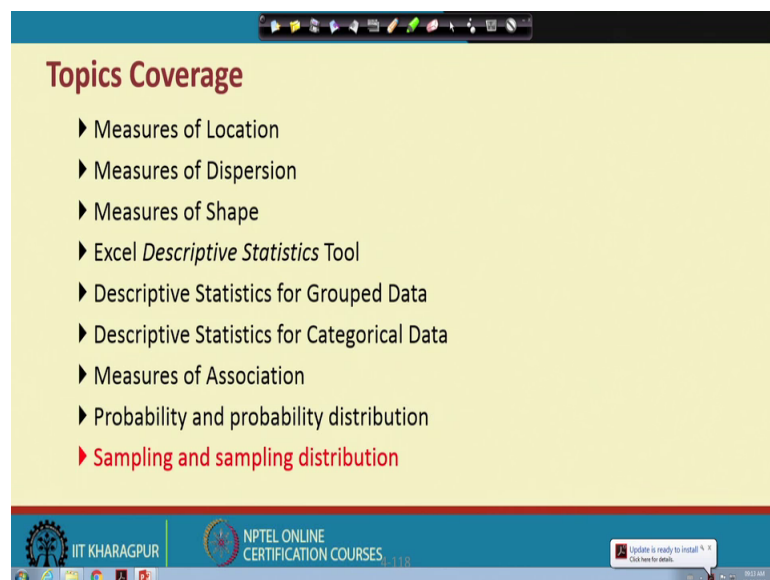


**Engineering Econometrics**  
**Prof. Rudra P. Pradhan**  
**Vinod Gupta School of Management**  
**Indian Institute of Technology, Kharagpur**

**Lecture – 15**  
**Descriptive Econometrics (Contd.)**

Hello Everybody. This is Rudra Pradhan here. Welcome to Engineering Econometrics. Today, we will continue with the Descriptive Econometrics and that to the coverage on sampling and sampling distributions.

(Refer Slide Time: 00:34)



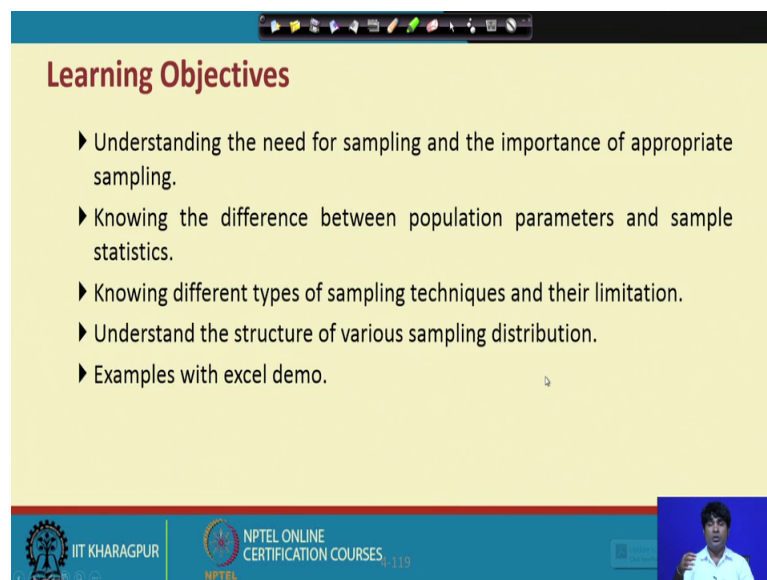
In fact, in this you need we have already covered couple of things, starting with measures of locations, measures of dispersions, measures of shape, and we have also connected to excel spreadsheet by using data analysis package. Then, we have covered descriptive statistic for group data descriptive statistic for categorical data and also we have discussed measures of association that to the need of covariance and correlation. And, in the last lectures we have discussed the requirements of probability and probability distributions. And, in this lectures typically we like to highlight sampling and sampling distribution.

Because, these are the components very much required for engineering econometrics and that to the use of regression modelling, for any kind of mean say is any kind of problems related to engineering in econometrics. And, in this lecture typically we like to highlight

how sampling and sampling distributions are so, important in the process of solving some of the engineering problems, by using engineering econometrics tools.

So in fact, we have already a discussed in details the need of the probability and probability distributions. In the same way, we like to highlight sampling and sampling distribution, because these are all very much interconnected. And without to having probability distributions it is very difficult to connect sampling distribution. Again, without knowing sampling distributions the concept of probability distribution is also meaningless, because, our requirement is to solve some of the engineering problems by using engineering econometrics tools. And these are the basics through which we can use directly and indirectly to work out the solution for some of the engineering problems.

(Refer Slide Time: 02:49)



**Learning Objectives**

- ▶ Understanding the need for sampling and the importance of appropriate sampling.
- ▶ Knowing the difference between population parameters and sample statistics.
- ▶ Knowing different types of sampling techniques and their limitation.
- ▶ Understand the structure of various sampling distribution.
- ▶ Examples with excel demo.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 119

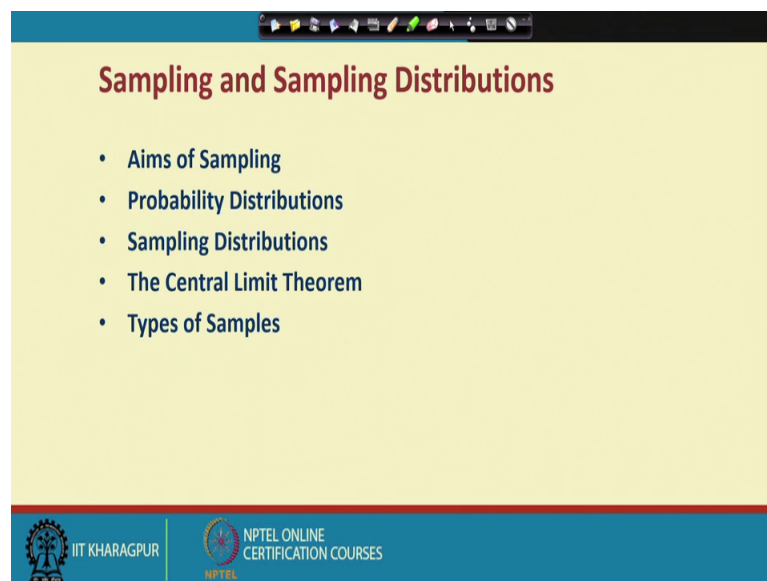
So, what we have already discussed? In the last couple of lectures, that there is a need of a sampling structure and in fact, there are 2 issues. So, the issue is what is the sample size? And, that to corresponding to the populations and what is the typical sampling distribution corresponding to the probability distributions?

So; that means, here the here the game is between sample and population. And, most of the cases we like to we would like to solve the engineering problems, by using some of the sample case only, because it is very difficult to go for population as a kind of representative to analyze the engineering problems.

So, what we what we supposed to do? We have to we have to select a particular samples or we have to pick up a particular sample to analyze a typical engineering problems. So, what is the requirement here to understand the need of need for sampling? And, the importance of appropriate sampling and then knowing the difference between population parameter and sample statistics. Against, we know we are interested to know different types of sampling techniques and their limitations. And, understand the structure of various sampling distributions and then we like to connect with the some kind of excel spreadsheet to a handle the problems relating to particular sampling distribution.

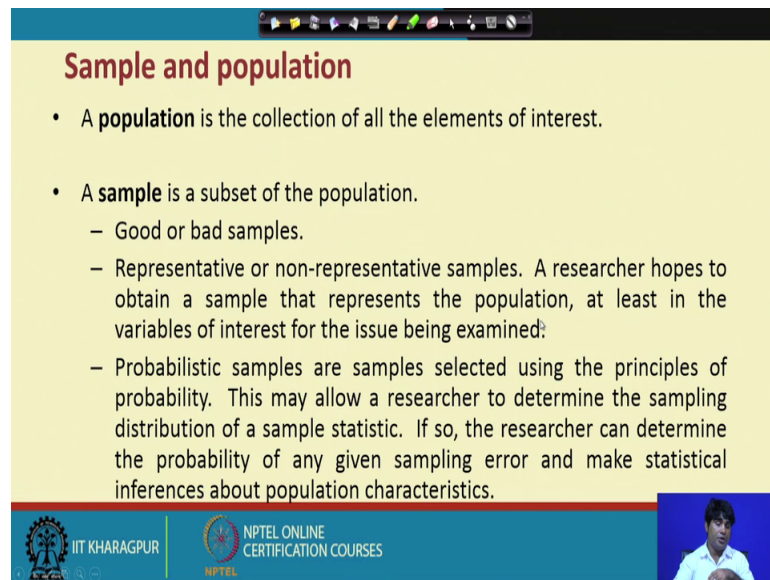
In the excel spreadsheet in fact, we have a kind of structures, where we can actually generate as random sampling as per the particular engineering problems requirement. Sometimes, we can generate the data then look for the kind of solutions, and sometimes we like to you not use the actual data followed a followed by a particular sampling technique to solve some of the engineering problems.

(Refer Slide Time: 04:57)



So, now what is happening here? So, these are the things we are supposed to highlights, aims of sampling, probability distribution, sampling distributions and the concept of central limit theorem and types of sampling. In fact, central limit theorem will give you the clue about the, optimality of the sampling structure to work out the solutions for the engineering problems.

(Refer Slide Time: 05:25)



### Sample and population

- A **population** is the collection of all the elements of interest.
- A **sample** is a subset of the population.
  - Good or bad samples.
  - Representative or non-representative samples. A researcher hopes to obtain a sample that represents the population, at least in the variables of interest for the issue being examined:
  - Probabilistic samples are samples selected using the principles of probability. This may allow a researcher to determine the sampling distribution of a sample statistic. If so, the researcher can determine the probability of any given sampling error and make statistical inferences about population characteristics.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, in fact, population is the set of all elements; like you know certain subsets so, we have the game between population and sampled. So, collection of all samples or sum of all samples we can say that population. And, within a particular population basket when we pick up particular samples, then it is called as a sample specific you know investigation and sample specific analysis. For instance, if you say that the production of a manufacturing industry as a population, then a particular industry production structures will take you to the concept of sampling.

So, what did it do? We have to pick up a particular sample or a particular case to analyze the particular problems. So, it is basically the generalization process of sampling. So, that is why we cannot target population directly. So, we have to pick out different samples, then you check the results on the basis of these sampling results we can comment about the population parameters. Whether, it is a perfect or not perfect, whether there is a need of change or need to be a constant. So, these are the things continuous process over the time we have to investigate and then every time we have to check the particular requirement.

