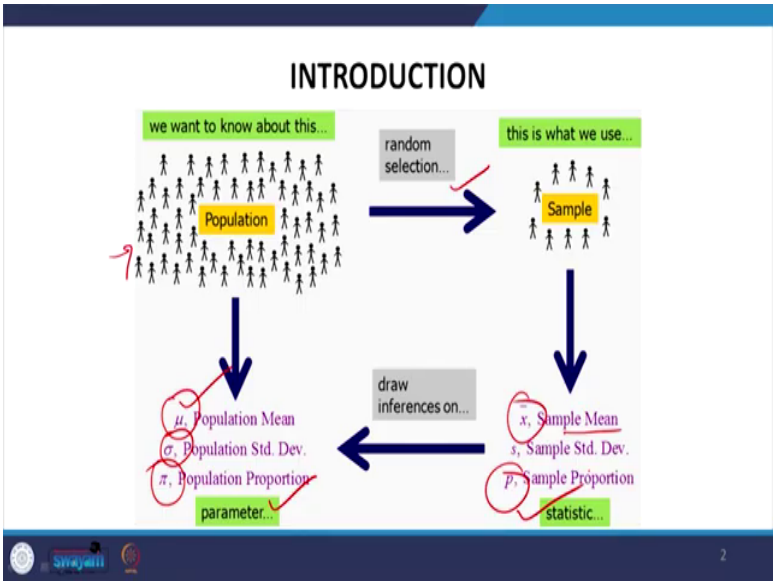
Handling Large-Scale Unit Level Data Using STATA
Professor Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee
Lecture 08
Sample Size

So, welcome dear students once again, to the NPTEL module on Handling Large-Scale Unit Level Data with STATA. So, STATA is the software, statistical software, we are going to use it from the next week onwards and we know some background will also be given on it, but in the week number 2, lecture number 8, we are discussing sample size. In the last lecture, I was discussing the distribution especially on sample, survey types, sample type probability sampling, and non-probability sampling.

And where sample size really matters, even if we follow multi-stage sampling technique, we do require the correct sample size. In both the cases, though the non-nonprobability sampling design requires little less the size because it is not going to generalize the population, it is very quantitative, it is quite pointed, but still, that requires sample size, in both the case sample size really matters. So, let us have a look at the sample design.

(Refer Slide Time: 01:46)



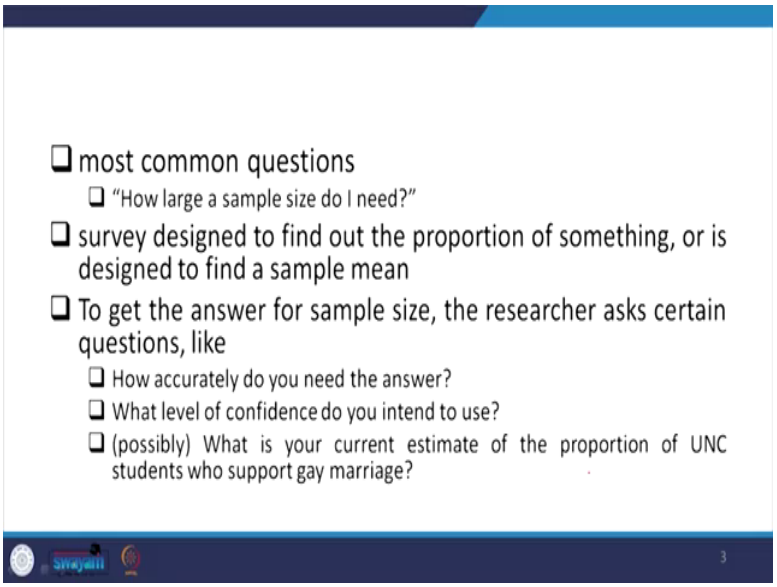
A pictorial format of sample and population where we are going to try. It may be the case that here in this particular picture, we have information about population here, we are saying some random samples are taken, if you are adopting a probability structure, subject to the case that a population is homogeneously distributed and if it is not homogenous, then it gets little complicated.

It matters when homogeneous, if it is not homogeneous and what kind of features are there which guarantees non-homogeneous and maybe the case that the data contains more distributional issues, it may contain more standard deviation, different standard deviation from one group to another group within the subsample of that entire population. If the variations are there, then the entire population is not homogeneous. So, the sample is designed accordingly or the sample frame is designed accordingly.

And here we are saying, first of all we derive certain sample and then we take their details, the population means that is generally noted as μ(mu), then population standard deviation as I said generally noted as sigma(σ) and population proportion if its information there we generally represent it with Pi, whereas, if the same representative parameters are taken, representative discussions are made in the sample, we say X bar for sample mean whereas, mu for population mean, small s for sample standard deviation whereas, in case of population it is sigma and in case of sample proportion it is P. Whereas, in case of population proportion, it is in fact Pi.

So, these indicators in case of population we say parameter whereas, in case of sample we say statistics. So, there are two ways of distinguishing these indicators. So, in case of sample these indicators are called a statistic. And in case of population, these are called parameters. So, I was just discussing from the beginning that population's features are explained to mu standard deviation or proportion. And accordingly, we derived sample, we already discussed sample type in our previous lecture.

(Refer Slide Time: 04:35)



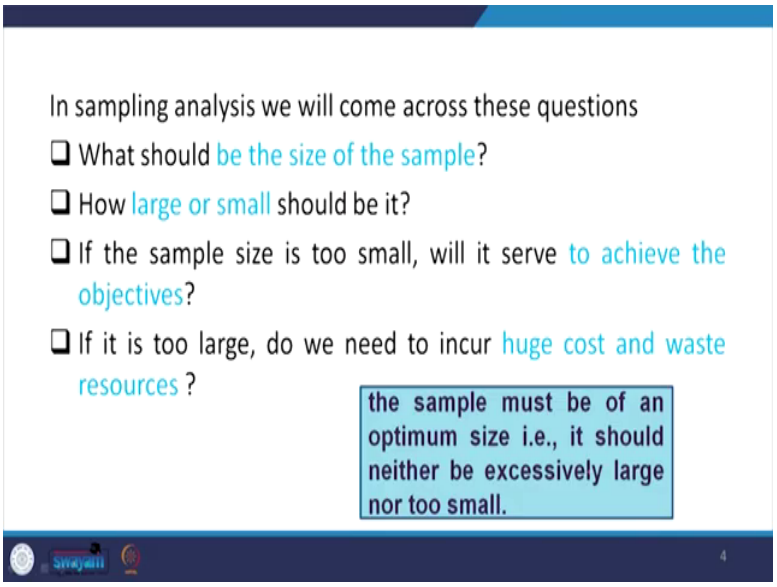
Let us proceed to understand the common questions we generally deal while discussing on the issues of sample size. So, first and most important question strikes to the researcher is how large a sample size do I need? So, a survey is designed to find a sample mean is very important and so far as the size is concerned. To get the answer for the sample size, the researcher asks certain questions like, how accurately do you need the answer? Regarding, how big the sample size I need. So, how accurately this is defined?

And what level of confidence do you intend to use? Suppose, I have collected some sample size whether it is confident. Well, whether it is meeting the confidence interval and confidence indicates the chances of error. If the confidence level is 95 or 90 percent or 99 percent. Usually, these three benchmarks are given and accordingly confidence tables are there or errors are there, reverse the confidence is nothing but errors. So, errors are indicated by alpha in our notation.

