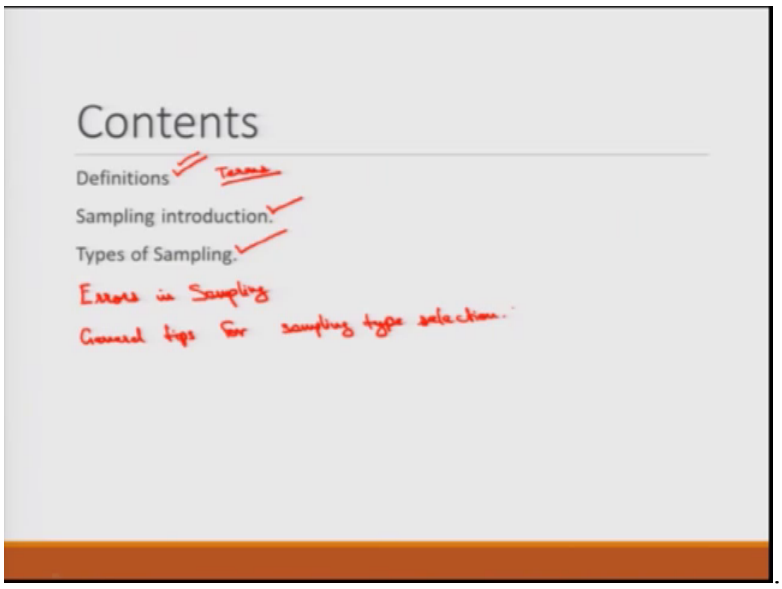
Practitioners Course in Descriptive, Predictive and Prescriptive Analytics

Prof. Deepu Philip
Dr. Amandeep Singh Oberoi
Mr. Sanjeev Newar
Department of Industrial & Management Engineering
National Institute of Technology, Jalandhar
Indian Institute of Technology, Kanpur

Lecture - 15
Sampling

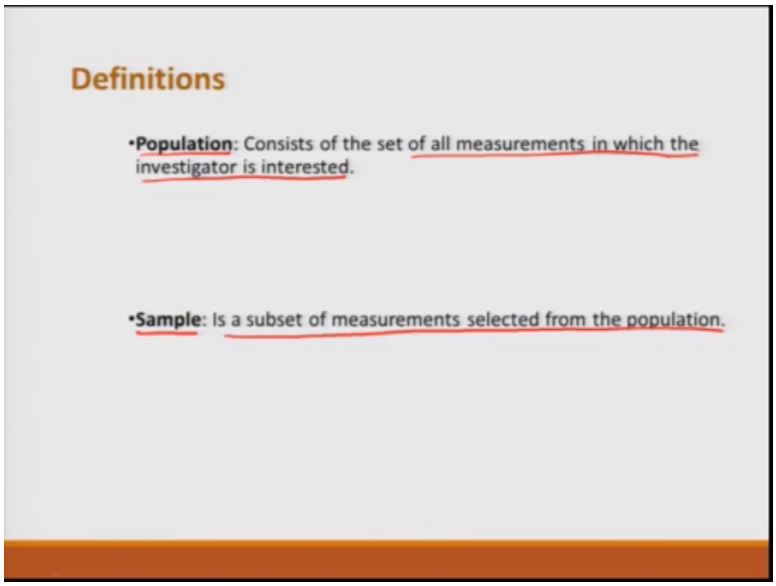
Good morning, welcome to the course on Analytics in which we are trying to study predictive, descriptive and prescriptive analytics. So, I am Doctor Amandeep Singh I will take this topic sampling here; So, better to call it sampling techniques.

(Refer Slide Time: 00:41)



So, the flow would go like this we will discuss about some definitions or terms which are used in sampling the introduction to sampling what is sampling then we will discuss various types of sampling; then also we will discuss the errors in samplings then I will try to give some general tips for sampling type selection.

(Refer Slide Time: 01:23)



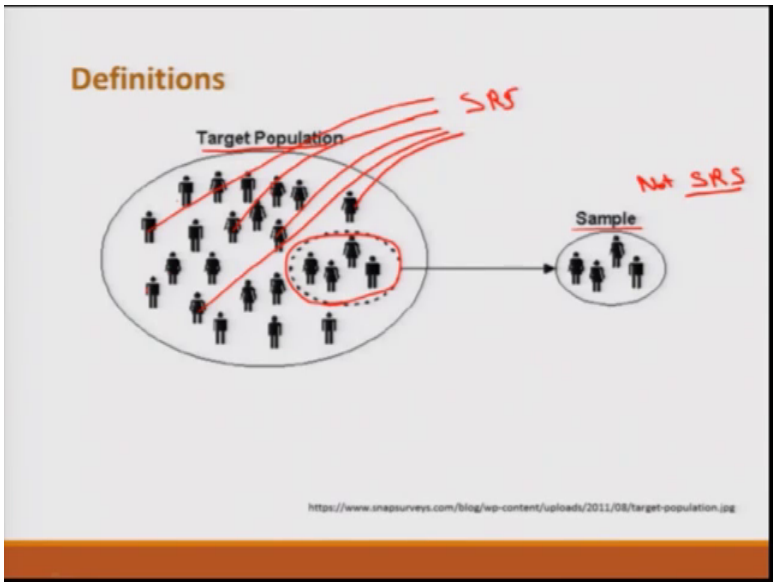
So, first a few definitions population and sampling population consists of this set of all measurement in which the investigator interested here. For instance if I am interested to study the number of students who are interested in doing engineering in the city Kanpur; so, my population is all the students in the city Kanpur who are in non medical stream that is my whole population.