So, sometimes I mean say the issue of sampling is you, whether it is a good sampling and random sampling the size of the sampling. So, there are typical structures or there are specific rules and regulations to optimize these things. If your sample specification or sample selection is not perfect then the choice of a particular technique or the solution

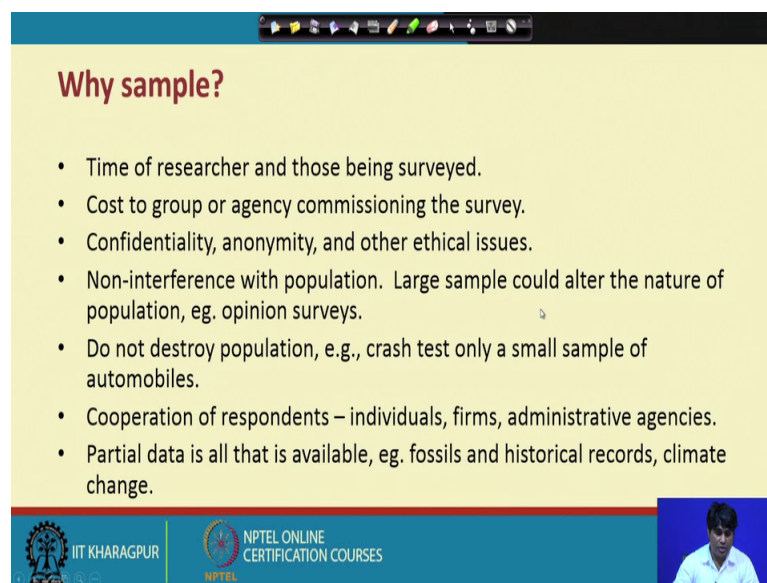


to a particular engineering problems may be not correct it will give you some kind of bias results.

So, that is how we are means we must be very careful about the sampling selection and the process of sampling. So, there are 2 things very important. So, what is the size of the sample and only means what is the basis of collecting these samples. So, these are very important. So, once these 2 things are taken care of then data side the data side, it is not a kind of problem.

So, we have to be very careful how to pick up a good sampling; with respect to size of the sampling and the collection of data for instance random sampling non random sampling. So, within the particular sampling we have a couple of structure. So, we have to be very careful how we have to be a means; we have to deal with all these you know issues and the kind of requirements.

(Refer Slide Time: 08:33)



**Why sample?**

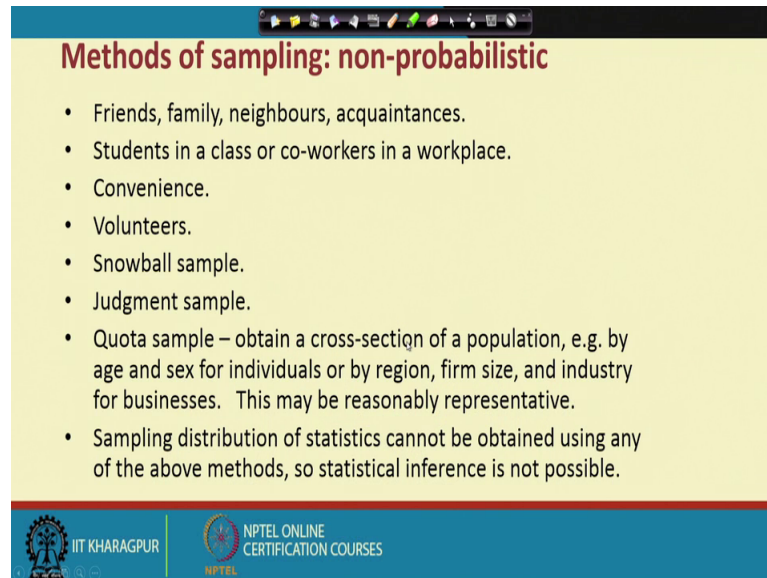
- Time of researcher and those being surveyed.
- Cost to group or agency commissioning the survey.
- Confidentiality, anonymity, and other ethical issues.
- Non-interference with population. Large sample could alter the nature of population, eg. opinion surveys.
- Do not destroy population, e.g., crash test only a small sample of automobiles.
- Cooperation of respondents – individuals, firms, administrative agencies.
- Partial data is all that is available, eg. fossils and historical records, climate change.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, let us see what are the process; and, a first of all why samples? As sample is a very good representative to the populations to give some kind of comments. So, it leads to confidentiality the other typically issues like you know a ethical issues. So, a noninterference with population then large sample could alter the nature of the populations like you know opinion surveys. And, like that we have a couple of things we have to calculate of course, it is not a kind of constant process it is a dynamic process.

So, every times when there is a kind of means when something wrong we can a change the particular structure and then pick up the correct sample correct samples and then we go for the analysis. Means a what we can say that it is a dynamic process? And until unless you get a kind of solutions which is very perfect as per the particular engineering a econometric requirement or engineering problems requirement.

(Refer Slide Time: 09:55)



**Methods of sampling: non-probabilistic**

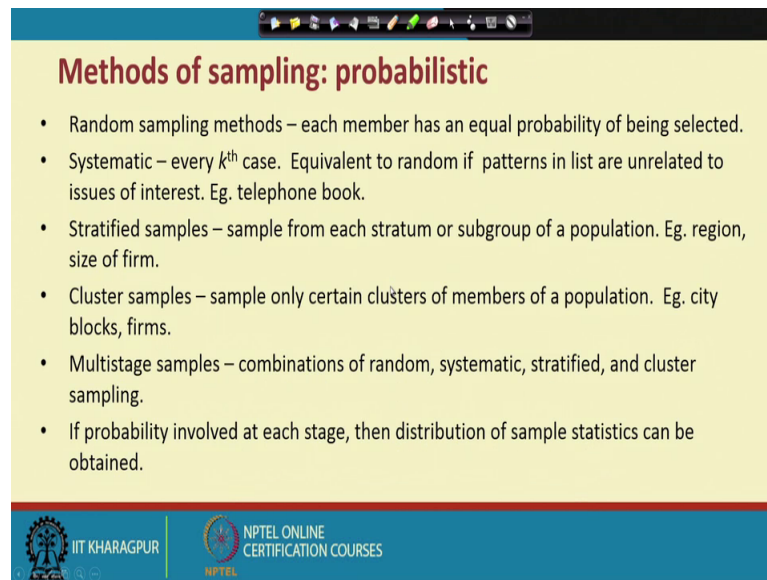
- Friends, family, neighbours, acquaintances.
- Students in a class or co-workers in a workplace.
- Convenience.
- Volunteers.
- Snowball sample.
- Judgment sample.
- Quota sample – obtain a cross-section of a population, e.g. by age and sex for individuals or by region, firm size, and industry for businesses. This may be reasonably representative.
- Sampling distribution of statistics cannot be obtained using any of the above methods, so statistical inference is not possible.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, sample and sample sampling selection is very important. So, far as are solving any kind of engineering problems. So, here there are you know methods. So, there are 2 types of methods typically; probability sampling and nonprobability sampling.

So, in the case of probability sampling, we have a connection to probability and in the case of nonprobability sampling. So, we have non-connection to probability. In fact, we have already discussed the issue probability and probability distribution and a both probability and probability distributions you know they have whole in the case of sampling and sampling distribution.

(Refer Slide Time: 10:30)



**Methods of sampling: probabilistic**

- Random sampling methods – each member has an equal probability of being selected.
- Systematic – every  $k^{\text{th}}$  case. Equivalent to random if patterns in list are unrelated to issues of interest. Eg. telephone book.
- Stratified samples – sample from each stratum or subgroup of a population. Eg. region, size of firm.
- Cluster samples – sample only certain clusters of members of a population. Eg. city blocks, firms.
- Multistage samples – combinations of random, systematic, stratified, and cluster sampling.
- If probability involved at each stage, then distribution of sample statistics can be obtained.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So you must be very careful about the particular structure. So, methods of sampling particularly in that case of probabilistic, we have random sampling methods.

So, here you to randomly pick up a particular sample to investigate a particular engineering problem. And, in my knowledge this is the best sampling selection means procedure of selecting samples to analyze some of the engineering problems, because it will give you some kind of unbiased kind of inference. If you do not follow random sampling then there is a high chance that there is a bias in the systems, but in order to avoid all these obstacles it is better to be cover random sampling rather than you know a specific you know sampling structures.

In fact, if it is required then you can follow that, otherwise a it is always you know suggestive to you know follow random samplings so, for as a size of the sample is concerned and the process of sampling is concerned

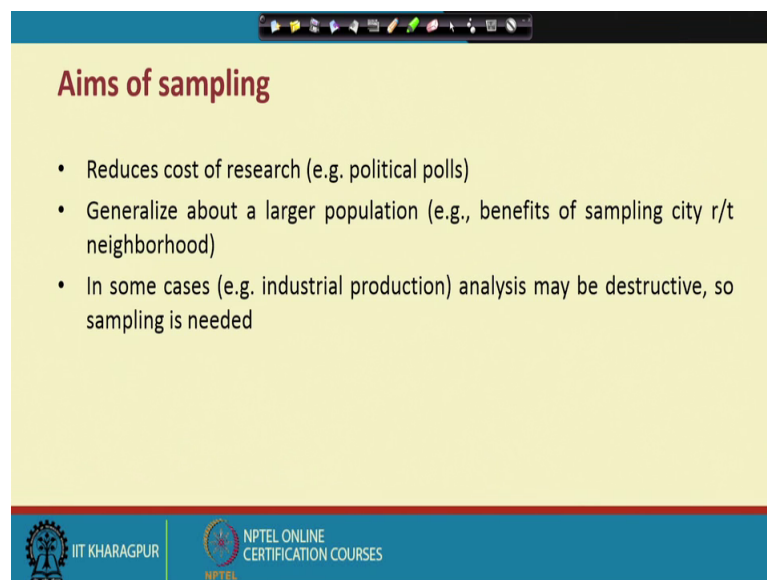
So, there is a structure called as a stratified sampling here the sample is a derived from a particular subgroups of a population. And, then there is a cluster sampling and here sample only on the basis of certain clusters of members of the populations, like city blocks, you know sectors etcetera. And, multistage sampling it is a combination of random systematic stratified and cluster sampling.

And a finally, a probability involved at each stage then the distributions sample statistic can be obtained. So, that means, these are all various sampling procedure altimetry whatever procedure you have to follow. So, there is a kind of process of a random one and that to this selection should be completely unbiased. Otherwise a otherwise a the data structures may not be correct and even if you select a good technique still if the result should not be unbiased in fact, it is not perfect.

So, that is how you must be very careful? So, while selecting a samples and the process of sampling. So, the things the thing is very clear here. So, if your input selection is very perfect, then the output that is engineering econometrics output will be very perfect. If you are input process is not correct then; obviously, the outproce output process will be also defect. So, that is why you must be very carefals how to choose the particulars you know structures. So, for is a sample selection is concerned and the size of the sampling is concerned.

Of course, we will give you details when you work out the particularly you know problems by using any of the engineering econometrics technique.

(Refer Slide Time: 13:28)



**Aims of sampling**

- Reduces cost of research (e.g. political polls)
- Generalize about a larger population (e.g., benefits of sampling city r/t neighborhood)
- In some cases (e.g. industrial production) analysis may be destructive, so sampling is needed

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

The aim of sampling is to reduce the reduced cost of the research. Of course, if you go for you know population as here kind of requirement to analyze the problem. So, or the cost will you know definitely very high. So, in order to reduce the cost it is better to take

few samples, analyze these samples. And then generalize and that will give you the kind of inference towards the population.