So, we are not discussing that in detail, but confidence is usually bracketed in three that is 90 percent, 95 percent, and 99 percent. Accordingly, errors are 1 percent 5 percent, or 10 percent. But that does not mean it is these are the three possibilities of error there are N number of possibilities of error, even we can define 3 percent of error, we can define 2 percent of error, 9 percent error, any number is possible.

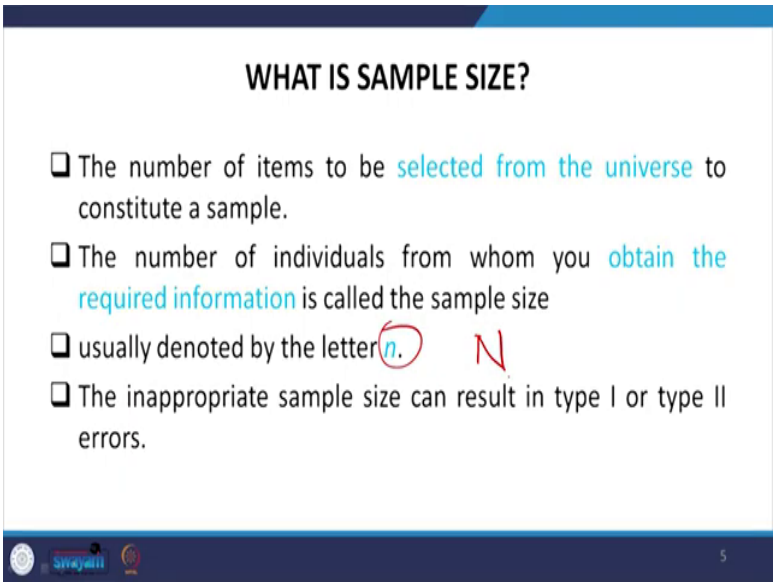
Usually, more than 10 percent error in any framework is not good, because if you do an experiment or you do any kind of testing, your error is more than 10 percent. So, it may not meet the requirement, may not solve the purpose, the results might be erroneous. So, the third indicator of understanding the sample size is, what is your current estimate who support the proportion of students in our case here, we are taking the example of UNC students who support gay marriage.

(Refer Slide Time: 07:09)



We have an example and we will discuss that in a short while with that example, in our slide that was purposely important. So, in a sampling analysis, we will come across the following questions, they are what should be the size of the sample, how large or small it should be if the sample size is too small, will it serve to achieve the objectives, or else if it is too large, then does it incur unnecessary cost or waste our resources and the sample size or the sample must be an optimum size. So, optimum should meet those indicators we have started discussing, I will discuss correctly whether that is meeting those indicators or not in our next slide.

(Refer Slide Time: 07:59)



But we are discussing the number of items to be selected from the universe is called sample and and the number of individuals from whom we obtain the required information is called a sample size. Usually, sample size denoted with small n, whereas, the population size is capital N. So, these are the indicator we will be using a number of times.

So, if our sample size results differently or inappropriately to the population and our results might be misleading and that can be presented with type one error and type two error. Sometimes we indicate with alpha and beta notation for type one and type two error respectively. For example, you have selected a sample but your results give wrong information. So, rejecting a null hypothesis when it is true is called a type one error. In reverse accepting a null hypothesis when it false is called type two error.

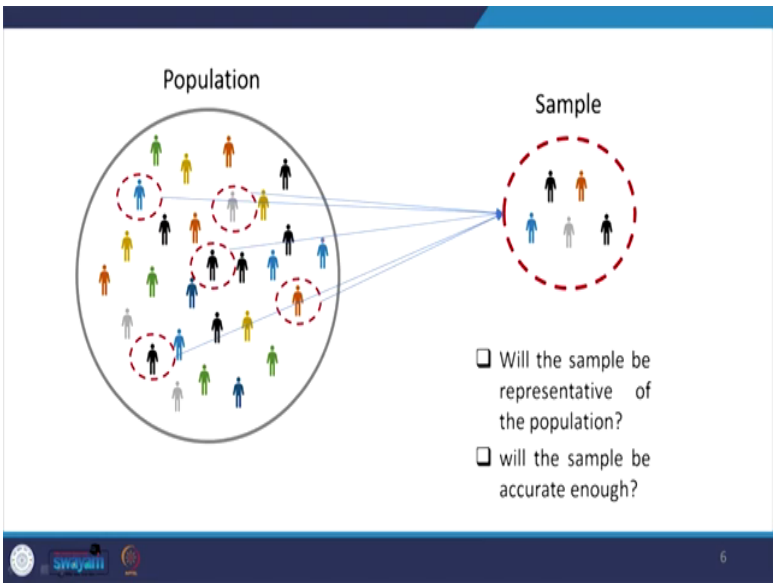
You have taken certain sample. From a sample you tested against the population, your population suggests that something is right, but your sample said it is not right. So, rejecting based your sample.

For example, in a consignment some packets are being transported from one place to another place, the excise officer just checked in between what exact content you are taking or even in a simple example like in a manufacturing unit for cold drinks beverages the content different contents are there, like chemical components are there, some percentages are allowed.

If they just simply taste whether it is harmful, whether your chemical composition is going to be harmful for body or not. And some of the samples are collected if those samples are erroneous. Unfortunately, those samples are erroneous or are picked up based on that what you said your entire package is bad. And these are not suggested for further manufacturing. So, based on a sample, it is simply rejected. So, it may be the case that these 3 4 indicators are bad or others are true.

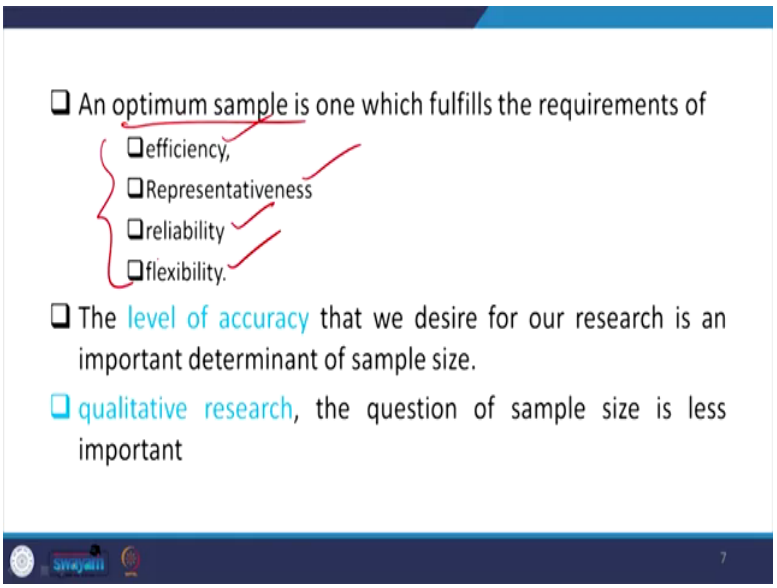
But your entire manufacturing, entire basket was true, but you have rejected the basket based on a sample. So, rejecting the null hypothesis when it is true is called type one error. Similarly, you can follow type two error, we will discuss many times in our next couple of modules a couple of slides in our next maybe next to next schedule.