Sample is a subset measurement selected from the population to do this study, I cannot reach all the students there would be about more than 1 lakh students in the year. So, I will select some number of students some students may be I will might divide it into demographics regions or I might divide into the kinds of schools they study the status of the schools. So, that is kind of sampling I am trying to select a small number from a big universe that big universe is population the small number is sample.

So, the population is generally denoted by capital N sample is denoted by small n a sampling from the population is often done randomly such that every possible sample of n elements will have an equal chance of being selected that is called simple random sampling a sample selected in this way in the simple random way is a simple random sample or just a random sample. So, example here may be if I want to test a student of a complete school having about 2000 students 2000 is my population and then if I test a smaller group of students from may be 30 to 50 students.

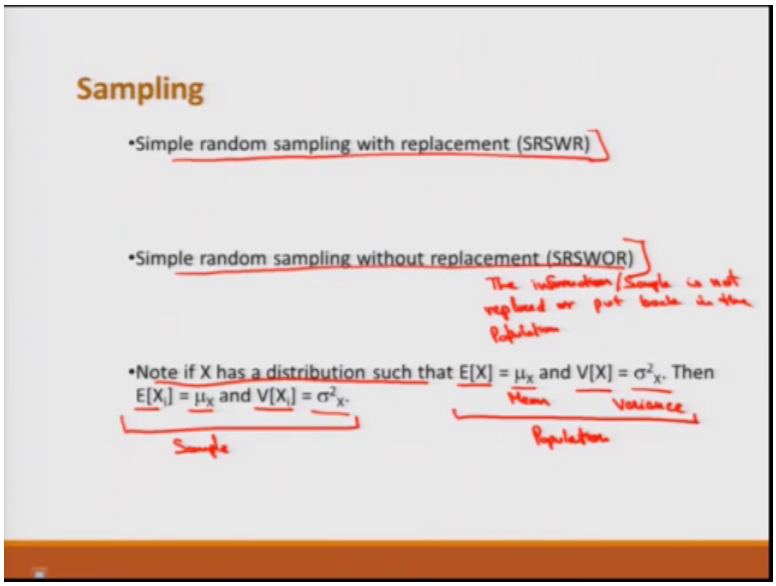
So, this is my small n small n by capital N is known as sampling ratio. Sometimes it is also mentioned in the percentage when it transept is in this case 50 by 2000 into 100 this is equal to 2.5 percent.

(Refer Slide Time: 04:15)



So, this is an illustration that explains we have a target population a sample is selected here. In this case this is not simple random sampling because if you see the students or the subjects here are distributed evenly it has just selected one cluster one group directly, but had is selected one from this side one from here one from here, one from here, one from here I might call it simple random sampling here we will see this things in the forth coming slides here.

(Refer Slide Time: 05:04)



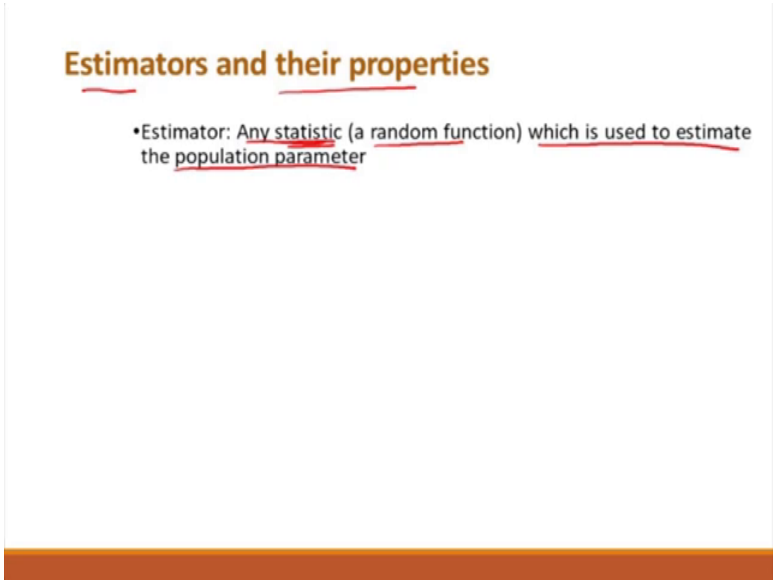
Then sampling can be simple random sampling with replacement simple random sampling without replacement. So, with replacement implies I am selecting out of the 2000 students which were out of my interest population is of my interest; I select 50 students and when I work on one sample that is sampling number from 1 from 50; 1, 2, 3 shown up to 50.