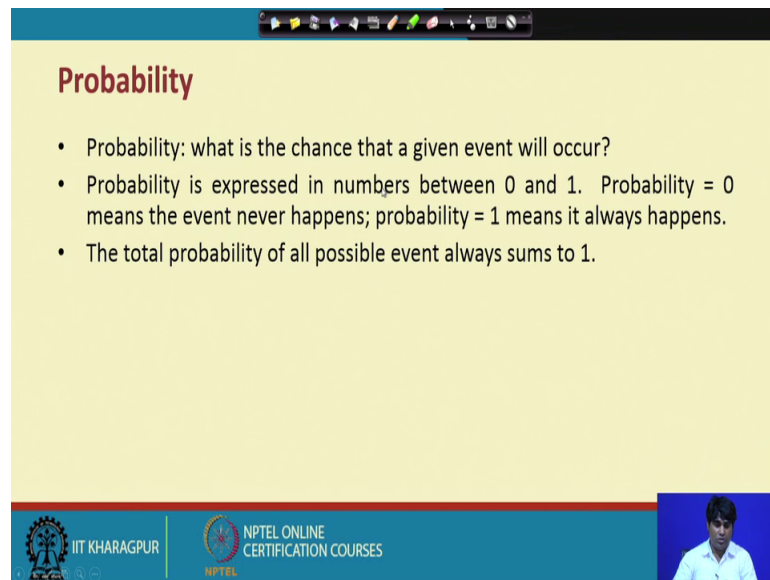
So, it is a process of generalizations about a large population. So, in some cases analysis may be destructive. So, sampling is very much needed. So, that means, technically. So, we means the suggestion is that we should go by a sampling process or you know sample selections and they kind of analysis, not the population specification. And, sometimes it is very difficult for a researcher to gather or all sorts of data to solve the particular engineering problems. If that is possible or that can be done so, that is called as population kind of inference, or else if you could not then it is better to take this a different samples and then connect to the population. So that means, even if you select some you know some of the samples.

Let us say a basket one basket 2 basket 3 depending upon the number of sample points say let us say 10 15 or something like you know 100, then you can you can analyze these samples individually and against you can pull these samples and again analyze as per the requirement. So, that means, that technically you have a sample specific analysis and against the kind of sampling distribution kind of analysis.

So, like we have actually a time series data, cross sectional data, then pulled data, final data. So, here also sampling procedure is also like that. So, we have a different you know cross sectionals that, that is what the sampling kinds of sample selection and then when you clog it then it becomes called as you know population.

So that means, technically you are moving actually simple case to little bit means a cross sectional to pull. So that means, is small sample to large samples and having a different kind of diagnostics or kind of shapes to analyze the problems as per the particular engineering requirement.

(Refer Slide Time: 16:12)



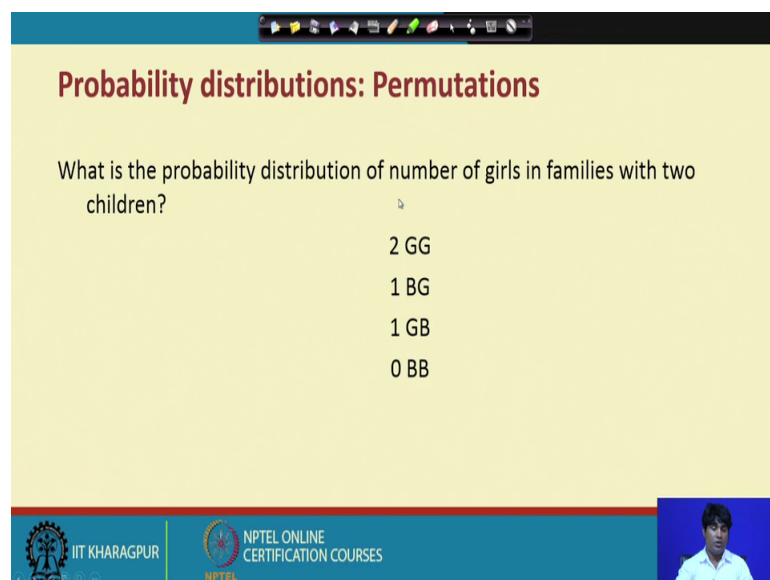
**Probability**

- Probability: what is the chance that a given event will occur?
- Probability is expressed in numbers between 0 and 1. Probability = 0 means the event never happens; probability = 1 means it always happens.
- The total probability of all possible event always sums to 1.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, this is what actually probability? So, what we have already discussed, it is a kind of chance that a given event means the event will occur. And, usually the probability value will be 0 to 1 and a, the sum of probability exactly equal to 1. So, with this you know I means conditions, we have to you have to select a particular samples and then the sampling distribution.

(Refer Slide Time: 16:41)



**Probability distributions: Permutations**

What is the probability distribution of number of girls in families with two children?

- 2 GG
- 1 BG
- 1 GB
- 0 BB

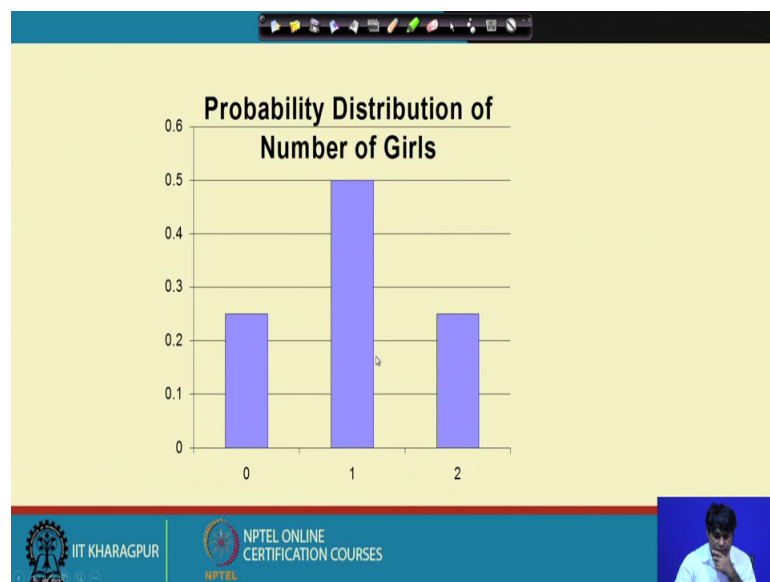
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, so, for instance let us say Probability Distributions. And, here what is the probability distribution of number of girls in a family with it 2 children's right?

So, means so, there are 2 family and what is the probability distribution number of girls?  
So, that means, this distribution will be either there is no girls or both the family having  
you know girl so, 2 extremes. So, if both the family having girls then it the starts with  
you know GG. So, like here and again if there is in no such you know happens then it  
will have a 0.

Otherwise in one family one girls and one boy again the reverse is also true. So; that  
means, there are 4 such possibilities can occurs obviously while analyzing this kind of  
problems or issues. So, permutation combinations you know they play very a excellent  
roles to create a kind of distributions as per the particular engineering requirement. This  
is how the distribution?

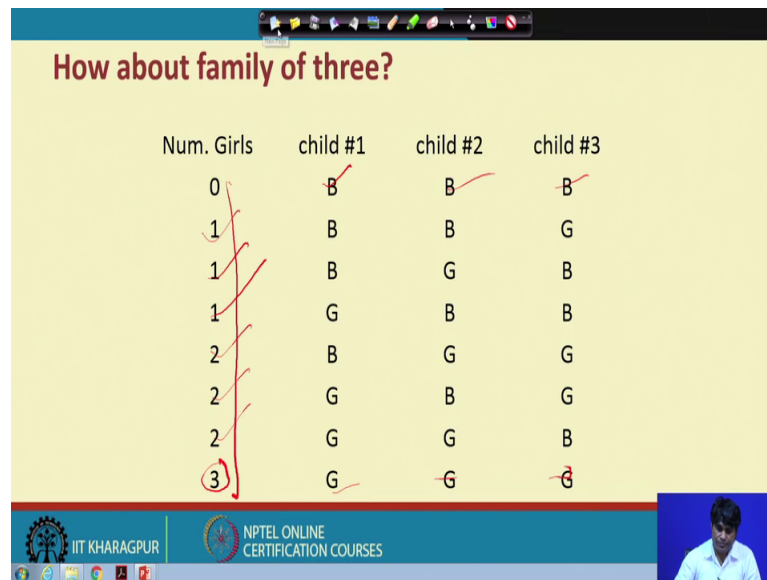
(Refer Slide Time: 17:59)



So, like you know and this is how 0 1 1 2, so then ultimately this is how the  
distributions?



(Refer Slide Time: 18:02)



Num. Girls	child #1	child #2	child #3
0	<del>B</del>	B	<del>B</del>
1	B	B	G
1	B	G	B
1	G	B	B
2	B	G	G
2	G	B	G
2	G	G	B
3	G	G	G

And to you know make it more you know interesting then you know you can extend this particular process you can start with it 2, 3 family, 4 family, 5 family, then the process released you know moves from simple to complex. So, if you are starting or the minimum requirement or minimum kind of structure is very perfect then extending the kind of processes you not so difficult.

So, so; that means, we start with the simple structure, then you know generalize it or you can extend it to you know as per the particularly you know requirement. For instance if there is a family of 3, then it starts with you know the one extreme will be GGG 3, then the 0 0. So, that means no girls and all girls.

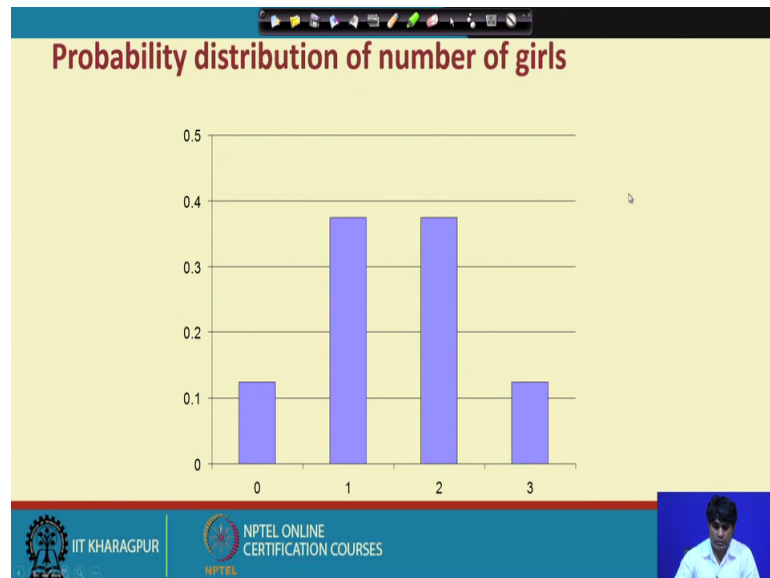
So, that means, in between 1 family having 1 girls and other 2 families having you know no girls then there are many different you know instances all together. And, these are the things are; here like you know what I can say that? So, this is these are all you know possible cases. So, maximum will be 3 girls from 3 family and minimum will be no girls that is you know 0 so, BBB.

So, it between there may be one girls out of this 3 family, then that depends upon you know each family against family 1 family 2 family 3. So obviously, since there are 3 families so, 1 1 case will be 3 means 3 such cases against there is a high chance that 2 girls you know out of these 3 families. So, there are 3 cases against. So, ultimately so, we have 8 different cases. So, it depends upon you know the kind of structures ultimately,

how many such cases can of course, it depends absolutely in the form of a concept collagen 2 to the power N.