(Refer Slide Time: 11:30)



Then what is here discussed regarding population and sample. Here, we say that the sample will be representative of the population, will the sample be accurate enough whatever you selected from this picture, like, some portion are selected these are based on our sampling design we already discussed some right sampling techniques may be applied to define whether these are representative or these are accurate enough. In order to discuss that sample size really matters, without sample size, we may not even discuss about whether these are accurate or not. So, these are our approaches to decide which sample size you should select.

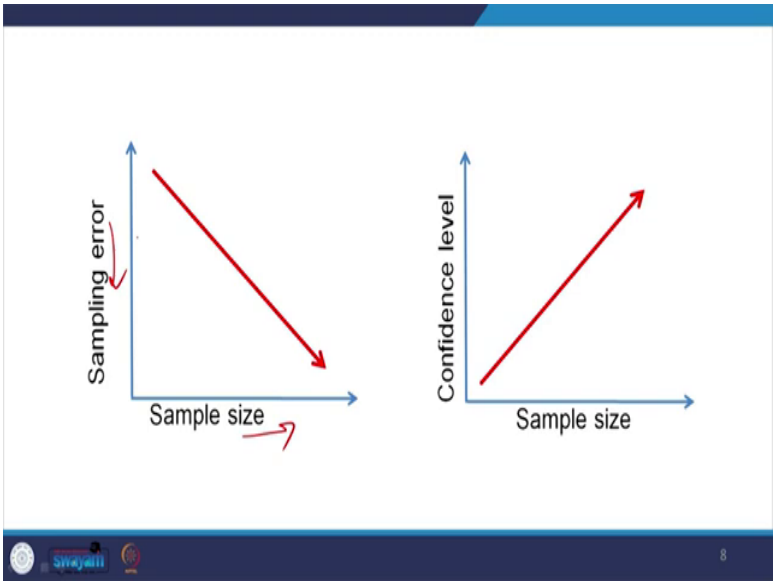
(Refer Slide Time: 11:30)



As I discussed already, what do you mean by optimum sample, what features is most fulfilled? There must be efficiency, representativeness, reliability, and flexibility. These four indicators are fulfilled. We have a dedicated lecture on this to understand the reliability of the

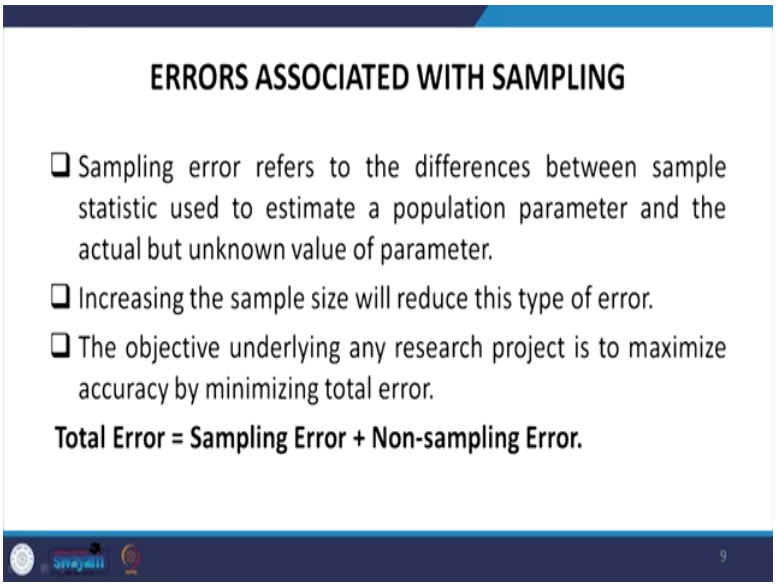
data and also will understand the representativeness of the data. But the level of accuracy that we desire for our research is an important determinant of sample size. In case of qualitative results, the sample size is less important.

(Refer Slide Time: 12:54)



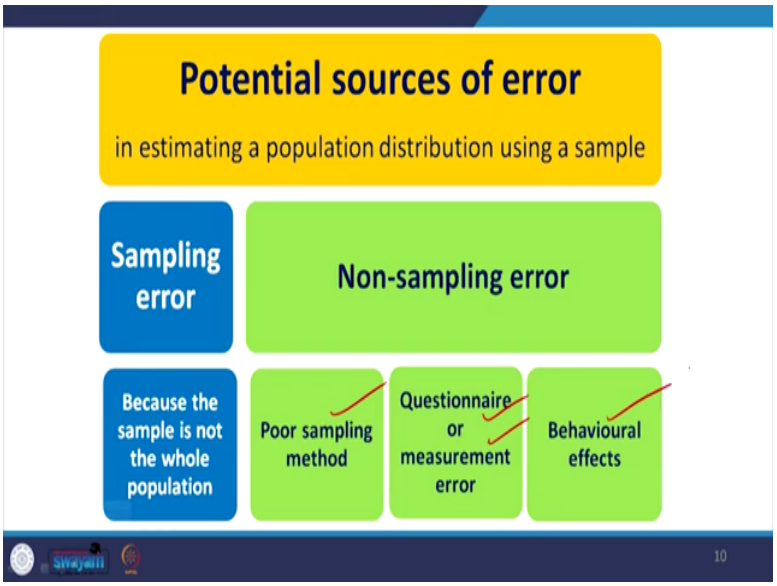
It is important to note that your qualitative variables must be different and must be complete. So, in this slide we are discussing when your sample size increases, error in the sampling format or design is going to be lesser. So, it is not negatively related. Similarly, when the sample size increases, your confidence level is expected to be higher. Today, we will discuss with further examples, when regarding sample size in this particular lecture.

(Refer Slide Time: 13:33)