I work on this sample first and then I replace this sample in my population again out of 2000 I will select second number. Then after working on this I will put that in my population again I will select the third this is with replacement. Without replacement is if I have worked if I have taken the information from sample 1; this is taken out in without replacement type the information or sample is not replaced or put back in the population.

In that case if I have worked on the sample number 1 I have taken the information from sample number 1 I would not replace the sample in the population out of the remaining 2000 minus 1 samples 2000 minus 1 elements I will select sample number 2 this is call without replacement. So, this is a general representation here.

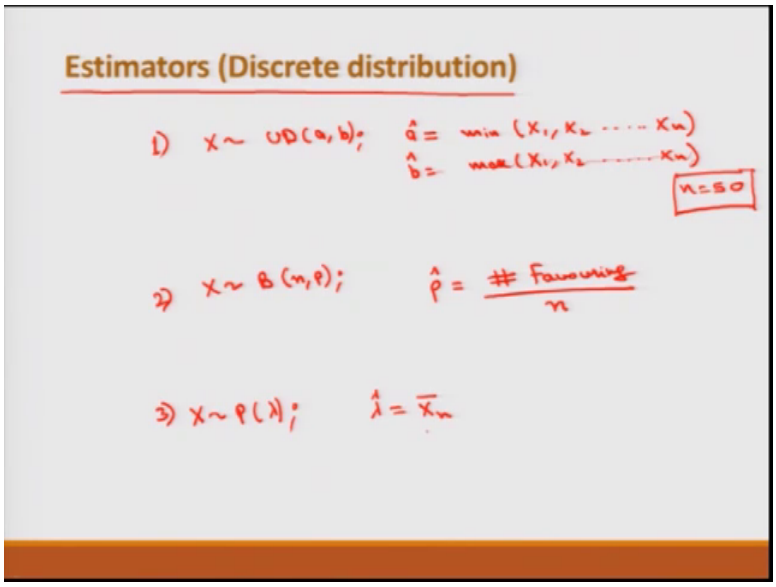
If X has a distribution such that expected value of X is mu X expected value is generally mean and variance of X is sigma square X this is variance, this is for population. Then expected value of X i would also be mu X and variance of X i would also be sigma square X that is the sample behavior would be same as that of the population in case of random sampling we are talking.

(Refer Slide Time: 07:50)



So, what are estimators and their properties estimator any statistics for a random function which is used to estimate the population parameter is known as a statistics; So, in this case this mu and sigma square r my statistics.

(Refer Slide Time: 08:19)



So, let us recall something about the probability distributions we had this estimators in discrete distributions. We had uniform distribution then binomial distribution in uniform distribution we have the smallest and largest value in binomial distribution have I have n

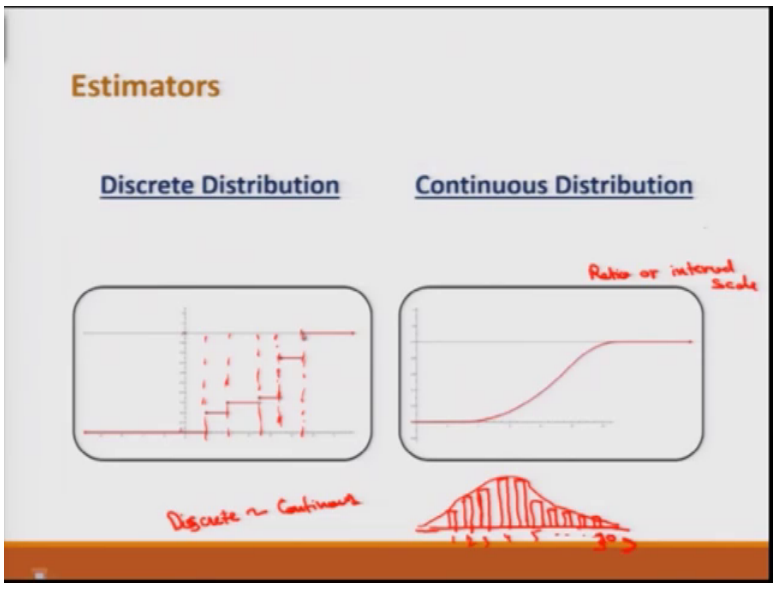
and probability of success and another discrete distribution is Poisson distribution in which I have lambda as arrival date.

