Applied Econometrics
Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology - Kharagpur

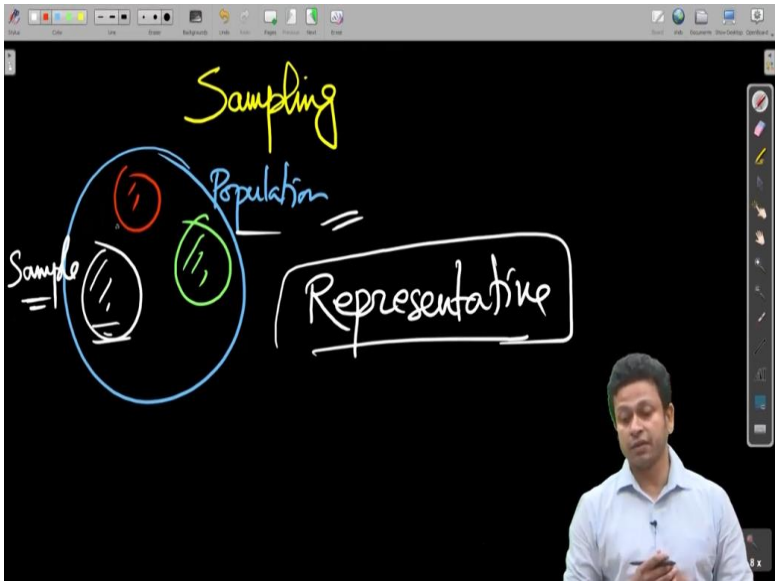
Module - 3
Lecture - 23
Sampling

Hello and welcome back to the lecture on Applied Econometrics. We are into the third week of our lecture. And this week, we are going to continue with the discussion on probability distribution. We are actually going to see some of the most important distributions that we are going to deal with; but before that, in this particular lecture, we are actually going to talk about sampling.

And sampling is something, we see the processes how we draw samples from a population, basically, broadly speaking; and how that really matters when we talk about the distributions. And more importantly, we are also going to talk about couple of natural laws in this lecture; one is central limit theorem, and another is law of large number. And we will see how the idea of sampling really matters to when we talk about the central limit theorem or law of large number, which are natural laws and which are really very important; because, with these, we will have certain assumptions, some relaxations, and we are going to see the importance of such laws when we construct the distributions.

Now, coming back to the topic of this lecture, which is sampling, we need to first understand what is sampling and why you do that. So, as you said, basically speaking, sampling is like to draw some samples from a population.

(Refer Slide Time: 01:43)



So, let us say, I have this big population here, and I want to draw a sample from here, let us say. So, this is my population and this is my sample. Now, before we actually go more in detail into it, first we need to understand why do I need to actually sample. I understand that I am trying to get some insight out of this population; but why not just look at the population, why to actually draw a sample and try to understand the sample, we can simply look at the population itself.

The reason why we do sampling is that, as such, getting entire data from the population is almost like, could be very difficult, it might be impossible in some cases and it is really prohibitively expensive, like it consumes lot of time, lot of manpower, and also the money of course is involved. So, to give you an idea, every country, they conduct the population census only once in a decade, and rest of the data need, where the government actually comes in and does a survey; so, it is basically survey.

For example, in India, we have National Sample Survey, National Family Health Care Survey, Human Development Survey and so forth. So, there are many surveys that is conducted. To give you an idea how, to what extent we actually do the survey; so, in India, for example, National Sample Survey, they usually for employment, unemployment consumption around, they usually would collect maybe information from around, let us say 400,000 households or 450,000 households in a country of like 1.3 million people.