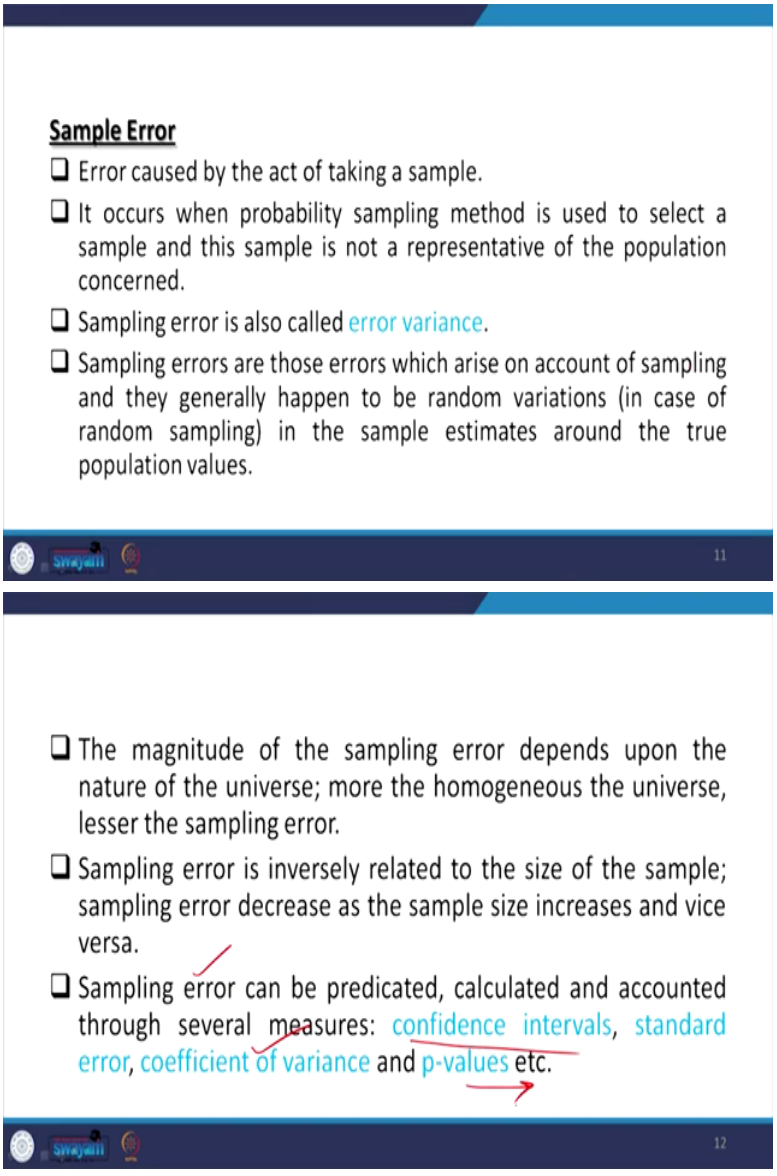
What about errors associated with sampling? What do you mean by sampling error? Sampling error refers to the difference between sample statistics used to estimate a population parameter and the actual but unknown value of parameter, as I already discussed. So, sample error measures the gap between sample statistics and population parameters. Increasing the sample size will reduce this type of error, we have already discussed. The objective underlying any research project is to maximize the accuracy by minimizing the total error, when the error is lesser, we can able to maximize our accuracy. So, the total sample error is derived by adding sampling error and non-sampling error.

(Refer Slide Time: 14:26)



So, here is the potential source of errors given. Usually, there are some reasons behind sampling error and non-sampling error, sampling error has resulted because the sample is not representing the whole population whereas, a non-sampling error sometimes due to poor sampling methods, questionnaire or measurement error, or behavioural aspects. A certain behavioural aspects do not give the right indicators to the server.

(Refer Slide Time: 14:26)



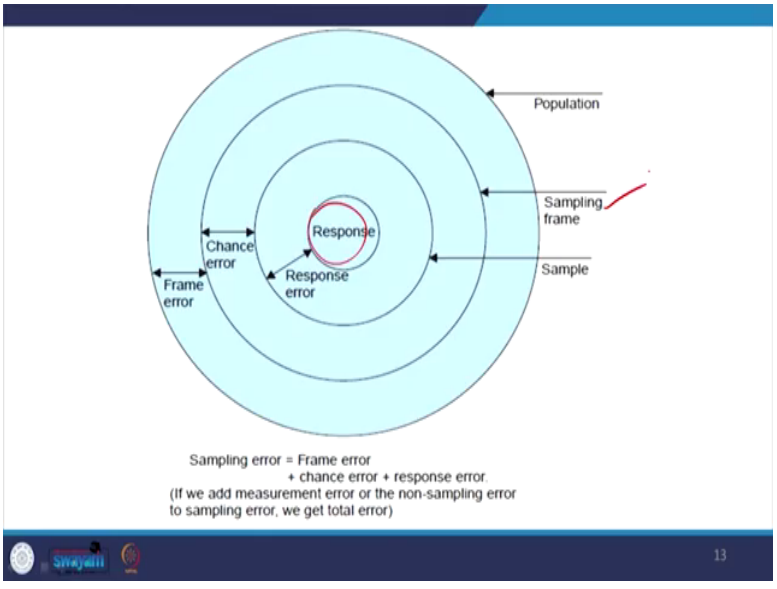
So, in a nutshell sample error caused by the act of taking the sample, it occurs when probability sampling method is used to select a sample and this sample is not representative of the population concerned. We have followed a probability method, but it is not represented correctly to the population then sampling error occurs, sampling error is also called error variance because there is much variance from one sample to another sample. If you are taking

one sample from the same population, you have taken once again another set of samples from the same population, another set of samples from the same population, error variance is expected to be higher when the error variance is much higher, this is also an indicator of sample error.

Sampling errors are also those errors which arise on account of sampling. And they generally happen to be random variations, mean the sample estimates around the true population values. The magnitude of the sample error depends on the nature of the universe and more the homogeneous universe, less the error, as I already discussed. Sample error is inversely related to the size we have already discussed.

Sample size we have discussed sampling error decreases as the size increase and vice versa. Sampling error can be predicted, calculated, and accounted through several measures like confidence interval, we will discuss those indicators to discuss the sampling error where size also matters, one indicator is confidence interval standard error, standard deviation of the sampling distribution is called standard error, coefficient of variation, p-values, probability value of the sampling distribution is very very important, this keeps the decision factor for the researchers at what p-values we are going to stick we will just clarify this in our module.

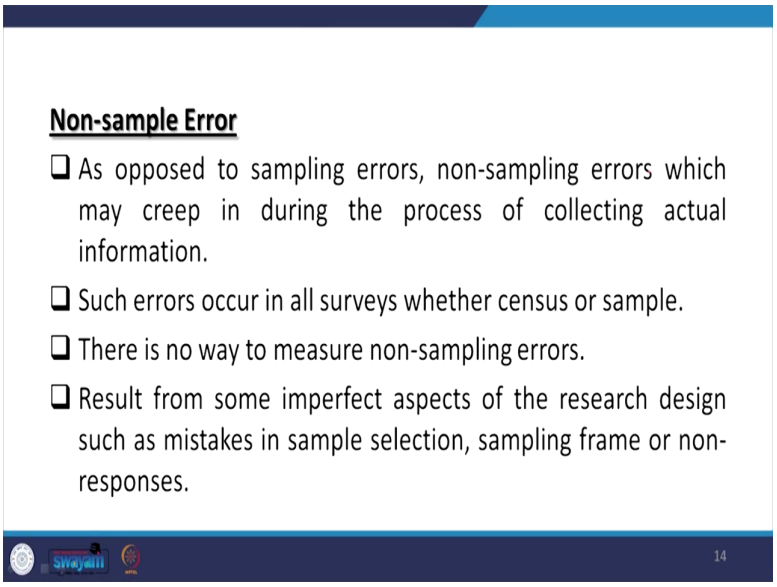
(Refer Slide Time: 17:00)



So, in this picture, this indicates that population here in the larger circle, then sampling frame is made, then the exact sample group is defined in the smaller sample within that once even the sample is defined where to take the sample the response may not be entirely the sample units that might be less responses. So, responses are figured under at the center, which is