So, in this case here a cap is equal to minimum value of X 1 X 2. So, on up to X n and b cap is maximum of X 1 X 2. So, on up to X n (Refer Time: 09:41) if I say marks of the students in the class are uniformly distribution. So, the minimum marks of the sample at is selected for example, if I talk about this 2000 students and if I am looking for the marks out of the of the 50 students which are of my interest sample size.

So, out of 50 students the minimum marks would be a cap here for the sample and the maximum marks would be b cap here. So, here is equal to 50 in this example in binomial distribution the p cap probability of success is the number that is favoring over total sample size.

So, Doctor Deepu Philip have already discussed the distributions probability distribution. So, I will just recall that and giving an overview here in this case this lambda cap is same as average of the n samples.

(Refer Slide Time: 11:03)

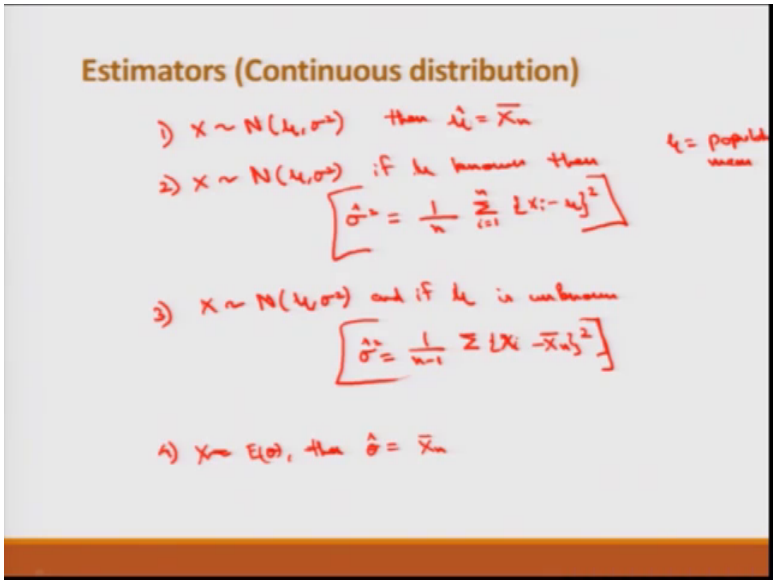


Then we have continuous distributions. So, before that I will like to show you this is discrete distribution this is continuous distribution discrete, this continuous distribution is like weight can very continuous it can take any numbers it is actually the ratio or interval scale ratio or interval scale.

So, this is discrete distribution it is not connected this one is continuous. And we will see that if the sample size is too high in case of I will when is see the plots is the sample size is too high the histogram plots are like this in normal distribution. if I am having very large sample size then this discrete distribution tends to become a continuous or behaves like a continuous distribution in this case.

So, this is histograms which are separate, but we can draw a line that joins these and which can call this as a normal continuous distribution provided that these numbers 1, 2, 3, 4 are more than 30 this is in general.

(Refer Slide Time: 12:42)



Continuous distribution we have again normal distribution in which we have two statistic mu and the sigma square; the mean and variance here mu cap that is for sample is equal to average of the sample. And if in normal distribution mu is known then variance is equal to 1 by n summation i varies from 1 to n X i minus mu square. If variance is unknown these degrees of freedom reduces the degree of freedom becomes 1 by n minus 1 and the other relation remains same.

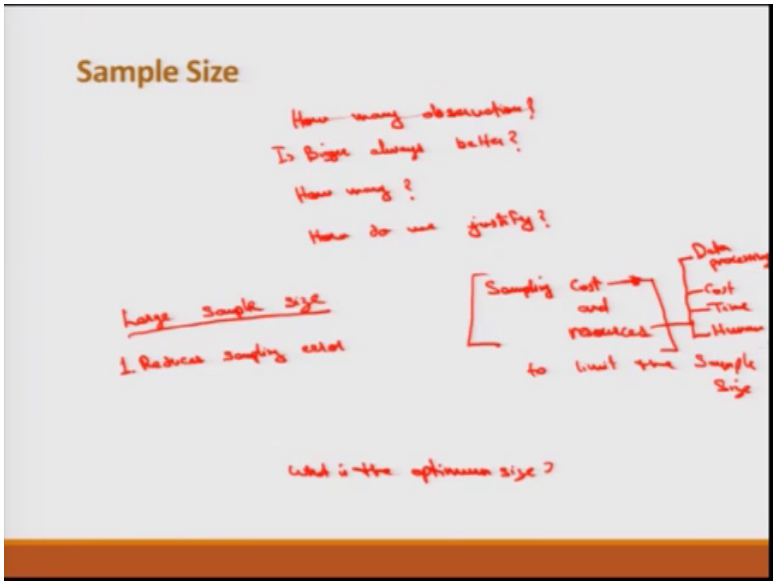
This is X i this is the third case when a normal distribution has mean mu variance sigma square and if mu is unknown. So, this would be more clarified when we will see when we will apply various kind of tests various kind of where we will do analysis of variance will apply T test will ap will do degration and similar kind of mechanism when we will learn. We will see that if the population mean is known the degree the freedom is n

minus 1 and the variance is calculated by this relation if populations means remember mu is equal to populations mean if populations mean is known then this is the relation.