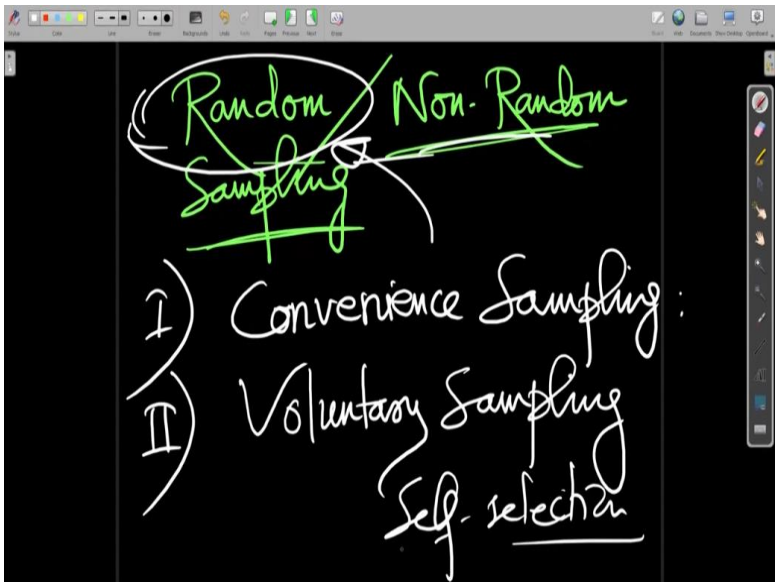
So, that basically shows we only deal with a very small fragment of the whole population when you are collecting data, and that is because, it is very difficult, almost impossible for us

to actually collect data from everyone, every time, because people simply will not give data, and as I said, it is really expensive. Now, we understood that, the reason we have to actually go for sampling.

Now, how do we ensure that my samples are actually telling the story that I want to hear from the population? As in the story that my sample is saying, is the same story that my population would have told me. So, I have to ensure that basically, that the sample is representative; the word we use is representative. Sample has to be representative of the population. The point is, how do I ensure that?

If I cannot ensure the sample as a representative of the population, so, it will tell something that is not true about the population. So, the whole purpose of doing the sampling would be defeated. So, I want to ensure that the sampling that I am doing for the purpose I am doing, that is actually hold true. Now, to actually understand how we go from sample to population or how we ensure this representativeness, we have different types of sampling, and it is a good idea I think, to actually start with those different ideas of sampling, and see how the sampling, the different sampling techniques are actually talking about different things.

(Refer Slide Time: 05:12)



So, let us say the first broad category of sampling is, one is random sampling and another is non-random sampling. Now, I will start with non-random sampling, just to give you an idea what is not a good sampling technique. Non-random sampling is usually not the best technique, usually you consider random sampling as a better technique. So, we will try to understand; it is like trying to understand the light, and it is in the absence of darkness.

So, we first explain the darkness; we will see what are the difficulties and challenges involved with non-random sampling; and from there, we will get to random sampling. So, let me give you a couple of examples to explain what is non-random sampling. So, this is one problem that you often see among our students. We call it convenience sampling. So, I will just explain that.

So, usually, we give projects to our students, where they need to do some sort of survey. They need to understand the, let us say, market pulse; let us say they want to understand the choice of people between online and offline market, when they are going to buy, let us say, some shoes. Now, what our students usually do? So, they will float a questionnaire; they will have a questionnaire; they will float that questionnaire around, maybe among their friends or their family or their past organisation or their childhood friends, social network they are active in.

So, they will kind of share this questionnaire among those people; or somebody can also go to LinkedIn; so, again, that is a network he has built over time. So, it is basically depending on his or her own convenience. So, I know these people; I am actually surveying these people. So, this is what we call convenient sampling. Now, the example I have given that, let us say, the behaviour of your regular use Crocs shoe purchase or some sort of like sandal you want to buy.

Now, there people, like different people actually may have very different choice. So, for a person who really is occupied with so many different works, so, he might simply not have enough time to go to the shop and actually buy a regular home based used or sandal or Crocs shoe. So, he can simply just go online and buy it. Whereas, in a rural setting perhaps or where people are more used to go to the brick and mortar shops, and they are not yet, have enough faith in the online market, because they might feel that these online companies might have been cheating them.