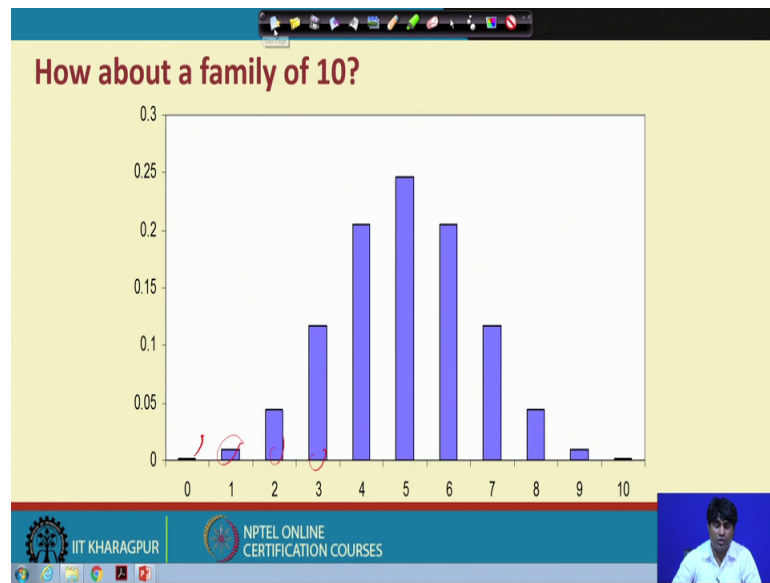
So, obviously, we can have the kind of structure and create a distributions.

(Refer Slide Time: 20:17)



So, this is again if you plot like this, this is a 0. And, that means, if this is what the probability? And, against in the middle case 1 1 girls into 2 girls it is more or less same that is you know 3 3 situations, then this is also same case. So, as a result in mid in the middles both are same and the left skewed and right skewed the kind of probability is also same and if you add all these probability then it becomes you know one.

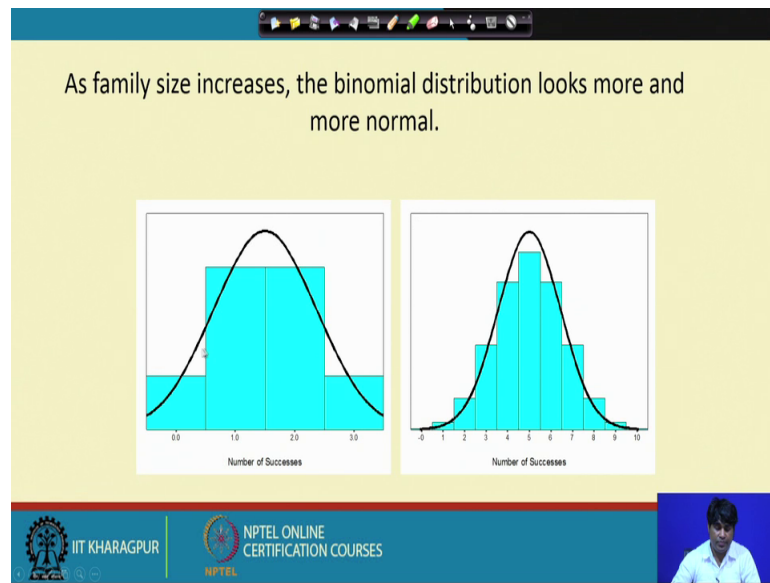
(Refer Slide Time: 20:49)



So, again how is the structure of 10 family? So, that means, it will extent like this. So, against the minimum will be 0, then maximum will be a you know tense that is one such case then in between we have a 1 girl case, 2 girl case, 3 girl case likewise we have a various kind of options you know starting with you here say you know this is actually 0, then 1 girl case 2 girl case 3 case.

So, these are all number of possibilities. And, then if we really actually plot then it will generate a particular distributions. In fact, if you look this distribution it look is looking like you know normal distribution. And, it is good for the particular analysis.

(Refer Slide Time: 21:37)

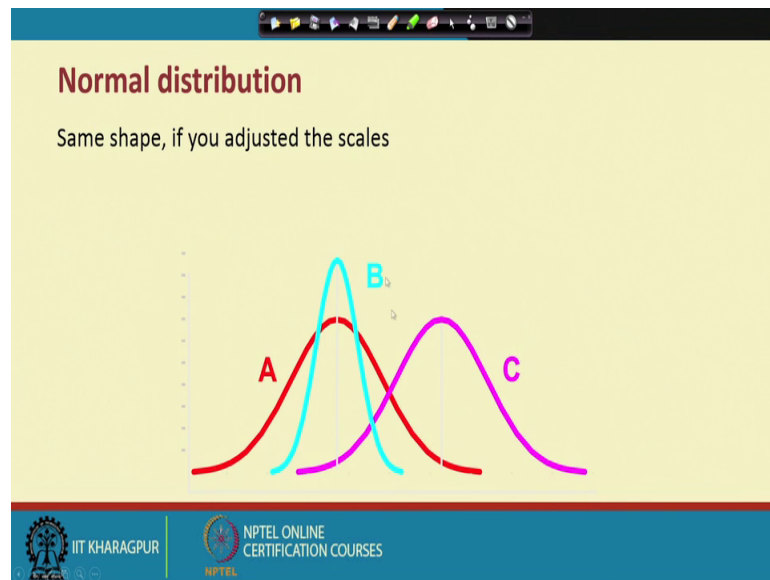


So, likewise you can actually extend these problems. And, then if you join midpoints of this particular probability that is what we call for technically called as frequency, then you will get a normal distribution curve. So, so, ultimately when we had in the process of in a analyzing the problems, we like to you know have a samples or you know size of the samples and the kind of distributions, it should be a perfect one and it should be as per the particular requirement. Usually normally distributed some sampling or sample is very good one for any kind of analyses.

So, we try to you know structure restructure in such a way the particular process should follow the normal distribution, that is what we call as symmetrical distribution? And, if it is possibly you can try if not possible then you can compromise and then you connect with the different I mean other different distributions, but it is always very good so, if you the process of sampling or the kind of intersection or sampling should of follow the normal distributions.

But, we have a, but at the same times we have the kind of structure, if it is not normally distributed then how do you analyze this problem? So, that is how we have a different distribution? And, in addition to normal distribution you should know other distribution simultaneously and connect the modeling a accordingly as per the particular requirement.

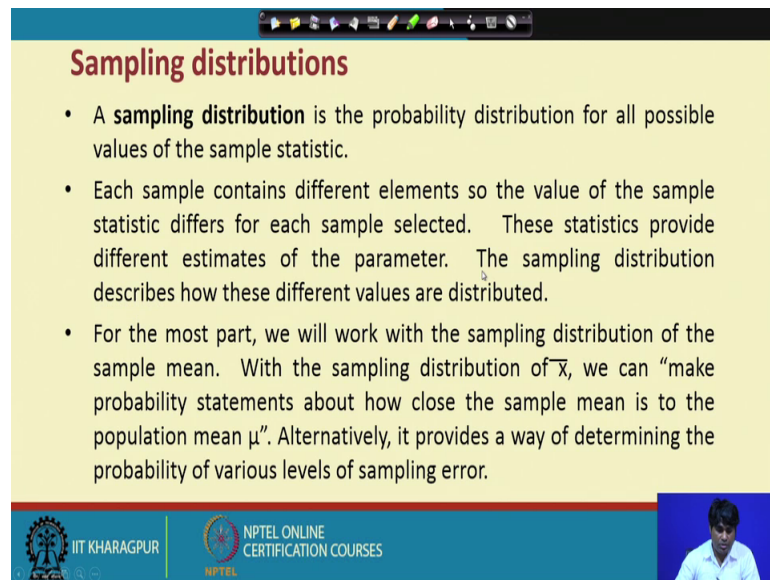
(Refer Slide Time: 23:14)



So, this is another way of looking at the particular distributions. All are known to be normally distributed, but they differ with respect to their shapes, that is what we have already indicated with the help of Skewness and Kurtosis. And so, sometimes I think the red one is a very good one. And, this is a little bit of a very kind of as different kind of shapes all together.

And, in fact, you must be very careful how you have to deal with the situation? In fact, in all these cases it follows the normal distributions, but which particular shape is perfect depends upon the kind of means and depends sample to sample and the kind of the size of the sampling.

(Refer Slide Time: 24:19)



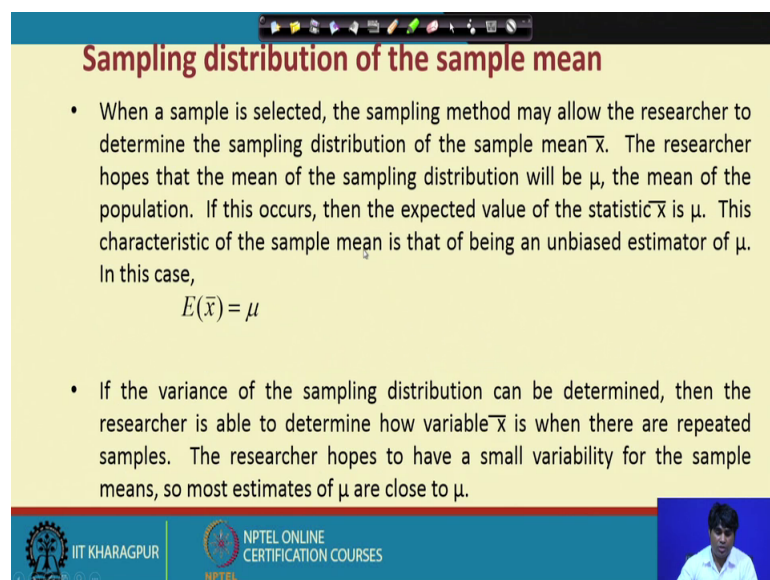
### Sampling distributions

- A **sampling distribution** is the probability distribution for all possible values of the sample statistic.
- Each sample contains different elements so the value of the sample statistic differs for each sample selected. These statistics provide different estimates of the parameter. The sampling distribution describes how these different values are distributed.
- For the most part, we will work with the sampling distribution of the sample mean. With the sampling distribution of  $\bar{x}$ , we can “make probability statements about how close the sample mean is to the population mean  $\mu$ ”. Alternatively, it provides a way of determining the probability of various levels of sampling error.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

Sampling distribution is the probability distribution for all possible values of the sample statistics. So, that means, it is actually a extending kind of things. So, we start with this as a samples, then the cluster of all samples called as a sampling distributions, then the cluster of all sampling distribution will called as population, this is how we have to extend our you know understanding and extend our analysis.