Also another distribution is (Refer Time: 15:31) then theta cap is equal to average of sample. So, I need to mention here that to select the distribution it needs detail knowledge of how distribution behave. We actually see the data the data that is given for the sample or for the populations, we prod the data with try to fit the distributions here and whichever distribution is the closest fit.

We select that you see that this is the distribution fit the coefficient of determination coefficient determination is the closeness to the distribution that is the coefficient of determination is very high for this distribution and this thing is very close to the sample. So, that will practice when we go to the our section when will see the our course for this one.

(Refer Slide Time: 16:38)



Next is sample size. So, we can raise the following question while sampling that how many observations now second thing I can ask is bigger always better is bigger always better as we are testing the sample and we will see whatever the behavior of sample is would reflect the behavior of population.

And if sample is not selected properly not a correct sample is not selected those can lead to errors if from a lot from a populations from a good populations that sample is selected

for a from a group of intelligent students only weak students are selected as a sample in case and because of those weak students, the whole group of good students can be rejected.

And in the other way from a group of weak students if intelligent student or intelligent a good sample is selected from a weak populations that whole weak populations can be selected. So, when a good lot is rejected that is the producers error, when a bad lot is selected that is a consumer error (Refer Time: 18:02) seen a the manufacturing if we a manufacturing this pen out of the 1000 pens which are manufactured I select is sample of 50 pens.

And the whole lot was good, but these 50 pens does not did not behave well and I have rejected the whole lot who is at loss? The producer and the other case can be if the whole lot is not good and 50 pen which I have select from 1000 populations size is good that is behaved well the whole lot is selected; that means, that would be sold to the consumers then who is at the loss? Consumers.

So, this error of selection a bad o this error of selecting a bad lot is known as consumer error. We will see when we plot the operation characteristics curve that acceptance sampling. So, is big always better is this is the question are the bigger static studies are always better then how many households villages; if you are trying to do study a kind of a survey how many households villages blocks should I study when I do a research on. How do I justify that to the donors or head of institutions how do we justify because sampling has cost associated with it sampling cost.

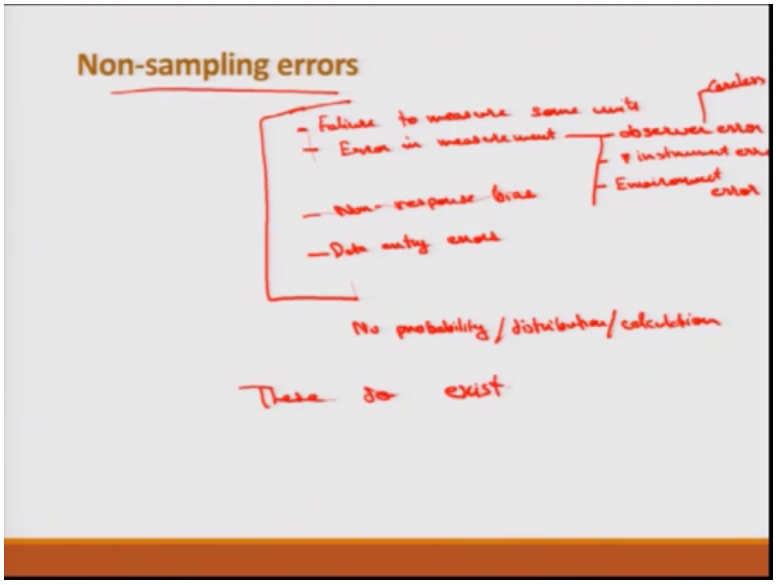
If it is a kind of destructive testing for example, if I want to test this strength of this pen depend body I need to apply some mechanical force here I need to break this thing; so, these 50 samples would be destructed. So, there is a cost associated with that all this 50 body or the pipes or the outer coverings would be wasted. So, sampling cost is associated and in surveys or in (Refer Time: 20:16) studies or in surveys also the cost is there. For each sample the researched as has to contact the respondent he has to give him a questionnaire, then get it back the time is consumed, the resources has con consumed and the cost is also there.