So, there are so many different reasons why people may not just go to online. And depending on the samples you have chosen, it can actually tell different stories. So, suppose if this student, he is coming from an urban setup, his friends and families and peers, they are all in an urban setup; so, then, what will happen is, we are going to see something that is basically

or totally representative of a group of people which is actually not talking about the entire population.

So, here, let us say these samples, they are basically rich people sample; and here it might be a poor people samples; here it might be your sample of middle class. So, all these different types of combinations or the compositions we have within that population. If I end up picking a particular group, then it will tell a story about that group; it will not talk about the whole population. So, that is a problem. So, then it is not representative.

In the convenient sampling, that is the reason it is not representative. Now, I will give you another example, and it is exactly opposite to convenience sampling; because, in convenience sampling, what you are doing is, you are actually, the researcher himself or herself is going out to people and deciding who is going to be in his or her sample set. And this one is called voluntary sampling.

Here, you are actually allowing the participants to voluntarily come and be a part of your survey. So, here, what people are doing? The candidates who are interested to be a part of your survey, they are actually self-selecting themselves. The word is self-select; self-selecting themselves; self-selection. So, they are deciding to be in that group. Now, who are the people who will decide to be in your group?

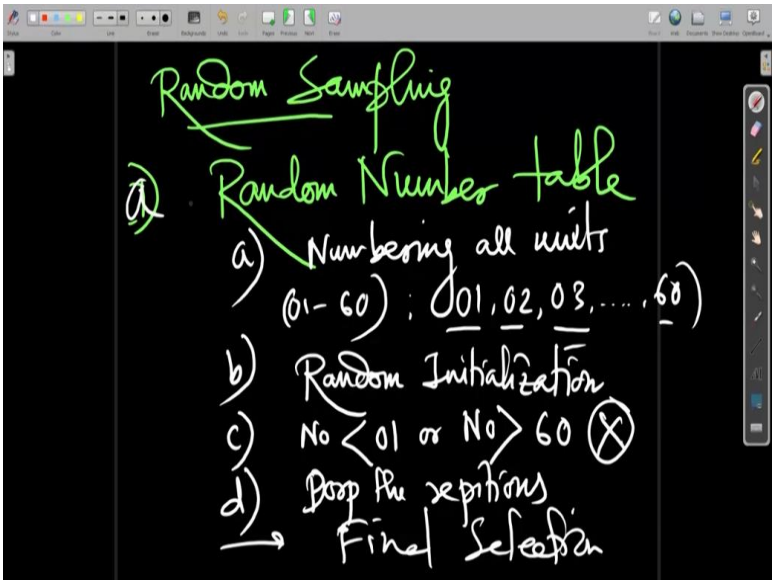
And depending on who are the people who actually wants to be a part of your survey, that can actually give you a story which perhaps is very far from the truth. And I can give you an example. Let us say you want to actually understand how busy people are in New York City at 10 a.m. in the morning, say, on Monday morning. And you decided to, let us say, stand next to a corner, and you saw all the people going, and you ask people whether they want to participate in a survey.

You want to understand if they are really busy on a Monday morning, 10 a.m. Now, most people say, no, I really have some work, I cannot really participate in your survey. And some people might just come in and they; yeah, I can definitely talk to you for a couple of minutes. And they will tell you that, no, perhaps Sunday is a very relaxed day for me, I just go around, I have nothing much to do.

So, the thing is, the people who really do not have much to do, they will come and actually answer you. And that is the reason; what will happen is, you will get a picture that Monday morning 10 a.m. New York City is really relaxing; I mean, people are really not that busy. So, that is actually going to give you a completely distorted picture, just because the people who have self-selected themselves, they perhaps are the people who really do not have much work; rest of the people actually really busy.

So, then, you can see how both these convenience sampling or voluntary sampling could be really misleading. And that is the reason we prefer a sampling called random sampling. And I am just going to explain what is random sampling.