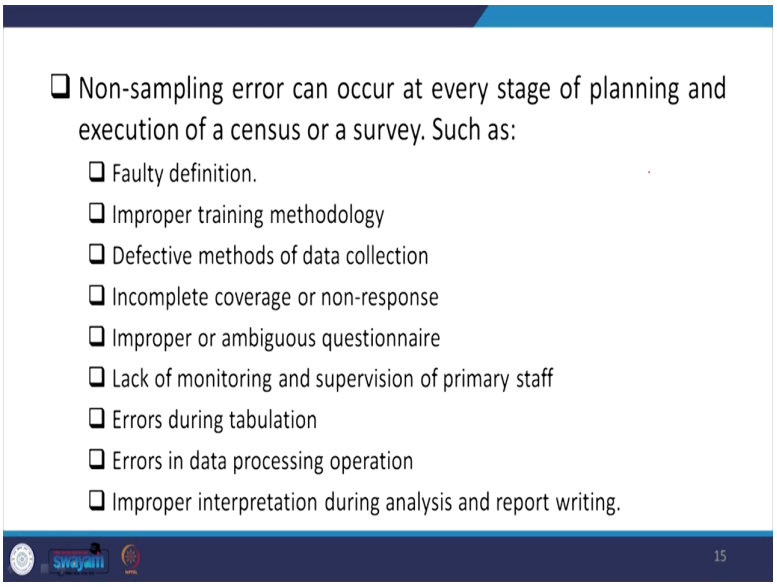
usually the less units. So, sampling error as I said, it might be due to frame error. So, why do we make a frame? Frame might be erroneous, chance error, response error, and if we add measurement error or the non-sampling error to the sampling error, we get the total error. Once we add or other indicators, we will get to the total error.

(Refer Slide Time: 18:02)



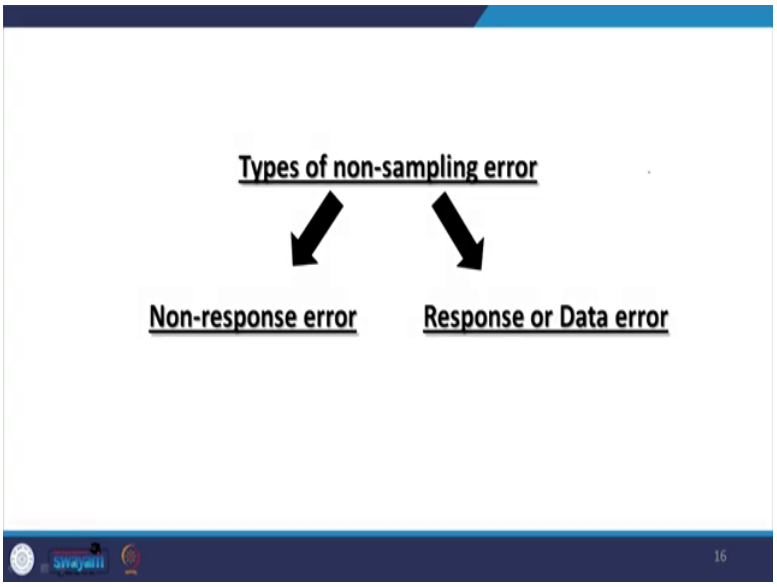
So, then what do you mean by non-sample error? Non-sample error is against the sample error, which may creep in during the process of collecting actual information. So, the process of collection is very important, such error occurs in all surveys, whether it is census or it is a sample, there is no way to measure sampling errors result from some imperfect aspects of the research design such as mistakes in sample selection, some sampling frame or non-responses are important.

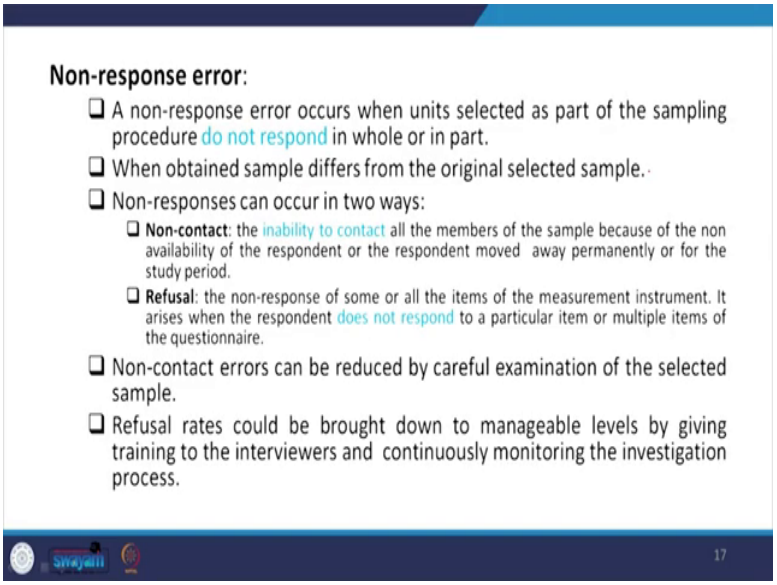
(Refer Slide Time: 18:35)



Non-sampling error can occur at every stage of planning and execution of the census or survey. Such as faulty definition, improper training methods, defective method, methods of data collection, incomplete coverage, or non-response improper or ambiguous questionnaire, lack of monitoring, and supervision of primary staff errors during tabulation, errors in data processing and its operation, improper interpretation during the analysis and report writing these are all possible non-sampling errors.

(Refer Slide Time: 19:10)

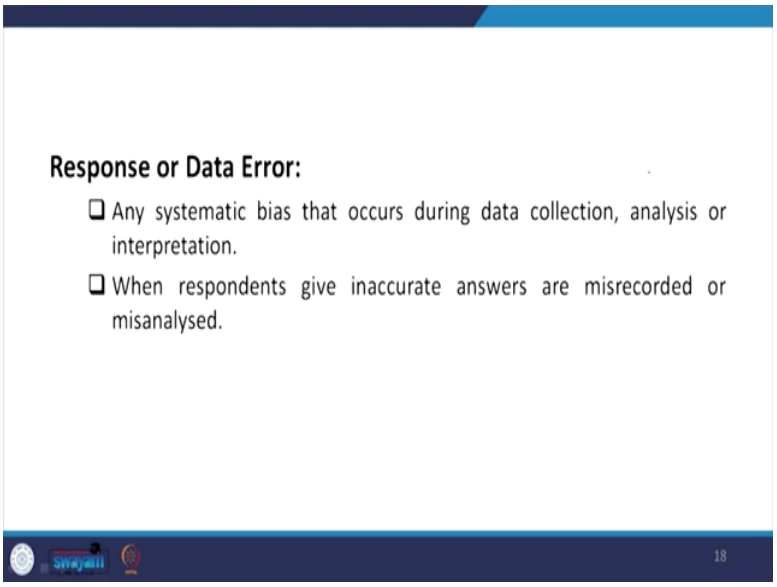




So, non-sampling error are of two types, one is non-response error and response or data error. So, what do you mean by non-response error? Non-response error are those where units are selected as part of the sampling procedure, please do not respond in whole or part. So, and when obtain the sample differs from the original selected sample. When there are differences then this also occurs. None responses can occur in two ways. One is non-contact or refusal.