So, better to put here sampling cost and resources; so, these are the key to limit the sample size. So, increased or big sample size large sample size has only a few

advantages that I list here large sample size that reduces sampling error, but it has many disadvantages that is the non sampling error is also there; the cost of data collection is high that is money time resources are one is cost, I put it here again then time, then human resources then I have here is data processing.

So, the question comes what is the optimum size; what is the optimum size of this sample? So, let us try to elaborate this farther bigger is not necessarily better sampling errors are their this. So, what is optimum size to work on this I like to mention a few non sampling errors.

(Refer Slide Time: 22:44)

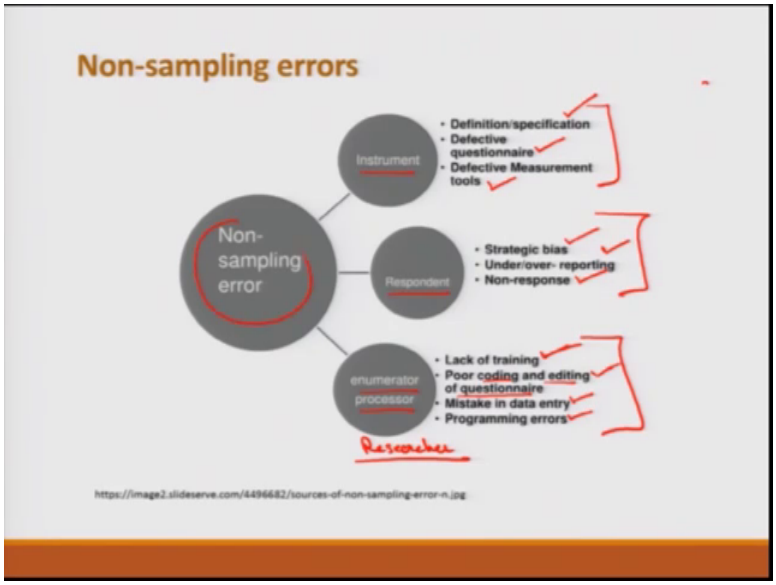


Non-sampling errors are the errors which do not have any probability which do not have any calculation involve in that, but those do exist. For instance the non amp sampling error may be the fill it to major sub units.

Then may be error in measurement this can be the observer error; the one who is doing this study or may be the instrument error some time the environment error as well. So, sometimes in surveys there is non-response bias people do not respond then sometimes the observer is may be careless or may be data entry error or sometimes miss representation of the facts and some other errors like telescoping a reference period etcetera.

So, these are non sampling errors which have no probability or distribution or calculation, but the thing is that these do exist.

(Refer Slide Time: 24:32)

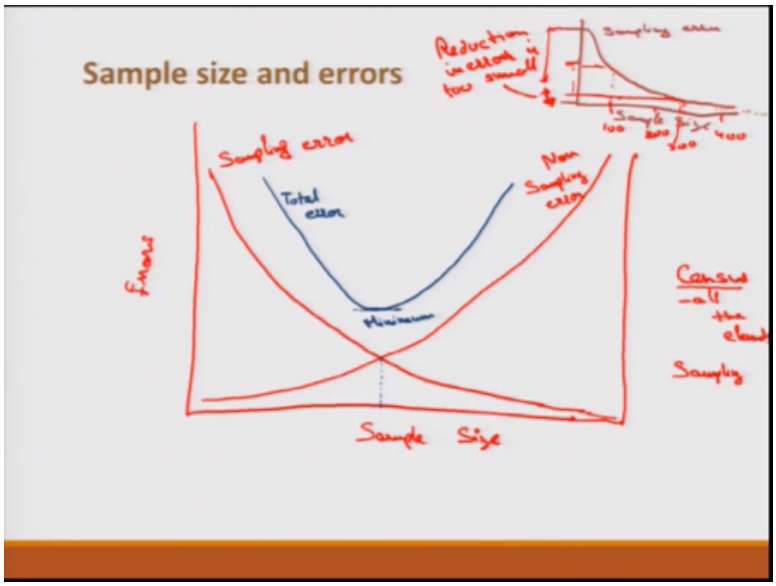


So, this is an illustration which is trying to explain the kinds of non sampling error. So, it can be due to the instruments, due to the respondent and due to the enumerator or processor so instrument the definition is specification of instrument can be wrong or defective questionnaire in case of survey defective measurement tools sometimes a tools measurement we some time use a questionnaire as a tool which is not easy for the respondent to understand.

So, sometimes the respondent has a strategic bias he under or over reporting also happens or non response occurs or something I think very casually about this studies and do not respond seriously. So, this are respondent errors the enumerator errors the lack of training of the that is the researcher is not very well confectioned with all the ways all the ups and downs those come during the practical survey.