(Refer Slide Time: 11:40)



And before I actually define it; random sampling; I will talk about the techniques; when I explain all the techniques, we will explain what is random sampling. We will go, get into the theory part after we see how it is done. So, there are different ways of doing it. The fundamental concept remains the same, but there are different ways of doing it. The first technique that I will talk about, it is like well established, economists have been using it, statisticians have been using it for very long; and that is basically random number table. And I have actually got a random number table for you here.

(Refer Slide Time: 12:29)

So, you can just simply search random number table in Google, and you will find these tables available there. So, this is a table, and I will just explain you in a while what this table is and how that is important, how that is helpful for ensuring my sample is random. Let us say, in our MBA programme we have 60 students, and we have suddenly recently launched a programme called, let us say, like Python bootcamp.

And I want to see, let us say, once they get into the job market, I want to see how this Python bootcamp is going to help them in getting a better salary. That is what my research objective is. And I am told that I need to ensure that there is no bias in the selection of samples, and that is why I have to select a random sample. So, that is what I am told. Now, I am also told; let us say, I do not know anything about random sample, but I am told that, you know what, you can actually use this random number table, and there is a way that you can ensure how you can have the sample which is random.

So, the procedure will go like this. So, first, let us say there are 60 students, and let us say I have to choose 30 among them. So, to do, basically, to use a random number table, I will just lay down the steps here. So, first, all the observations you have, the entire population. So, here, my population is my classroom, which is 60. So, entire population has to be ranked or has to be assigned a particular number.

So, numbering, let us say, numbering all units, let us say. So, here, I number all the units, all the students, and let us say I number them using their roll numbers. So, I have a number 1 to 60. Now, second; I will go to this random number table here, and what I will do is, I will

actually decide how many numbers I need to take. So, basically, I can take, let us say 3 numbers; let us say, I can take these 3 numbers; or I can take, let us say, only 2 numbers; or I can, let us say start from here, with 3 number; or I can start from there with 2 numbers.

So, there are all these possibilities are there. So I need to decide what I am supposed to do here. So, since I have my number of students are 60, which is a 2-digit number, so, I will only select; I will, perhaps it is a better strategy to take 2-digit numbers from the random number table. Now, so, to do that, perhaps it is a better idea when I select 1 to 60, to give a number 01. So, basically, the numbers are going to be 01, 02, 03, so forth, going to 60.

So, all the numbers, there will be 2-digits. Now, once it is done, from, in the random number table, what do we have to do? We have to select 2 numbers. And to select 2 numbers, we can actually initiate from any point. So, I can write it as a random initialisation. So, what it means is, I can just start from any random point. Let us say I want to start from, maybe the row number 12 and column number, this 5, 6.

So, row number 12, column number 5, 6 is, what is this number? 59, 89; it is not very clear. Let us say this is 89, let us say. So, first, I get 89; then I get 32; then I get 51; then I get 15; then I get 27; then I get 21; then I get 00; then I get 33. So, remember, I am taking 2 numbers at a time. Then I get 93; then I get 03; then I get 97, then I get 13, then I get 40, then I get 12. So, the idea here is, we will take those individuals who are, the numbers, these numbers are same as their roll number.

So, that means, in this selection procedure, I have to ensure that any number that is below 1 or below, above 60, we exclude those numbers. So, 00 will be excluded. Number below 01 or number above 60 are to be excluded, because we do not simply have those roll numbers present in our entire population. So, then, what we will select here is that, we will select 32, roll number 32. So, roll number 89 is not going to be a part of it, because 89 is above 60.