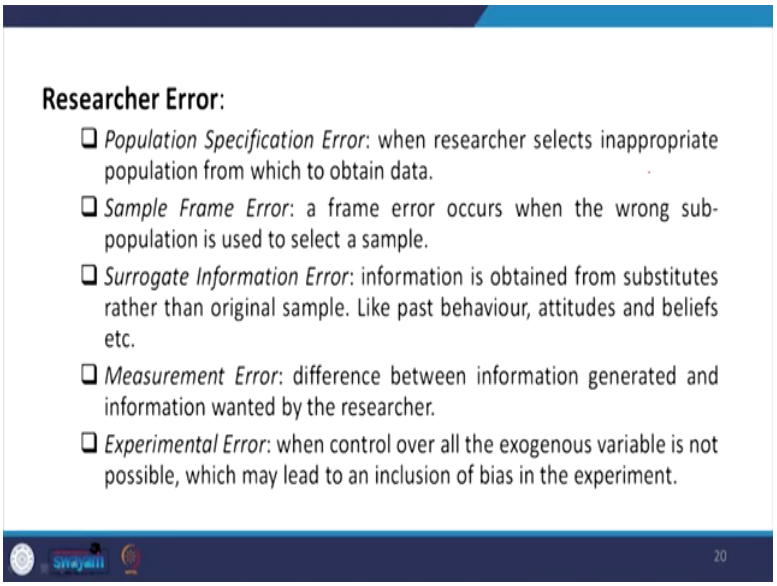
Non-contact errors can be reduced by careful examination of the selected sample whereas the refusal rates could be brought down to a manageable level by even training to the interviewer and continuous monitoring the process of investigation.

(Refer Slide Time: 20:14)



So, what do you mean by response or data error? It is usually a systematic bias which occurs during the data collection or analysis or interpretation stage and that too when the respondents give inadequate answers that mislead the information and sometimes these are miss recorded or miss analysed or called data related error. So, data error again classified into majorly of three types researcher error, interviewer error, or respondent error. So, what do you mean by the researcher's error?

(Refer Slide Time: 20:53)



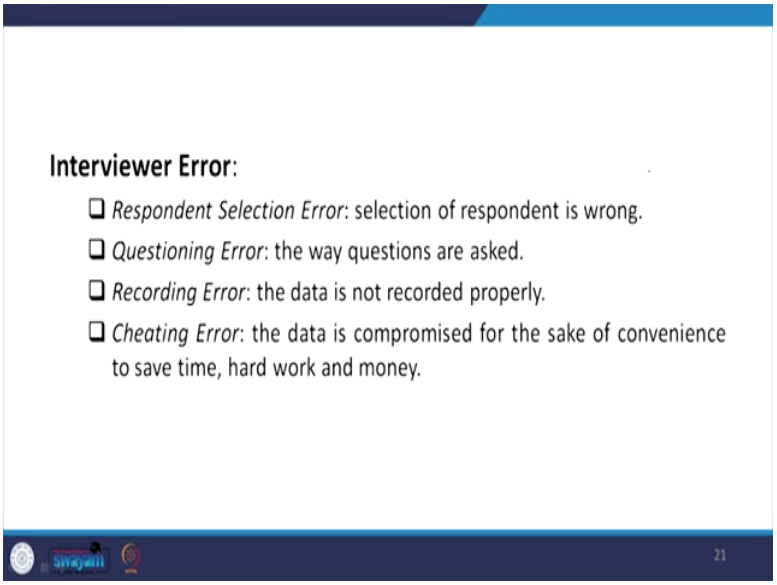
The researcher's error may be mentioned as population-specific. An error when a researcher selects an inappropriate population from which we obtain the data. So, these are called population specification. Similarly, sample frame error, we have already discussed that.

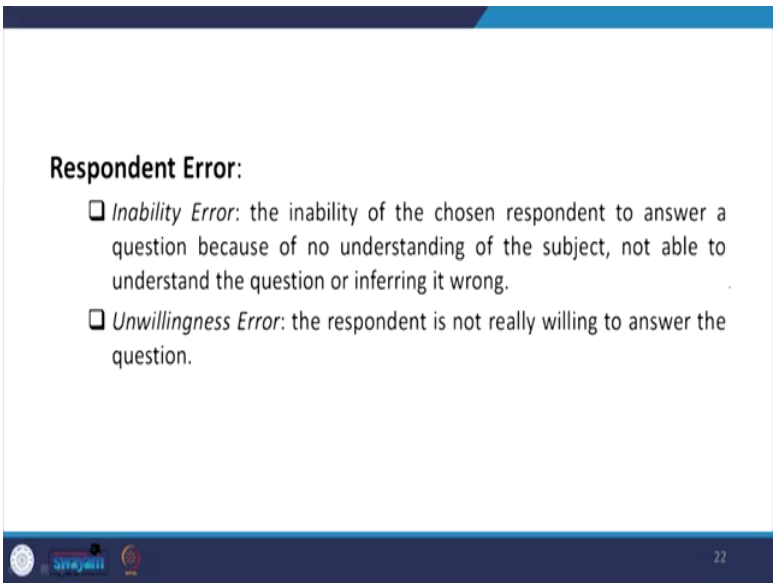
Sample frame occurs when the wrong subsample is used to select a particular sample out of the total universe you have bracketed into different subsamples but sub-samples may not be perfectly fit, you wanted to study the health-related information for and their gender divide, but you have not been particularly categorized into gender differences, you have simply randomly selected health issues from the population.

So, gender issues are not correctly represented. So, those called sample frame errors. Surrogate information error is often from software tools rather than the original sample. For example, past behaviour if you consider attitudes of the person believes are generally creating errors. Similarly, measurement error difference between information generated and the information wanted by the researcher if the information required is not properly met with the collected, then there are measurement errors.

Similarly, experimental error when we control over the exogenous variable, which is not possible, we are trying to control but not possible may lead to an inclusion of bias in the experiment. So, there are lots of bias in the experimental design create the error.

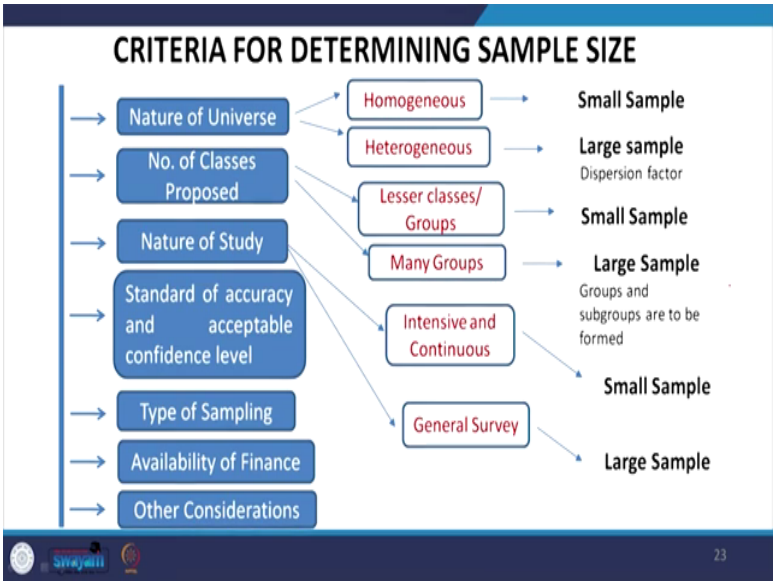
(Refer Slide Time: 22:32)





So, the interviewer error, respondent selection error, selection of responded is wrong questionnaire, questioning error questions there are problems with the questions, recording error, cheating error like data is compromised for the sake of convenience to save time, this is usually, the case in the field survey. And to minimize money also that creates some kind of error. Respondent error is another error which our due to inability and due to unwillingness.