So poor coding and editing of questionnaire getting the missing data, data cleaning all that does not happen very good mistake in data entry programming errors. So, these are non-sampling errors which comes into play.

(Refer Slide Time: 25:59)



So, because of these errors what should be the sample size? So, I like to put that in a diagram here we have errors here and we have sample size. So, as we increase the sample size the sampling error decreases this is a sampling error that is when the sample size increase increasing if I say this is my population; this is the total populations the sample size is increasing and it will be 0 at this case it will touch 0 at this case if the whole that is census study is done senses is all the populations all the elements I will put it is the complete populations census study second is sampling study.

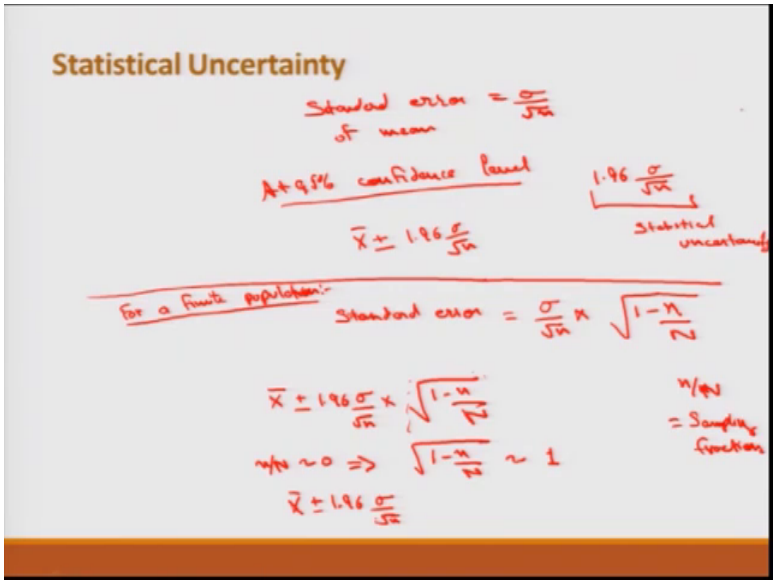
So, the non sampling error because of the high sample size; so, these errors could come in to play instrument respondent researcher it is increasing here this is non sampling error. So, what to do? The best strategy here could be select this point at this point this is actually the total error the blue line. Total error is the sum of sampling and non sampling error and this is the point where total error is minimum. I will also try to prove this statistically further that the sample size or the sampling error behaves like this. So, let me say if the populations is way far from this and this is the sample size, this is sampling error you see this errors war this error was this high at the beginning.

. So, in the later stage the error is reducing the error is reducing at this stage the error is this much. So, at this stage the error is already very less after this the gain this much of gain that is the reduction in error is too small is too small; even if I go here if I suppose this was if I divide put a scale here or let me put it 100, 200. So, if I put a scale here 100,

200, 300, 400 and so, on; if I increase my number of respondence from 300 to 400 the gain would be about 2 percent, but if I increase my number of respondence from 100 to 300; this gain might be about if I say this is 100 percent this is 100 percent this is this gain might be about 40 percent. So, it is not recommended to have very large sample size.

So, even if the populations is whole country or whole city lakhs of people are there a sample size between 300 and 400 is a good number to work on we will see how statistically this is true. So, next comes the statistical uncertainty.

(Refer Slide Time: 30:49)



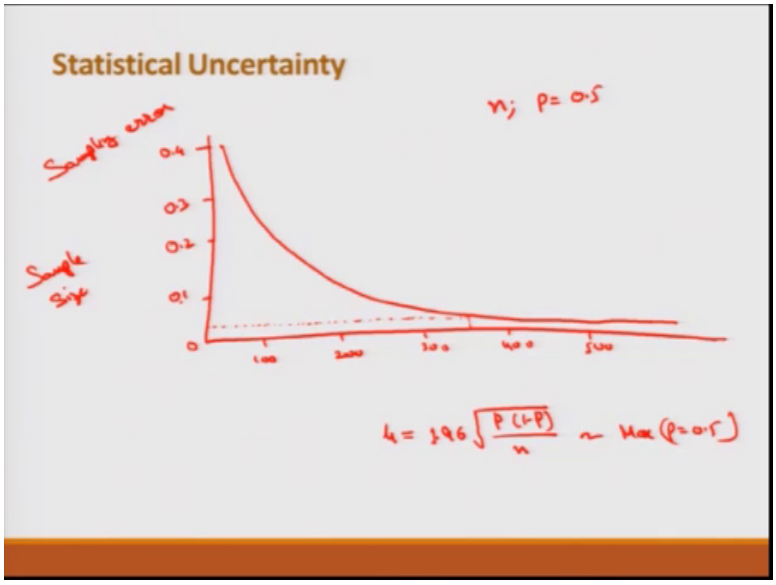
So, if this is standard error sigma by under root n is the standard error of mean. This is the deviation standard deviation of an average obtained by dividing the original standard deviation with square root of the number of data values of n. Then at let me say at 95 percent confidence level 95 percent confidence level, this tendered error is 1.96 here 1.96 is the normal deviate here 1.96 into this tendered error. So, we can say X bar plus minus 1.96 sigma by under root n is our 95 percent confidence level.