(Refer Slide Time: 24:49)



### Sampling distribution of the sample mean

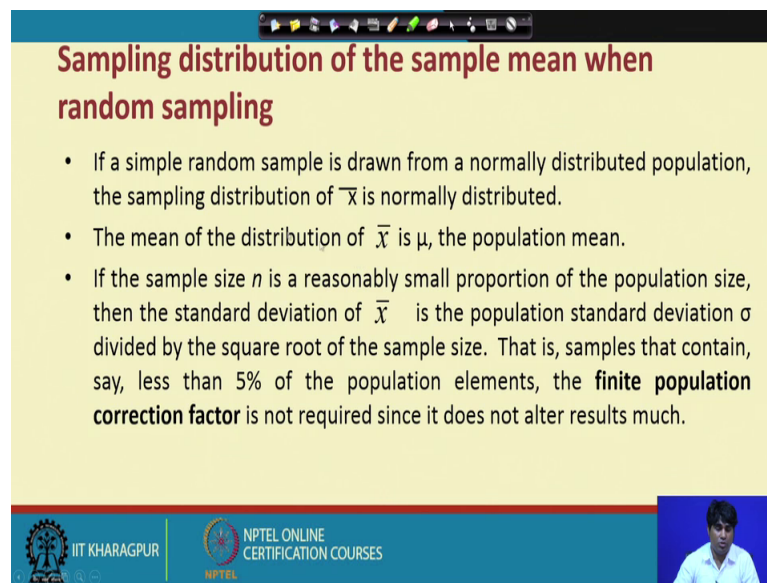
- When a sample is selected, the sampling method may allow the researcher to determine the sampling distribution of the sample mean  $\bar{x}$ . The researcher hopes that the mean of the sampling distribution will be  $\mu$ , the mean of the population. If this occurs, then the expected value of the statistic  $\bar{x}$  is  $\mu$ . This characteristic of the sample mean is that of being an unbiased estimator of  $\mu$ . In this case,  
$$E(\bar{x}) = \mu$$
- If the variance of the sampling distribution can be determined, then the researcher is able to determine how variable  $\bar{x}$  is when there are repeated samples. The researcher hopes to have a small variability for the sample means, so most estimates of  $\mu$  are close to  $\mu$ .

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So; that means, technically every time you may have a 3 types of analysis; sample specific, sample distribution specific, and then towards the population kind of requirement.

So, ultimately touching population is a very difficult, but doing the analysis at the sample and sampling distribution is a very easy. And, that is actually means that is the process, which you can called as best process of the estimation and inference.

(Refer Slide Time: 25:24)



**Sampling distribution of the sample mean when random sampling**

- If a simple random sample is drawn from a normally distributed population, the sampling distribution of  $\bar{x}$  is normally distributed.
- The mean of the distribution of  $\bar{x}$  is  $\mu$ , the population mean.
- If the sample size  $n$  is a reasonably small proportion of the population size, then the standard deviation of  $\bar{x}$  is the population standard deviation  $\sigma$  divided by the square root of the sample size. That is, samples that contain, say, less than 5% of the population elements, the **finite population correction factor** is not required since it does not alter results much.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

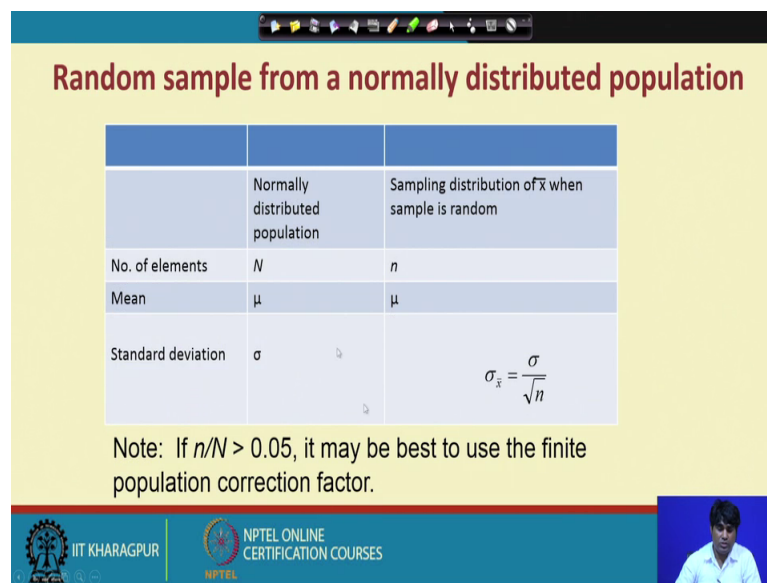
NPTEL

Video of a presenter in the bottom right corner.

So, sampling distribution of the sample mean when it is a kind structure called as random sampling. And, if not it is called as non random sampling, sometimes we are connecting with you know a sample to population. So, we have a finite samples and infinite samples, but most of the cases we really deal with a kind of finite sample case to analyze the particular problem.



(Refer Slide Time: 25:53)



### Random sample from a normally distributed population

	Normally distributed population	Sampling distribution of $\bar{x}$ when sample is random
No. of elements	$N$	$n$
Mean	$\mu$	$\mu$
Standard deviation	$\sigma$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

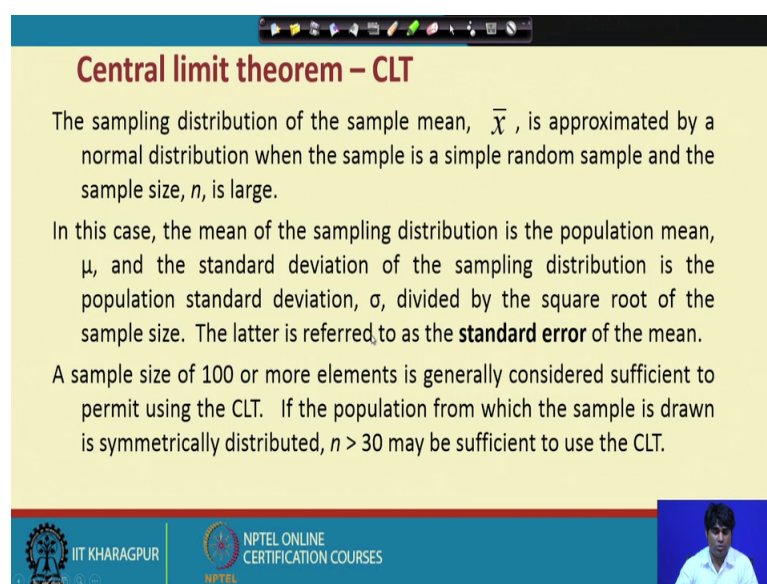
Note: If  $n/N > 0.05$ , it may be best to use the finite population correction factor.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, this is how kind of linking between sampling to populations. And in fact, when we have a kind of finite population and infinite populations and that; that means, technically there is a huge connection between sampling distribution to the population. So, on the basis of samples sample statistics or sampling distributions; so we can generalize to the populations.

So, that that is how the structure we have to follow?

(Refer Slide Time: 26:25)



### Central limit theorem – CLT

The sampling distribution of the sample mean,  $\bar{x}$ , is approximated by a normal distribution when the sample is a simple random sample and the sample size,  $n$ , is large.

In this case, the mean of the sampling distribution is the population mean,  $\mu$ , and the standard deviation of the sampling distribution is the population standard deviation,  $\sigma$ , divided by the square root of the sample size. The latter is referred to as the **standard error** of the mean.

A sample size of 100 or more elements is generally considered sufficient to permit using the CLT. If the population from which the sample is drawn is symmetrically distributed,  $n > 30$  may be sufficient to use the CLT.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, then in order to justify the best sampling or the kind of need of the normally distorted structures. So, we can connect with the central limit theorem. So, the structure of the sample you know central limit theorems that is what called as CLT is that, if you increase the sample size indefinitely, then the particular distribution will follow normal distribution. And for the clarity most of the distribution will converse to normal distributions, once you have you know once you increase the sample size you know indefinitely till you reach that particular a requirement.

So, central limit theorem will give you such kind of clarity and you know confirmations, whether the particular distribution is the normally distributed and what is the kind of in a requirement so, that the particular distribution would converse to normal distributions.

(Refer Slide Time: 27:22)

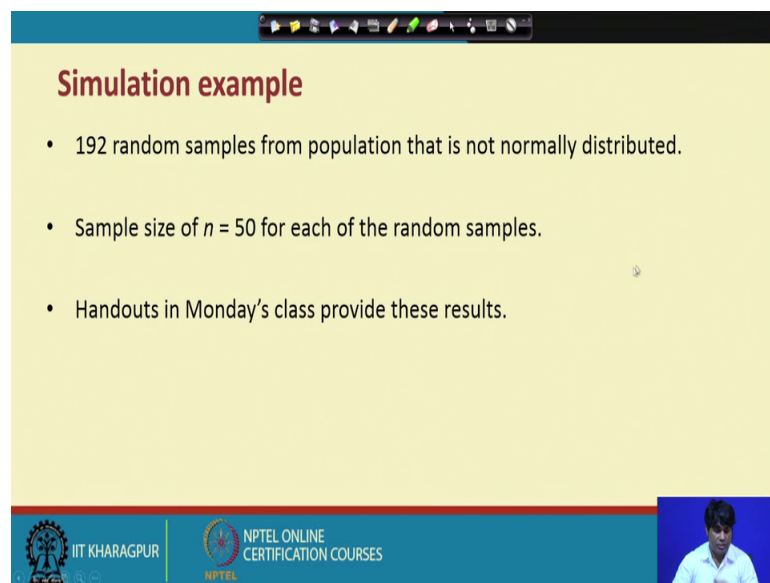
Large random sample from any population		
	Any population	Sampling distribution of $\bar{x}$ when sample is random
No. of elements	$N$	$n$
Mean	$\mu$	$\mu$
Standard deviation	$\sigma$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

A sample size  $n$  of greater than 100 is generally considered sufficiently large to use.

So, here again same things it is a question of populations to sampling distribution. So, in the case of sample the parameters are recognized as a small  $n$   $\mu$ , and in the case of population the parameters are capital  $N$ . And again  $\mu$  and in that case of population the some population variance is called as standard deviation  $\sigma$ . And, in the case of sampling distribution it is called a  $\sigma_{\bar{x}}$  which is nothing but actually  $\sigma$  by root  $n$  that is what the with respect to population? That means, from this formula also it is clear that there is a link between sample to sampling distribution and sampling distribution to the population.

So, theoretically if you just add up you know substantially and then nearly converse towards the populations and against with the help of sampling distributions you can actually comment about the population parameters. Sometimes some of the cases population parameters are known and then you analyze the issue and sometimes in you know population parameters are not known. So, with the help of sampling distribution statistics or sample statistic you can actually predict the population requirement or you know population structure.

(Refer Slide Time: 28:46)



**Simulation example**

- 192 random samples from population that is not normally distributed.
- Sample size of  $n = 50$  for each of the random samples.
- Handouts in Monday's class provide these results.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

And, this is how the kind of process?