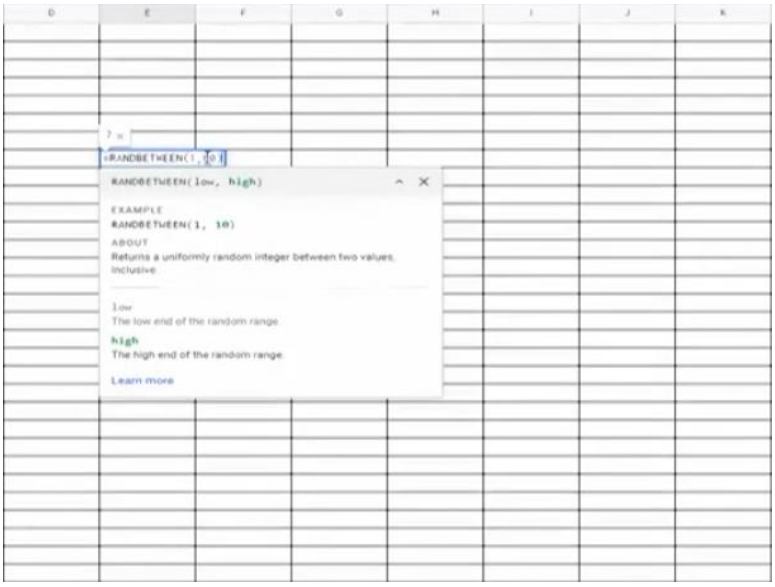
Roll number 51 is going to be a part of it. Roll number 15 is going to be part of it. Roll number 27 is going to be a part of it. Roll number 21 is going to be a part of it. 00, absolutely no point; 33, of course; 93 is not, because there is no 93. Then, we can have 03; 97 is not; 13 is; 40 is; 12 is; and so forth. So, you can have; if you want more students, which we want, so, we can go to the next line and you can keep on selecting like this.

You can also have repetitions. So, for example, you can have 13 twice, perhaps; we have to see, 13, it might appear again; and in those cases, you have to drop those numbers. So, that is how you basically select your individuals using random number table. So, it is absolutely that your roll number and this random number table, there is no correlation, nowhere they came from the same source.

And since the numbers are already there and you are using it that way, so, you can basically select the individuals. Basically, you drop the repetitions. And you create the final selection. So, we try to ensure that there is no way that there could be any bias from the researcher or from any other source when you are selecting individuals, because that is the only way we can ensure some randomness.

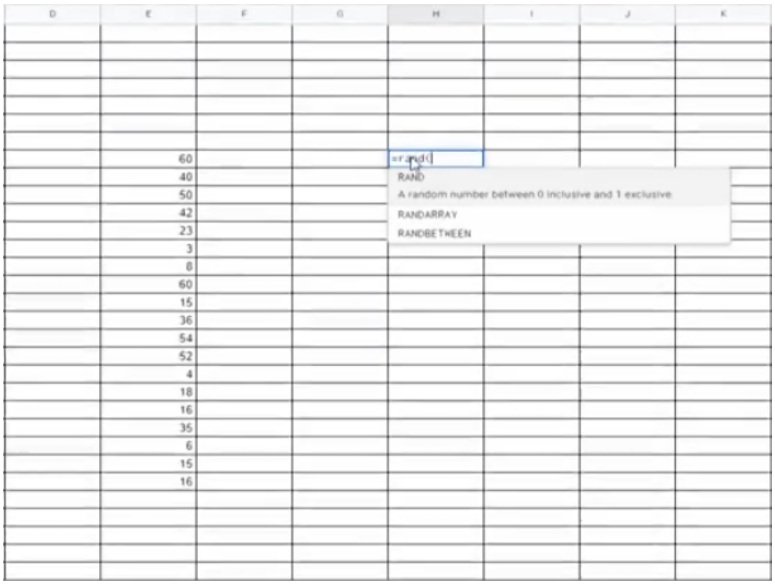
So, this is one procedure that people have been using for long; but once we have Excel available with us, there are other relatively easier way of doing it. So, let me just show you how using Excel, we can actually create some random numbers.

(Refer Slide Time: 19:35)



Let me just show; a Google spreadsheet, and I was just typing down this. So, there is a function in Excel. So, there is a function RAND, and there is another function RANDBETWEEN. So, let me first start with RANDBETWEEN. So, the idea of RANDBETWEEN is that, it will create random number between 2 whatever number, upper limit and lower limit that you provide. Let us say, I provide 1 to 60. So, it will create random number 1 to 60. So, let me first do that; 1 to 60. And I see what? 45.