(Refer Slide Time: 23:09)



And, what are the criteria of determining sample size. So, the criteria broadly after discussing the sample error and understanding the sample size and its framework, why sample size really important. So, let us understand the criteria of the sample size. So, the criteria request you in understanding the nature of the universe the entire population of our targeted survey, number of classes proposed within the subgroup. How it is defined? And natural universe

particularly, we are interested broadly here to know whether the universe is available in a homogenous format or in a heterogeneous format.

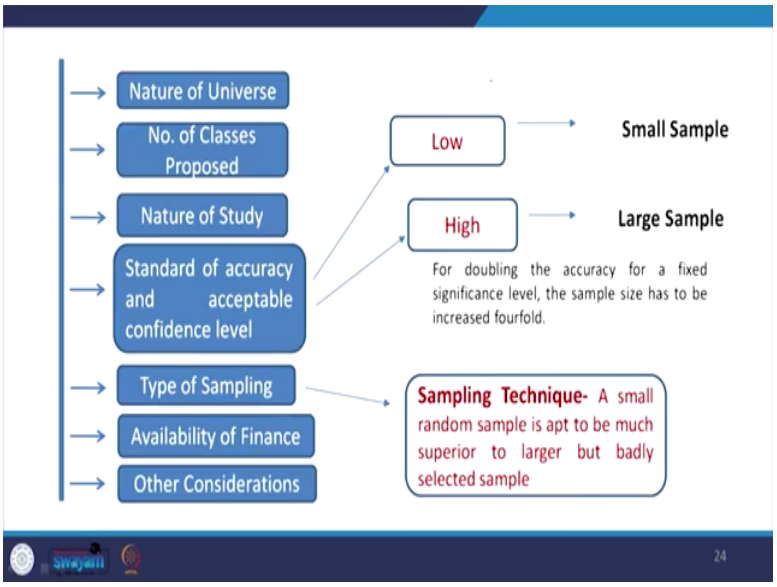
Suppose, it is homogeneously distributed, the universe is homogeneously distributed, then as I mentioned from the beginning of our sampling design, that we require less sample because less sample can represent your entire universe when it is homogeneous. But if it is heterogeneous, there are huge variants from person to person or group to group within the universe, then you are supposed to take more sample size because you are supposed to divide more sample groups or subgroups.

So, when the universe is heterogenous you are supposed to collect more sample size that is a standard rule, because of the dispersion factor. What do we know about sample size in the case of the nature of the study? Nature of study is important whether your nature of the study is very intensive, very core, then you especially require qualitative information. So, in that case, you need not cover more sample size, you may simply cover a small sample size. So, the standard rule is to stick to small sample size and study intensity. If it is a general survey, obviously it is quite a large sample.

So far is the number of classes proposed, number of subgroups is important, the subgroup is there if and only if your subgroups are defined and subgroups are lesser. There are less number of subgroups obviously, you are supposed to collect less sample size. If more subgroups are there, then large sample is required and more subgroups are defined as I mentioned, it is due to variability or dispersion factor as well.

Other indicators like standard of accuracy or acceptability related to confidence level if the confidence level is higher, then in order to make your confidence level higher you are supposed to collect more sample size. Then type of sampling what kind of sampling you are adopting? Available to finance and other constraints.

(Refer Slide Time: 26:13)

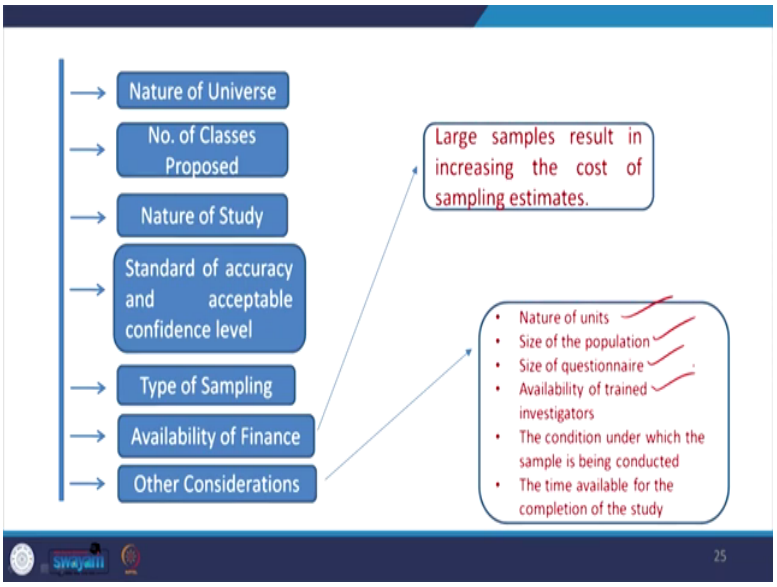


Let us explain some of those. Regarding standard confidence level or accuracy, either it will be low or high, if you want it to lead with a lower level of confidence, then you need not require more sample size, sample sizes is low in that case.

Whereas, to make better confidence among your researchers or in a study that nobody can reject your answer or result. In that case, you are supposed to take a large sample for doubling the accuracy for a fixed significance level, the sample size has to be increased fourfold, is a standard rule defined by statistician. In order to double your accuracy, you are supposed to four times increase your sample size, then regarding the type of sampling you are supposed to follow a small random sample is apt to be much superior to larger but badly selected sample.

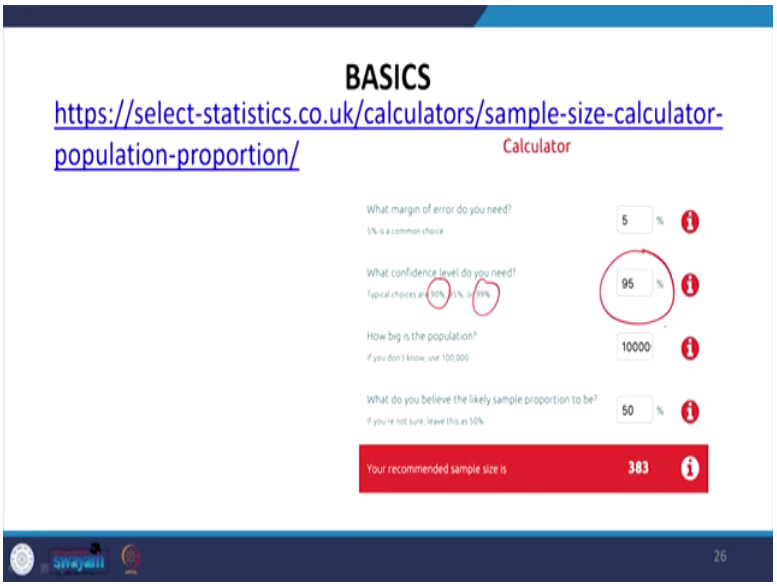
So, type of sampling is very important, small versus large-small is maybe very apt also depending upon again your universe of Genesis. And regarding the availability of finance, this is important because this result in a cost to the hardware cost estimates is very important from the beginning.

(Refer Slide Time: 27:51)



So, availability of finance is very very important. The sampling framework or the other indicators, other restrictions or considerations, like nature of units we are collecting, size or population, size of questionnaire, the availability of trained investigator. Trained investigators are there with the less sample information they can able to estimate result the condition under which the sample is being conducted, the time of surveys also important to deal with this aspect.