(Refer Slide Time: 28:50)

### Sampling distribution in theory and practice

- Population mean  $\mu = 2352$  and standard deviation  $\sigma = 1485$ .
- Random sample of size  $n = 50$ .
- Sample mean,  $\bar{x}$ , is normally distributed with a mean of  $\mu = 2352$  and a standard deviation, or standard error, of

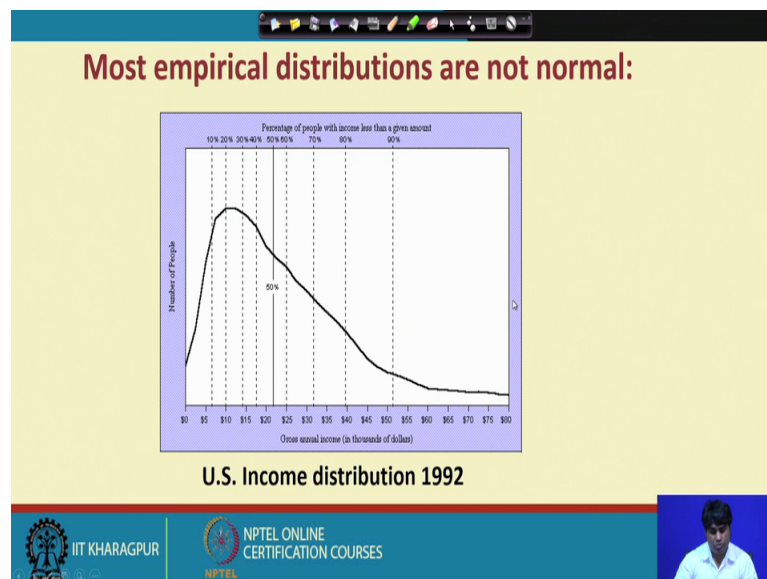
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1485}{\sqrt{50}} = \frac{1485}{7.071} = 210$$

In the simulation, the mean of the 192 random samples is 2337 and the standard deviation is 206.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

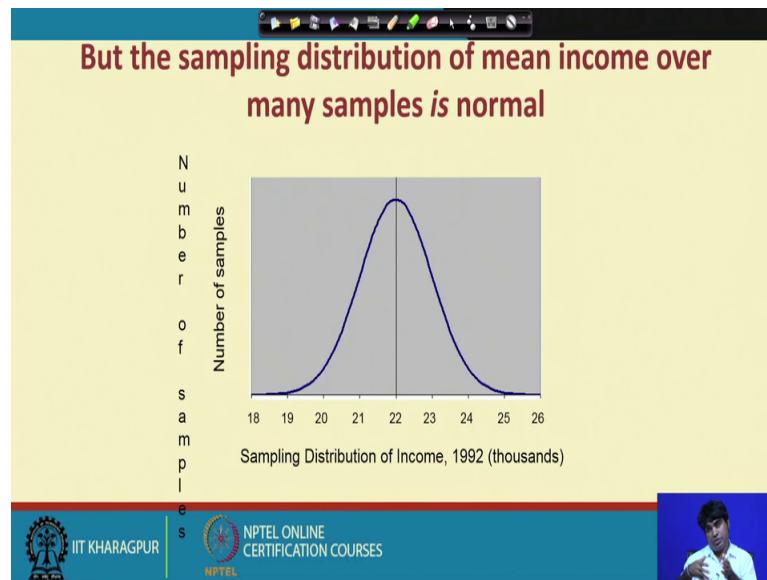
And means these are all different kind of look.

(Refer Slide Time: 28:53)



How actually extents you know a sample size and then check the pattern of the distributions.

(Refer Slide Time: 28:56)



So, if you take a case and in case the samples and then you know take the shape, then you will find I mean most of the instances. In case of sampling will definitely create a path for you know normal distributions this is what the cited examples?

(Refer Slide Time: 29:20)

**Random Sampling from Probability Distributions**

Example: Sampling from the Distribution of Dice Outcomes

Probability distribution			Intervals for random sampling	
$x$	$f(x)$	$F(x)$	Interval	Outcome
2	0.028	0.028	0 to 0.028	2
3	0.056	0.083	0.028 to 0.083	3
4	0.083	0.167	0.083 to 0.167	4
5	0.111	0.278	0.167 to 0.278	5
6	0.139	0.417	0.278 to 0.417	6
7	0.167	0.583	0.417 to 0.583	7
8	0.139	0.722	0.583 to 0.722	8
9	0.111	0.833	0.722 to 0.833	9
10	0.083	0.917	0.833 to 0.917	10
11	0.056	0.972	0.917 to 0.972	11
12	0.028	1.000	0.972 to 1	12

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES

And so, this is the classic examples here. So, these are all probability distributions and with respect to individual informations. And this is with respect to intervals, but in any case you can create a kind of distributions.

(Refer Slide Time: 29:39)

## Random Sampling from Probability Distributions

Example Sampling from the Distribution of Dice Outcomes

=RAND( ) generates random numbers in Excel

Sample	Random Number
1	0.681423018
2	0.835253743
3	0.438867243
4	0.11755569
5	0.731253287
6	0.58484908
7	0.450591681
8	0.119527366
9	0.778954333
10	0.833953932

Outcome = 8 since 0.681 is between 0.583 and 0.722

Outcome = 4 since 0.119 is between 0.083 and 0.167

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES - 145

As per the particular generally you know a requirement ok.

(Refer Slide Time: 29:44)

## Random Sampling from Probability Distributions

Example Using the VLOOKUP Function

- Generate a random sample of Changes in DJIA.
- First compute  $F(x)$
- Assign intervals to outcomes
- Generate random numbers using =RAND( )

=VLOOKUP(H2, \$E2:\$G\$10, 3)

Change in DJIA	$f(x)$	$F(x)$	Interval	Change in DJIA	Random Number	Outcome
-20%	0.01	0.01	0 0.01	-20%	0.681423018	>5%
-15%	0.05	0.06	0.01 0.06	-15%	0.835253743	10%
-10%	0.08	0.14	0.06 0.14	-10%	0.438867243	0%
-5%	0.15	0.29	0.14 0.29	-5%	0.11755569	-10%
0%	0.2	0.49	0.29 0.49	0%	0.731253287	5%
5%	0.25	0.74	0.49 0.74	5%	0.58484908	5%
10%	0.18	0.92	0.74 0.92	10%	0.450591681	0%
15%	0.06	0.98	0.92 0.98	15%	0.119527366	-10%
20%	0.02	1	0.98 1	20%	0.778954333	10%
					0.833953932	10%

IIT KHARAGPUR NPTEL ONLINE CERTIFICATION COURSES - 146

So, this is randomly generated a excel sheet you know Analyze a particular problems.

(Refer Slide Time: 29:48)

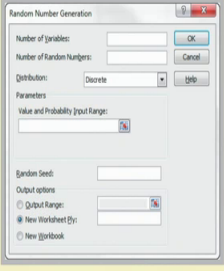
## Random Sampling from Probability Distributions

Example Using Excel's Random Number Generation Tool

Generate 100 outcomes from a Poisson distribution with a mean of 12.

*Data*  
*Data Analysis*  
*Random Number Generation*

Number of Variables: 1  
Number of Random Numbers: 100  
Distribution: Poisson  
Parameter: Lambda = 12



The dialog box shows the following settings: Number of Variables: 1, Number of Random Numbers: 100, Distribution: Poisson, Parameters: Value and Probability (input Ranges), Random Seed: (empty), Output options: ☒ New Worksheet By: (empty), ☐ New Workbook.

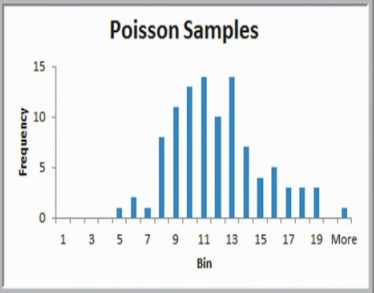
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 147

(Refer Slide Time: 29:50)

## Random Sampling from Probability Distributions

Example Using Excel's Random Number Generation Tool

Histogram of 100 random outcomes



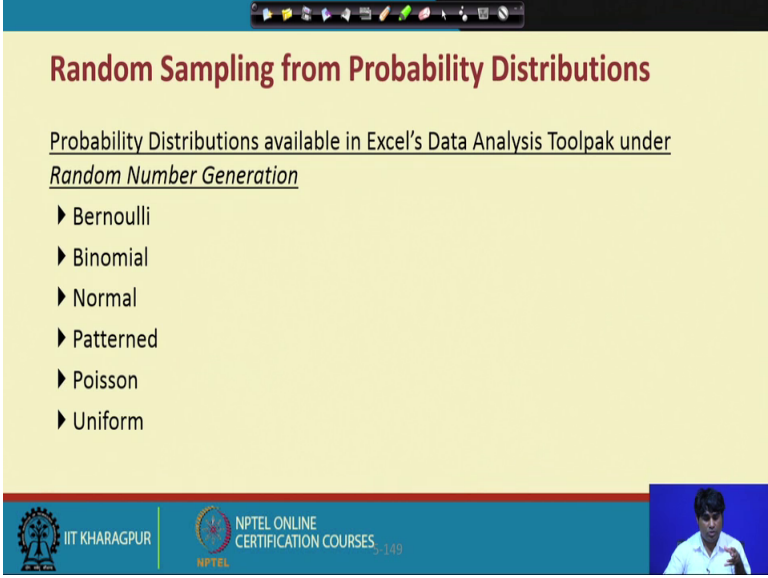
The histogram is titled "Poisson Samples". The x-axis is labeled "Bin" and ranges from 1 to 19, with a "More" option. The y-axis is labeled "Frequency" and ranges from 0 to 15. The distribution is unimodal and slightly right-skewed, peaking at bin 12 with a frequency of 14.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 148

And in fact, sampling distribution corresponding to probability distributions, we have the kind of binomial normal Poisson and uniform.



(Refer Slide Time: 29:53)



**Random Sampling from Probability Distributions**

Probability Distributions available in Excel's Data Analysis Toolpak under  
Random Number Generation

- ▶ Bernoulli
- ▶ Binomial
- ▶ Normal
- ▶ Patterned
- ▶ Poisson
- ▶ Uniform

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 149

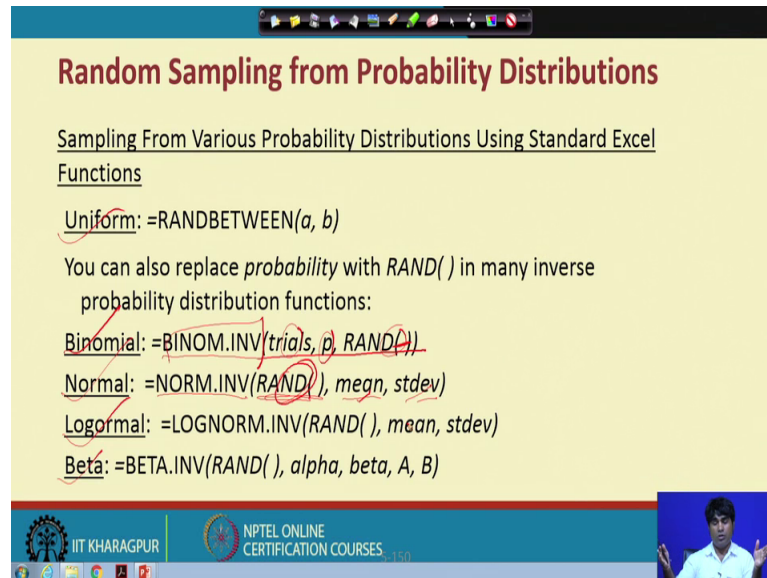
So, that means, technically here we need actually is samples or sampling to analyze the problem and some sometimes we use actually probability distribution to generate the sampling. So, if the problem is connected with you know let us say binomial. So, then you use the binomial distribution and then specify the parameters, then it will give you the kind of spreadsheet to analyze the engineering problem again, if the problem is connected to Poisson.