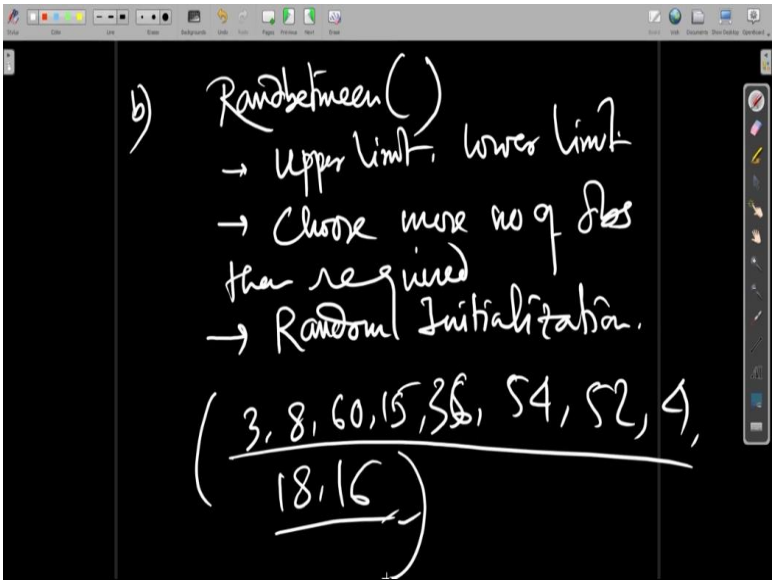
(Refer Slide Time: 20:18)



So, it will include both 1 and 60 for RANDBETWEENs, and it can also repeat the same number twice, like we have seen previously. So, and we can see, 60, 40, 50, 42, 23, 3, 8, 60, 15, 36, 54, 52, 4, 18, 16, 35, 6, 15, 16. So, basically, let us say you want to select not 60, but 10. And then, you will end up selecting roll 60, 40; let me actually write down looking at this table.

So, I will just take the first 10, or I can also decide to start from here; it is totally up to my decision, my choice; I can start from column, row number 13 or I can start from row number 8, whichever way I want. So, let us say I start from row number 13, and I go from here; so, 3. So, I will have;

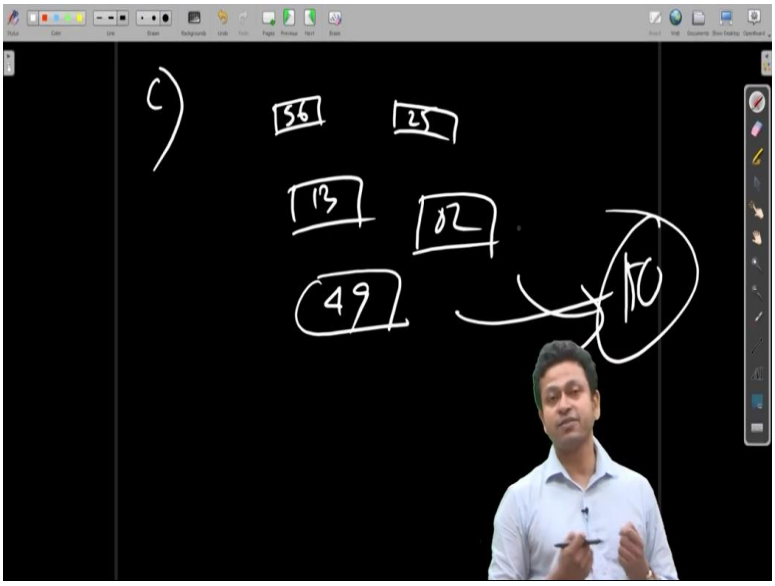
(Refer Slide Time: 21:23)



RANDBETWEEN function: So, this is a second way of doing. Here, I will have; a; this is my b. So, if I use RANDBETWEEN function, I will have; I have to ensure, I have to basically, the upper limit and lower limit, have to give. And since there could be repetitions, so, always choose more number of observation than required, because I might have to drop some numbers.