So, this value this value is my statistical uncertainty the standard error here is for a finite populations the standard error is sigma by under root n into 1 minus n by n here capital N is finite number and that is it the number of (Refer Time: 33:04) in the populations we know small n by big n this n by n as I said before this is sampling fraction.

So, the formula for the confidence interval can be modified in a way X bar plus minus 1.96 sigma by under root n into 1 minus n by n when the sample is very small that is n by n tends to 0; this implies the relation under square root that is 1 by n by small n by capital N tends to 1 then this relation X bar plus minus 1.96 sigma by under root 10 tends true.

We will see the practical significant of the statistical uncertainty when we will do the r code for that I will show the plots that how statistical uncertainty is important in determining the sample size and the kind of sampling strategy we should use.

(Refer Slide Time: 34:42)



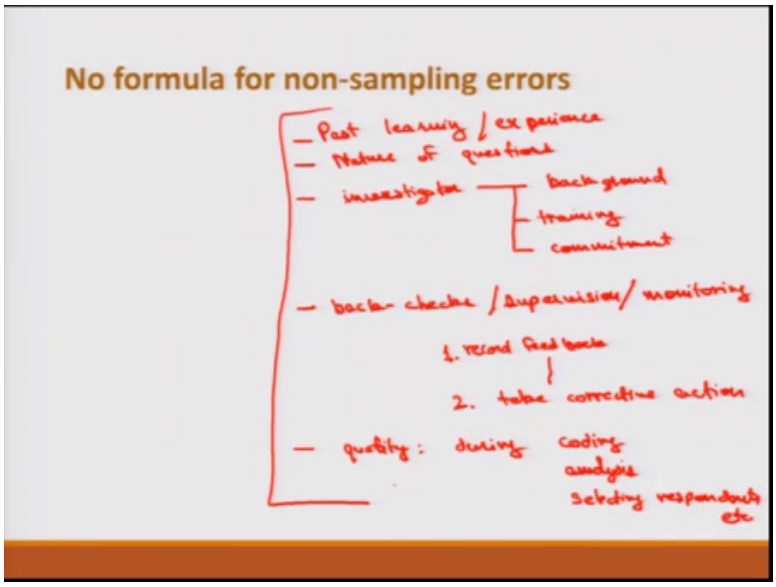
So, what is this statistical uncertainty I would like to put this illustration here again. This is actually the statistical uncertainty which we call as sampling error. So, if I say this is 0 maximum it could be 1 uncertainty if I say this is about 0.4; 1, 2, 3; 0.1, 0.2, 0.3.

So, if the sample size is 100, 200, 300 and 400 and so, on. So, this is statistical uncertainty; the general formula for statistical uncertainty that is denoted by u is 1.96 under root P into 1 minus P by n. So, this the P is probability of success; so, we can see this is maximum when P is equal to 0.5 and this is a chart for a sample size n and P is equal to 0.5 here it is seen that the size between 300 and 400 is a good number.

So, big sample size does not mean we will have very less error. So, the uncertainty is this much error about 350. So, this elaborate what I said that even for very big studies the

sample size of 300 to 400 is a good number. So, this was about the sampling error this is all about the sampling error; we are actually working on this sample size sample size what should be the sample size we working on this thing.

(Refer Slide Time: 37:21)



So, for non sampling error there is no formula available the decisions you reduce them are based on the past learning or experience. So, that depends upon the nature of questions; then for on this the researches or the investigator; his background that is the amount of knowledge he has about the study, then his training then the commitments that he makes. Also because there would be always some ups and downs, there would be always something different in the actual performance than this casual. So, back checks are to be made back checks or I could call it super vision or monitoring this study is important.

So, what do you do here? Take feed back or you put record feedback and take corrective action. Then during coding or during analysis of data the quality is to be maintained; quality during coding analysis selecting respondents etcetera. So, no wonder there is very little discussion about non sampling errors in the sampling theory; in the in our platform there would be no discussion about the non sampling errors. But these contribute these have a big role on the quality of the study on the kinds of the results we obtained.

So, in this lecture discussed about the terminology in the sampling, the populations samples, sample size, discrete and continuous distribution, then samples size depends

upon this sampling error, non sampling error. How are this interrelated? All these thing we have seen; in the second part of this lecture I like to discuss the types of sampling the sampling strategies.

So, I will take a break here; so, let us meet in the next lecture.

Thank you.