So, you use Poisson statistic for Poisson means use Poisson parameter see first you know peaks and then give the kind of random proxy to generate a kind of spreadsheet to analyze the particular, you know engineering problem. So, that means, here to create as samples or the kind of sampling. So, we have to take care or we have to you know connect with a particular probability distributions, because the generous of sampling definitely has a connection with a particular institution.

Since, we have a couple of distribution. So, definitely you know the particular problems will be connected to a particular distributions. So, here no such you know possibility where you know a particular problem will not connect to any distribution. So, every problems will be for you know will be converts or you know connected to either binomial or normals or Poisson or a kind of uniform distribution. Likewise, there are a couple of other distributions, but these are all standard this you know distribution

through which you can analyze the engineering problems by using some of the engineering econometrics tools.

(Refer Slide Time: 31:50)



**Random Sampling from Probability Distributions**

Sampling From Various Probability Distributions Using Standard Excel Functions

Uniform: =RANDBETWEEN(*a*, *b*)

You can also replace *probability* with *RAND()* in many inverse probability distribution functions:

Binomial: =BINOM.INV(*trials*, *p*, *RAND()*)

Normal: =NORM.INV(*RAND()*, *mean*, *stdev*)

Lognormal: =LOGNORM.INV(*RAND()*, *mean*, *stdev*)

Beta: =BETA.INV(*RAND()*, *alpha*, *beta*, *A*, *B*)

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES

So, if you go to the actually excel spreadsheet and these are all actually various commands and that to work out the particular distributions. So, in the last lectures in fact, we have discussed about the binomial you know probability structure, normally normal distribution structure, log normal beta distributions, uniform distributions, every distributions you have a population parameters. And, then regarding the sampling you know sample generation you have to specify the indications, then automatically it will give you a you know spreadsheet for the kind of empirical estimation process.

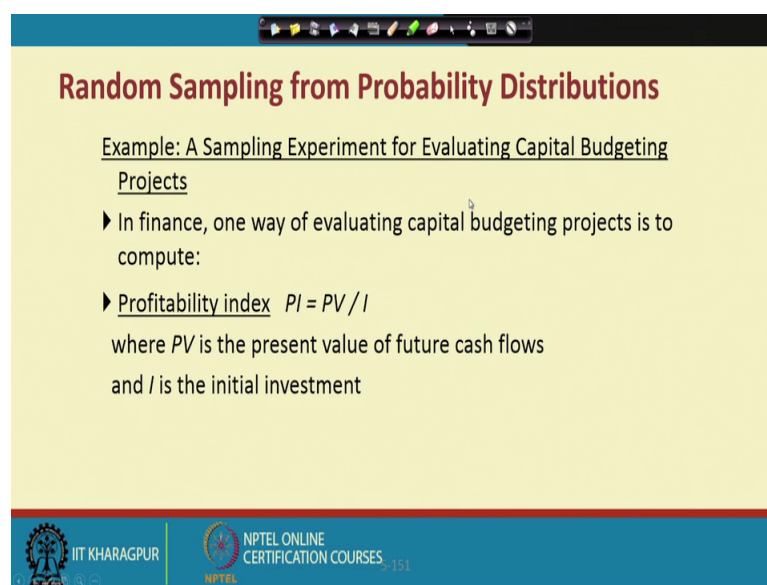
For instance, if you need to actually generate the binomial distributions and then the following command can be used. So, just you go to the excel spreadsheet and they put equal to signs and then by default by being for binomial distribution this command will be coming. So, you have to just feel of the trials, then the cap on particular probability and then random you know a randomized figures, then automatically, then particular distribution of decorated.

Similarly, for normal distribution so, this is what the command? So, a random structure and then mean standard deviation, because mean  $\mu$  and standard deviation  $\sigma$  is the population parameters for normal distribution. So, this is how the kind of variation will be you can make for instance, you know you may create your 1 sample case, 2 sample

case or multiple sample case. So, mean standard deviations will be constant and then you can change the randomize figures between 2 to 3 or you 2 to 2 5 4 you know something like that. You know on the basis of that you can have a more kind of spread and a different kind of sampling to analyze the engineering.

Similarly, for log nom log normal distributions and beta distribution. So, that means, technically you have to create a sampling on the basis of different probability distributions.

(Refer Slide Time: 034:05)



**Random Sampling from Probability Distributions**

Example: A Sampling Experiment for Evaluating Capital Budgeting Projects

- ▶ In finance, one way of evaluating capital budgeting projects is to compute:
- ▶ Profitability index  $PI = PV / I$   
where  $PV$  is the present value of future cash flows  
and  $I$  is the initial investment

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 151

So, these are all you know different kind of examples for instance a sampling experiment for evaluating capital budgeting projects that is a mostly corporate finance problems. And, one way of evaluating capital budgeting problem is to compute profitability index; that is that is the quality that is equal to present value divided by initial investment. And, then for that how you have to create a kind of, because capital budgeting is nothing but you know give you some kind of predestine structure about the future.

So, whether the project is the feasible one or not feasible ones or if it is a feasible one to what extend it up to what point of time? So, all depends upon you know the kind of probability and the kind of distribution.

(Refer Slide Time: 35:03)

### Random Sampling from Probability Distributions

Example (continued)

- For PV: =NORM.INV(RAND( ), 12, 2.5)
- For I: =NORM.INV(RAND( ), 3, 0.8)

$PI = PV/I$

Profitability  
Index mean  
= 4.76

	A	B	C	D	E
1	Profitability Index Analysis				
2					
3		Mean	Standard Deviation		
4	PV	12	2.5		
5	I	3	0.8		
6					
7	Experiment	PV	I	PI	Mean
8	1	8.396743042	3.573822001	2.349513601	4.762283
9	2	11.7446542	3.6654571	3.204067043	
10	3	11.76586852	3.554538257	3.310097619	
11	4	11.44456518	3.33708406	3.429510606	
12	5	9.373641185	3.692222659	2.538752955	
13	6	10.47906344	2.598868941	4.0321631	
14	7	14.31716958	3.203954788	4.46855289	
15	8	8.901052248	0.729081227	12.20858791	
16	9	13.99414343	3.180751244	4.399634852	
17	10	12.5758327	3.513579887	3.579207847	

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 153

So, likewise you can see here. So, this is the case and for that we have create a kind of normally distribution you know excel spreadsheet. So, here we have to specify the kind of parameters mu and standard deviation that is 12 and 2.5 and that is for present value. And, in the case of initial investment we put 3 and 0.8, then we are giving the randomized figures and by default will create it actually these are all various experiments. And, as a result we have a present value generated spreadsheet, then initial investment spreadsheet and then finally, the kind of profitability index.

(Refer Slide Time: 35:45)

### Random Sampling from Probability Distributions

Example (continued): A Sampling Experiment for Evaluating Capital Budgeting Projects

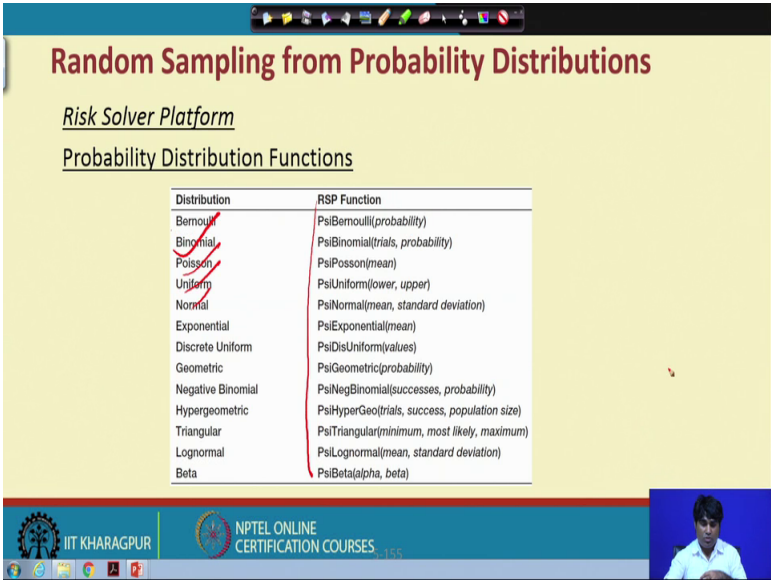
- Profitability Index is skewed to the right

Histogram of Simulated PI Values

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES 154

So, likewise you can actually work for various problems. And, if you plot all these things, then this will give you some kind of synopsis about the particular distribution, whether it is in normally distributed or it is a kind of skewed distributions, whether it is the right skewed or left skewed and what are the ways again the adjustment can have friends and that means, technically the right skewed and left skewed can be materialized. And then finally, we have a normally distributed figure to analyze the particular problem.

(Refer Slide Time: 36:17)



Distribution	RSP Function
Bernoulli	PsiBernoulli(probability)
Binomial	PsiBinomial(trials, probability)
Poisson	PsiPoisson(mean)
Uniform	PsiUniform(lower, upper)
Normal	PsiNormal(mean, standard deviation)
Exponential	PsiExponential(mean)
Discrete Uniform	PsiDisUniform(values)
Geometric	PsiGeometric(probability)
Negative Binomial	PsiNegBinomial(successes, probability)
Hypergeometric	PsiHyperGeo(trials, success, population size)
Triangular	PsiTriangular(minimum, most likely, maximum)
Lognormal	PsiLognormal(mean, standard deviation)
Beta	PsiBeta(alpha, beta)

So, these are all various kind of structure about the sampling and sampling distribution and that to connect with the some kind of probability distribution. In reality or in the statistical world or economic world the following distributions are you know very common and like starting with the Bernoulli distributions, Binomial distribution, Poisson uniform, Normal, Exponential, Discrete Uniform and Geometric, then Negative Binomial a Hypergeometric, Triangulars, Log normal and Beta. And, these are all actually excel commands and every distribution has a parameters and then and the choice of the random variable  $x$ .

So, you have to fix the parameters value first and then give a proxy to for the random variable  $x$  by default to recreate a sampling distribution or you know sample point. And, on the basis of that you can actually go for analyzing some of the engineering problems. So, that means, technically if you have you know actual data and with the help of a particular distributions, if you know that that particular problem is connected to a

particular distribution. And, with the help of some you know structure you know or some kind of feedbacks. So, you have to specify the parameters and then we can generate a sampling.

And finally, with the help of these samplings you can analyze this particular engineering problem.

(Refer Slide Time: 38:04)

### Random Sampling from Probability Distributions

Example: Using Risk Solver Platform Distribution Functions

- An energy company is considering offering a new product and needs to estimate the growth in PC ownership. The expected growth rates are:
- Minimum = 5%
- Most likely = 7.7%
- Maximum = 10%
- Generate 500 samples of PC ownership growth rate using:  
=PsiTriangular(5%, 7.7%, 10%)

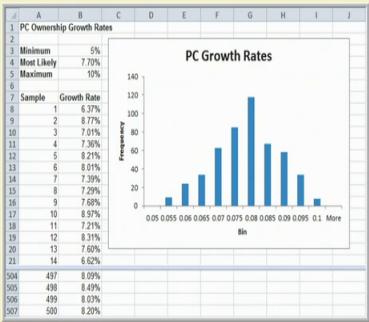
IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 156

So, that means, technically it is a very you know easy to do these things, because we have a different kind of scenarios.