So, if I just go back to the table, and let us say I want to start from here; so, let us say I want 10 students; let me just write down; 3, 8, 60, 15, 36, 54, 52, 4; and then I have two more number, 18, 16. So, incidentally, we do not have any repetition, but there could be repetitions. So, you can basically random initiate here; random initialisation, just like before, and we can get all these different, you can get a final list of candidates that you want to actually deal with. So, this is how you use RANDBETWEEN. This could be a very simple way.

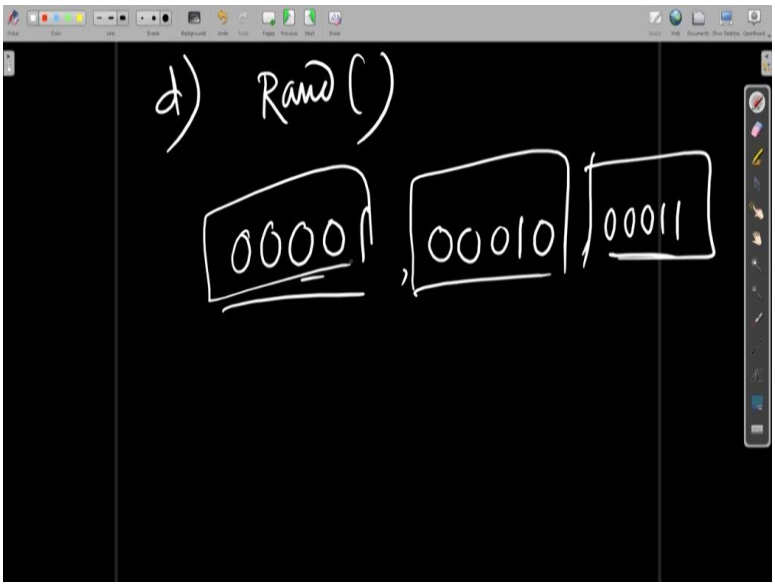
(Refer Slide Time: 23:19)



You can actually; what you can do is; perhaps, instead of doing all these things, you can simply write down the name of the students on a list basically, and you can create like small coupons. So, everywhere, you can just write the numbers 56, 25, 13, 02, maybe 49 and so forth; and you fold them, and like a lottery, you just shuffle them well inside a bowl; and then you decide to choose 10 of them.

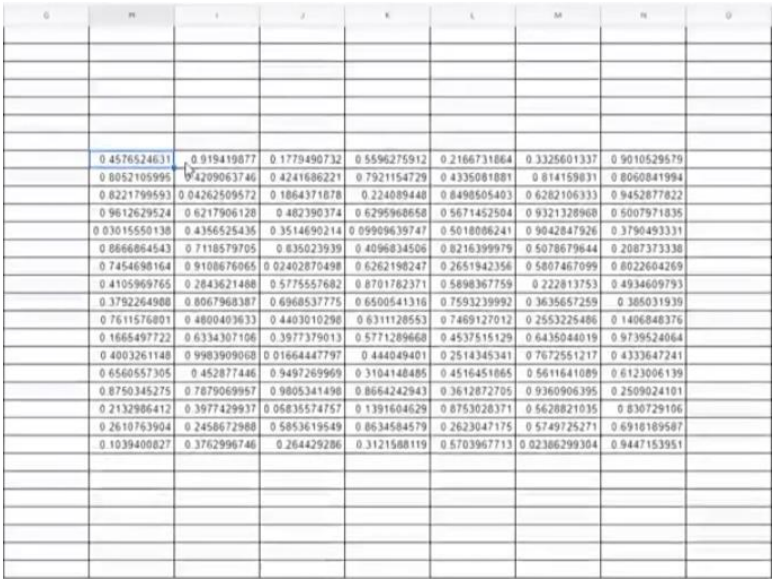
So, that is another way of first randomising. So, you can see, in all the processes, we are trying to see that, when we are choosing the individuals, we really do not have any hand in that; really, it is just happening on its own; it is just happening as like a natural choice.