(Refer Slide Time: 28:17)



Those who are interested more in understanding the sample size, I think so far in my analysis of till the date, every researcher is interested in understanding sample size, they are often confused what sample sizes should take. So, some of the indicators were already discussed.

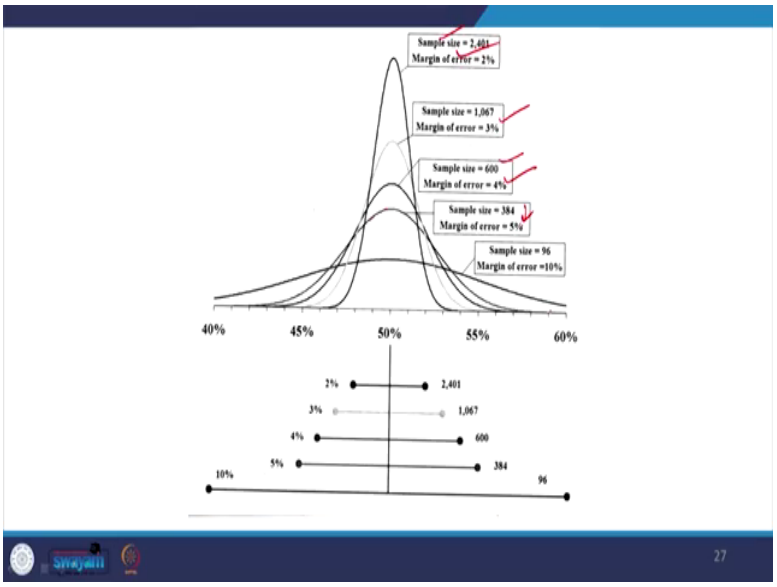
Accuracy and confidence level is very important. So, as I already said, accuracy level, maybe 95 percent, maybe 90 percent, maybe 99 percent these are all important. So, error report, margin error we will discuss with its calculation. Margin error usually decided at 5 percent level, is a common choice the error is acceptable. Error is acceptable at 5 percent.

And if your population in total based, on this website if you click on this link, select-statistics dot co dot uk and simply click this website and you will be directed to have these indicators to fill off this indicator like what is your margin error? Let us set 5 percent for it. What are your competence level? Usually, 95 percent for it and what is the size if it is 1 lakh you can write 1 lakh there if it is 50,000 you will enter 50,000 therefore our sample answer specifying here at 10,000.

What is the proportion of your sample to be included? What is your sample proportion? What do you believe the likely sample proportion to be? Usually, it is 50 percent, 50 percent accepted, 50 percent of the population have the chances of inclusion in your sample. So, what do you mean by that? We are not distinguishing population each person has an equal probability of selection by half, 0.5 is the probability of inclusion in the sample.

So, this is 50 percent means half. So, if you enter that it suggests a sample size called 383, we will discuss some formula Cochran formula, even there are certain other indicators of including like, as I mentioned, we have some alternative scenario also I will come to the discussion correctly that how much to be included and how we are calculated will go by the exact formula okay the formula to be derived here and discussed.

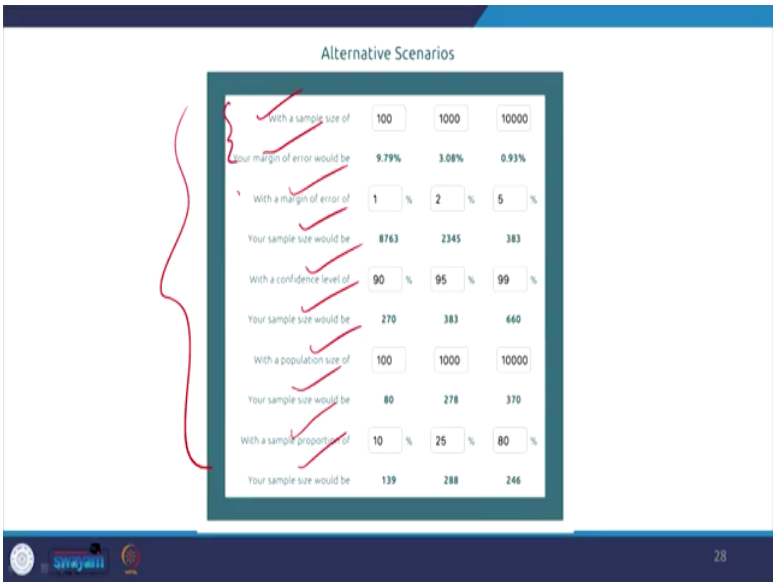
(Refer Slide Time: 31:23)



As we all know that sample size, margin error if it is 2 percent then, how many samples size to be there? If it is 3 percent sample size? Look at this, if margin of error is higher, you need not include more sample size, margin of error is more sample size reduce or if sample size reduces margin of error increases sample size again reduce margin error increases.

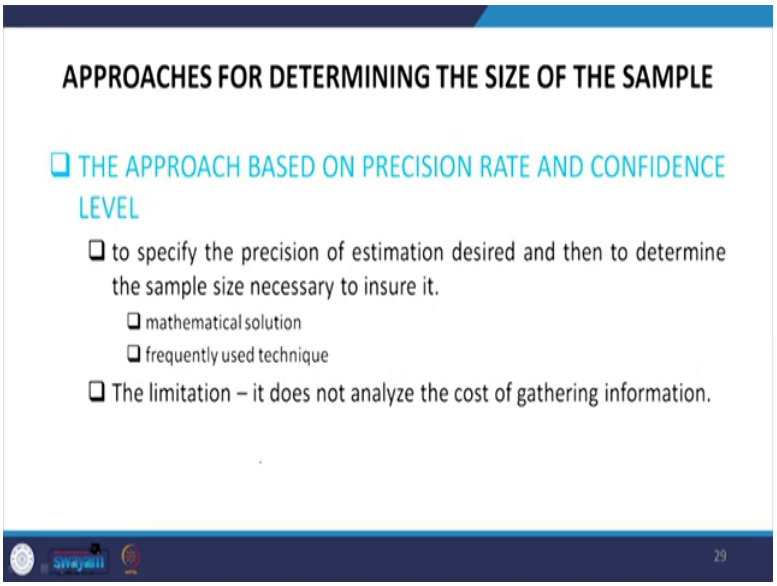
If it is 384 or 383 here margin of error is 5 percent. This is indicated in this diagram and the sample says 96 margins of error is 10 percent, even other indicators to be intact. So, the distribution is defined accordingly, we will discuss this normal distribution with different ways of whether it is symmetric or non-symmetric in detail in our sample size determination.

(Refer Slide Time: 32:17)



And what I just said is indicated in this particular box, sample size with margin error, then margin error with sample size confidence interval and sample size, population size and sample size, population proportion and sample size, you can have a comparison, these we can have a comparison you will understand correctly.

(Refer Slide Time: 32:42)



So, in the next class, we will discuss further on it. And we will carry forward the sample size discussion to the next class. And we will discuss these aspects like, which approaches are important in understanding or determining the exact sample size with equations, with the exact theory, and their calculation with a certain formula, we will discuss in our next class. So, with this let me stop here. So, I will catch you up in the next class. See you.