes. The principal diagonal is one. And the half the diagonal elements are the correlation coefficient, which is symmetric, so that is denoted by in the case for propagation of it. For the population, and when it is the sample, it is denoted by R.

(Refer Slide Time: 27:17)



So, I think I can erase this part. I will I even of the slide is being shown here. I will discuss these details of simple random something and all these things more detail in the next class, because I thought it was important. So, hence I am going a little bit slow. And I hope, you are you are understanding it. Now, coming back to the observations what I have written down.

So, if you remember, I put a yellow highlighted here, so there was a reason. Now, when you have a sample, and you want to find out the correlation coefficient, co variances and all these things: what you do is simply, you consider the sample. Consider the averages of the sample as such, so they would be the sample averages. So, the sample averages would be given as X bar 1. Similarly till the last one is X bar m: so for this is for this.

The variances would be given by S 1 square. I will come to that later on S m square. So, they would be discuss in detail within few class. And the correlations would be found on find out the averages here. So, they would be sample averages given here. Find out the co variances the co variances, which the word is not covariance it basically square of standard error or the variance of the sample. So, they would be given here. And you can find out the corresponding correlation coefficient and the co variances also.

With this, I will end the 12th lecture, but the slide which is in front of us, I have not discussed anything about probability sampling, this I will only discuss in a qualitative framework. I will come to that in the 13th lecture and correspondingly other lectures. But remember, when I have been talking about the correlation coefficient, co variances, mean value, standard deviations, for those samples corresponding to the case, we will be discussing the sampling distribution.

I will again come back to that in much further details. Have a nice day. With this, I will close this net 12th lecture.

Thank you very much for attention. Bye.