(Refer Slide Time: 38:08)

### Random Sampling from Probability Distributions

Example(continued): Using Risk Solver Platform Distribution Functions



PC Growth Rates

Frequency

Bin

5.00 5.25 5.50 5.75 6.00 6.25 6.50 6.75 7.00 7.25 7.50 7.75 8.00 8.25 8.50 8.75 9.00 9.25 9.50 9.75 10.00

504 497 8.09%

505 498 8.49%

506 499 8.03%

507 500 8.20%

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 157

And your problem can be connected to a particular scenario and as a result you can have a solutions as per the particular, engineering problems require requirement.

(Refer Slide Time: 38:18)

## Data Modeling and Distribution Fitting

Example: Analyzing Airline Passenger Data

- Sample data on passenger demand for 25 flights

Bin	Frequency
36	0
35	0
47	0
45	0
48	1
43	0
42	0
56	2
40	6
47	3
44	5
46	3
53	2
45	1
44	0
45	0
45	0
41	0
47	0
46	0
48	0
42	0
41	0
58	0

Can we assume normally distributed?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 158

(Refer Slide Time: 38:19)

## Data Modeling and Distribution Fitting

Example: Analyzing Airport Service Times

- Sample data on service times for 812 passengers at an airport's ticketing counter

Bin	Frequency
227	126.2783
83	3.691221
10	86
158	83
350	105.1836
15	11063.59
63	8.707526
224	2.413577
96	867
61	9
61	876
91	102538
133	812

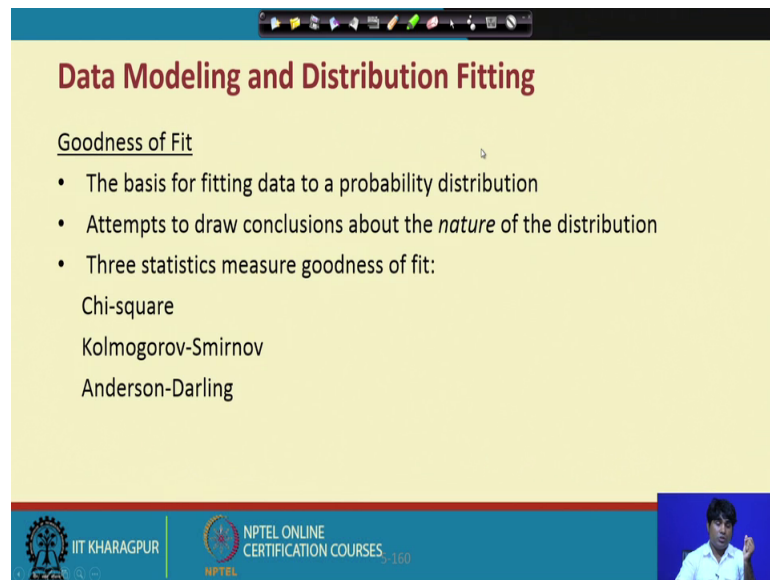
Can we assume normally distributed?

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 159

So, likewise different you know cases and then you have to check every times depending upon the particular requirement.



(Refer Slide Time: 38:31)



**Data Modeling and Distribution Fitting**

Goodness of Fit

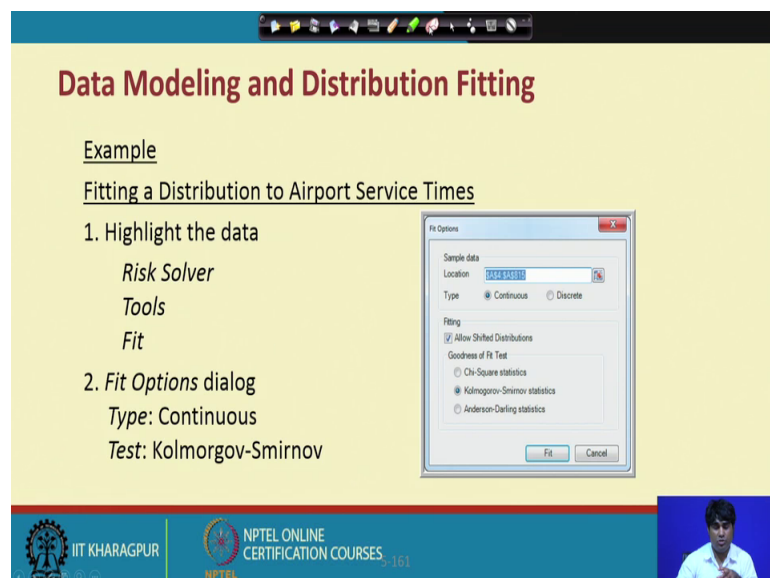
- The basis for fitting data to a probability distribution
- Attempts to draw conclusions about the *nature* of the distribution
- Three statistics measure goodness of fit:  
Chi-square  
Kolmogorov-Smirnov  
Anderson-Darling

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 160

And, ultimately a some of the issues are you know how to fix the sample points and the kind of sampling distribution, then comment about the populations the particular choice of the distribution like you know, Binomial Poisson uniform something like that. And, on the basis of that actually you have to be very accurate or very perfect about the sampling and that to for the particular engineering problems requirement.

So that means, technically. So, these are all actually you know different goodness fit to you know streamline the process of sampling and sampling distribution.

(Refer Slide Time: 39:12)



**Data Modeling and Distribution Fitting**

Example

Fitting a Distribution to Airport Service Times

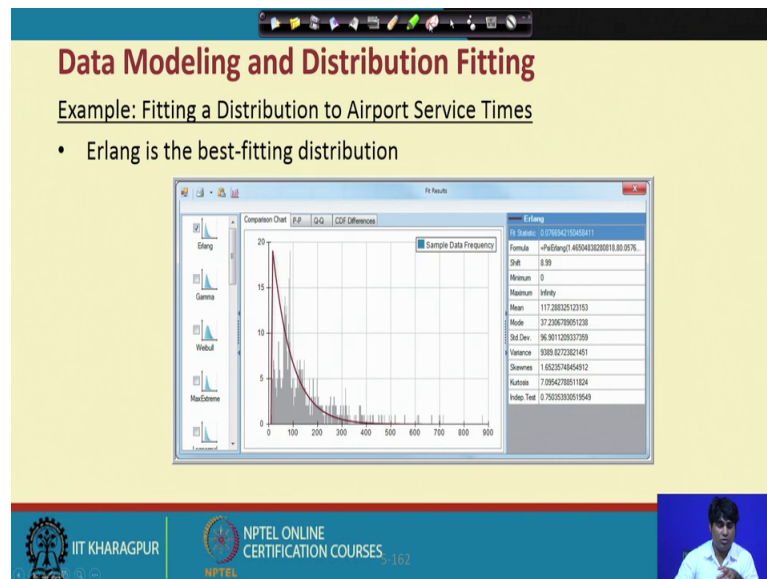
1. Highlight the data  
*Risk Solver*  
*Tools*  
*Fit*
2. *Fit Options* dialog  
*Type: Continuous*  
*Test: Kolmogorov-Smirnov*

Fit Options dialog box showing:  
Sample data: Location: 101.8 101.130  
Type: ☒ Continuous ☐ Discrete  
Fitting: ☒ Allow Shifted Distributions  
Goodness of Fit Test:  
☐ Chi-Square statistics  
☒ Kolmogorov-Smirnov statistics  
☐ Anderson-Darling statistics  
Buttons: Fit, Cancel

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 161

And these are all you know different modes again.

(Refer Slide Time: 39:14)



(Refer Slide Time: 39:16)

### Data Modeling and Distribution Fitting

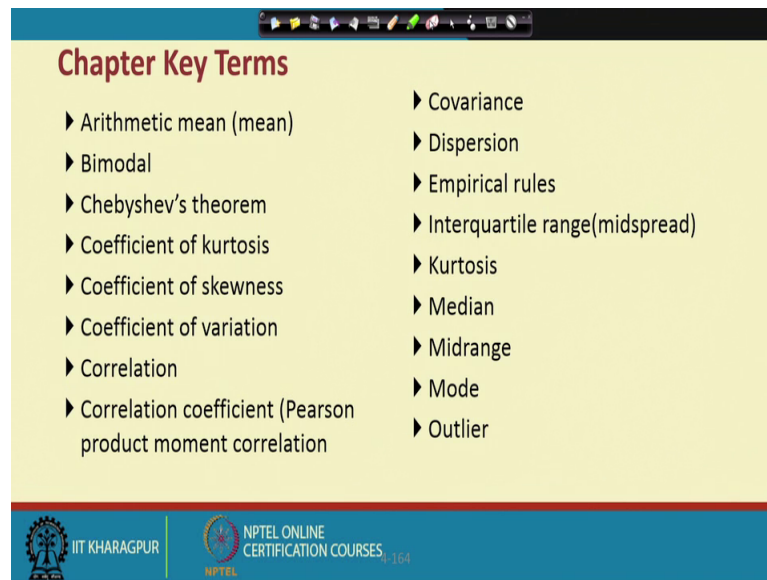
Analytics in Practice: The Value of Good Data Modeling in Advertising

- Gross's model:
- A mathematical model that relates the relative contributions of creative and media dollars to total advertising effectiveness.
- Often used to identify the best number of ads to purchase.
- Analysis found that the optimal number of ads can vary significantly depending on the shape of the distribution of effectiveness for a single ad.

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 163

So, ultimately.

(Refer Slide Time: 39:17)



**Chapter Key Terms**

- ▶ Arithmetic mean (mean)
- ▶ Bimodal
- ▶ Chebyshev's theorem
- ▶ Coefficient of kurtosis
- ▶ Coefficient of skewness
- ▶ Coefficient of variation
- ▶ Correlation
- ▶ Correlation coefficient (Pearson product moment correlation)
- ▶ Covariance
- ▶ Dispersion
- ▶ Empirical rules
- ▶ Interquartile range(midsread)
- ▶ Kurtosis
- ▶ Median
- ▶ Midrange
- ▶ Mode
- ▶ Outlier

IIT KHARAGPUR | NPTEL ONLINE CERTIFICATION COURSES | 154

So, in this particular unit we have gone through various you know items that means, these items are very essential for further processing or you know advanced requirement of using any kind of engineering econometrics and that to solve the engineering problems. Of course, a any hardcore modeling or any complex you know engineering problem solutions, we need to have a descriptive econometrics starting with you know the kind of measures of locations, then this persons say and the kind of sampling distribution probability probability distribution the kind of associations.

So, all these things are there and in between some of the inferential statistics or inferential econometrics you have to connect and then that to you know means the requirements of the validity of a particular models for analyzing any kind of engineering problems. In fact, we have now separate lectures for inferential econometrics; we will directly go to the predictive kind of structures where we use you know frequently the regression modeling. And, in the regression modeling 2 things are very essential in the process of analyzing any kind of engineering problems and that to by dealing with you know some of the engineering econometrics technique.

So, the first requirement is the descriptive statistic, the association statistics and then and the kind of inferential econometrics. So, descriptive econometrics, inferential econometrics are the basics for any kind of advanced engineering econometrics. Until unless you go through all these you know descriptive structure and inferential structure

you are not in a position to analyze any particular, engineering problems and to validate the results and come with a kind of solutions as per the particular requirement.

With this we will stop here. And in the next lectures we will start with regression modeling.

Thank you very much. Have a nice day.