(Refer Slide Time: 24:07)



Now, there is another way of doing it, and that is basically using the function RAND. And here, we simply do not write anything. So, how we can do it is, let me actually go back, and let me actually write down RAND. So, what it will do is, you just write RAND and then create an empty sort of function here; and what you will get is these numbers.

(Refer Slide Time: 24:40)



	0.45765246311	0.919419877	0.1779490732	0.5596275912	0.2166731864	0.3325601337	0.9010529679	
	0.8052105996	0.4209063746	0.4241686221	0.7921154729	0.4335081881	0.814159831	0.8060841994	
	0.8221799593	0.04262509572	0.1864371878	0.224089448	0.8498505403	0.6282106333	0.9452877822	
	0.9612629524	0.6217906128	0.482390374	0.6295968658	0.5671452504	0.9321328968	0.6007971835	
	0.03015550138	0.4356525435	0.3514690214	0.09909639747	0.5018086241	0.9042847926	0.3790493331	
	0.8666864543	0.7118579705	0.835023939	0.4096834506	0.8216399979	0.5078679644	0.2087373338	
	0.7454698164	0.9108676065	0.02402870498	0.6262198247	0.2651942356	0.5807467099	0.8022604269	
	0.4105969765	0.2843621488	0.5775557682	0.8701782371	0.5898367759	0.222813753	0.4934609793	
	0.3792264908	0.8067968387	0.6968537775	0.6500541316	0.7593239992	0.3635657259	0.385031939	
	0.7611576801	0.4800403633	0.4403010298	0.6311128553	0.7469127012	0.2553225486	0.1406848376	
	0.1665497722	0.6334307106	0.3977379013	0.5771289668	0.4537515129	0.6435044019	0.9739524064	
	0.4003261148	0.9983909068	0.01664447797	0.444049401	0.2514345341	0.7672551217	0.4333647241	
	0.6560557305	0.452877446	0.9497269969	0.3104148485	0.4516451865	0.5611641089	0.6123006139	
	0.8750345275	0.7879069957	0.9805341498	0.8664242943	0.3612872705	0.9360906395	0.2509024101	
	0.2132986412	0.3977429937	0.06835574757	0.1391604629	0.8753028371	0.5628821035	0.830729106	
	0.2610763904	0.2458672988	0.5853619549	0.8634584579	0.2623047175	0.5749725271	0.6918189587	
	0.1039400827	0.3762996746	0.264429286	0.3121588119	0.5703967713	0.02386299304	0.9447153951	

So, all the numbers you get here. So, perhaps, it is a better idea if we can, like we did for random number table; we can actually create something similar to that. So, here we can create completely random number table. And here also, like the previous one, we can actually see the numbers here 06; 04, 57, 65 and so forth, just the way; you basically create a random number table for yourself.

This is the RAND function and you can actually use that. Now, we are almost finishing the talk on this random sampling, and I am going to go to the theory part in the next lecture. But before I go there, there are few things we perhaps need to just mention is that, it could be, I have taken examples where I am taking only 2-digits, for example, the roll number of students; but it could be like 3-digit, 4-digit, 5-digit.

So, then what you do, usually you actually take something like maybe, if it is a 5-digit thing, you can take 00001; 00010, like it was 10; if it is 11, 00011; something like that. So, we basically ensure that the length of the numbers are going to be constant. We basically pick the same number of digits, whenever you pick the numbers, depending on how many numbers you need.

So, if there are that many numbers, so, you have to ensure that you maintain the consistency. So, with this, we will end the lecture. We have basically discussed about; we introduced sampling in this lecture; we talked about the non-random sampling first; and then we have spoken about the techniques of random sampling. And in the next lecture, we are going to talk about different types of random sampling and how they are different from each other, and particularly, we will emphasise what is simple random sampling. With this, we end the lecture here. Thank you.