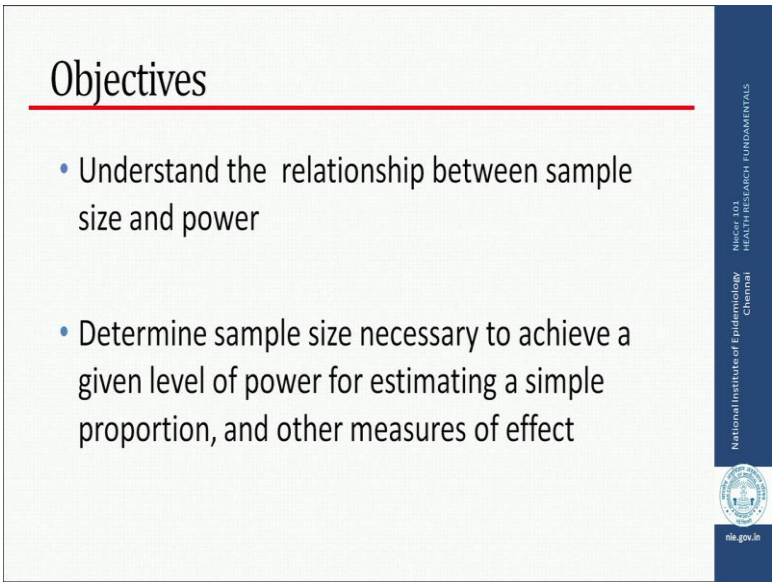
Health Research Fundamentals
Dr. R Ramakrishnan
ICMR - National Institute o Epidemiology, Chennai

Lecture - 12
Calculating sample size and Power

Hi. In this NieCer course on Health Research Fundamentals, in this module, we are going to see some aspects of sample size, how much subjects you require to do a study and some of the concepts that goes behind the Calculation of Sample Size.

(Refer Slide Time: 00:32)

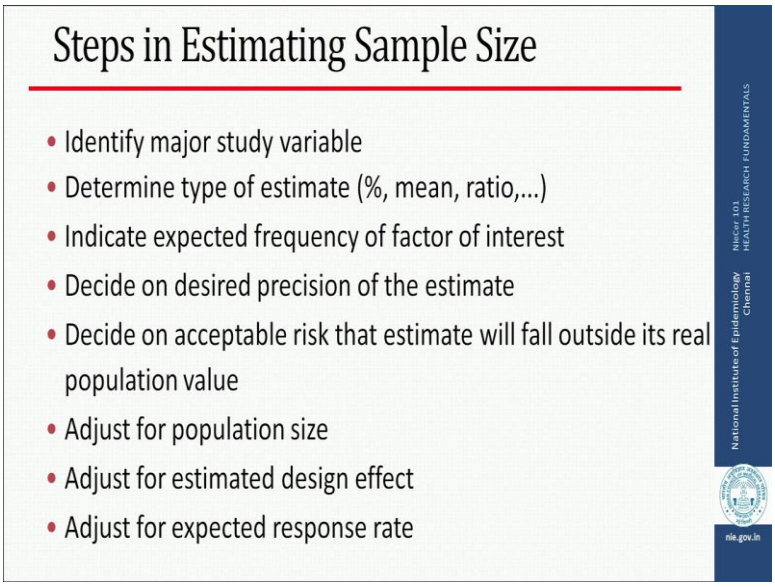


The usual question most of the investigators they have in their mind when they want to rather start doing a research study is, how much subjects that I should recruit in to my study? How many patients I should see? How many households that I should cover? And a simple answer for this question is there is no simple answer. This requires a little bit of logical thinking like, and usually it depends on some of the information that we already should know before we start our study.

So in this module, we will try to understand, what's the relationship between sample size and I am going to introduce you, to a concept called power. And, we also try to rather

determine the sample size, which is absolutely essential or necessary to achieve a given level of power for estimating may be a simple proportion or any other measures of effect.

(Refer Slide Time: 01:58)



As I was mentioning to you, there is no simple answer for estimating a sample size. We need to go in a systematic manner and let me rather take you step by step in the process of estimating sample size. First of all, you will have to identify, what is a major study variable you are planning to study? In any investigation, you will be seeing a number of things, you will have to identify which among them is the most important variable, which you want to rather study about. Say for example, when you are studying on may be scrub typhus in a community; your aim may be to estimate a prevalence of scrub typhus in which case, the variable whether a person has got scrub typhus or not is a major variable.

Suppose, if your interest, is not on the prevalence but on some of the associated factors, whether a person has been exposed to a forest or something like that, in that case that becomes the major study variable. Then the second step is to determine the type of estimate. Are you going to rather study at mean or a ratio or a percentage or proportion because accordingly we need to rather re-frame or have a formulae for computing the sample size.

Then one of the important things that comes out it is, you need to indicate the expected frequency of factor of interest, common sense says, suppose if you are going to rather study something very rare you need to rather have a large sample, unless you see a very large number of people you may not probably get sufficient number of people with the factor of your interest. On the other hand, if you are going to rather study something which is very common, in that case you do not need a large sample even and a small sample you may be able to rather give fairly a good sufficient and a precise estimate of your factor of interest.

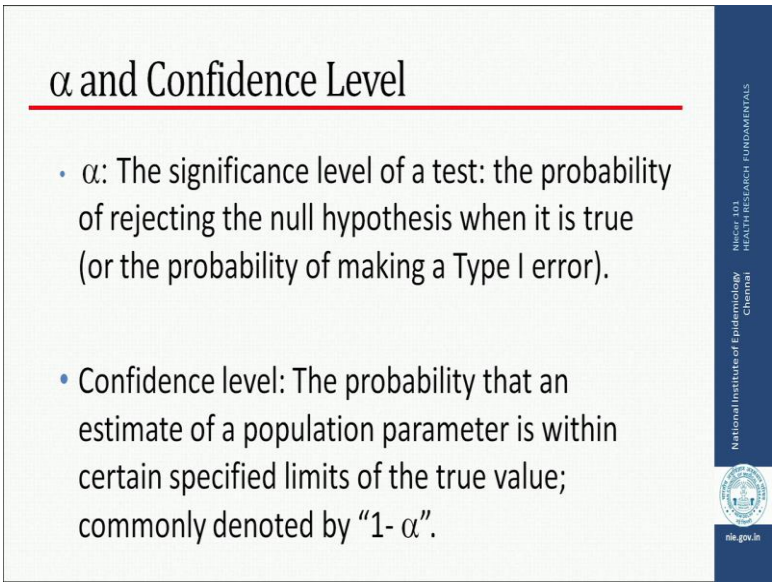
Then the next factor is, the decide precision of the estimate. How precise you want your estimate to be? You want your estimate to be within 5 percent this side that side or within 10 percent this side that side. As you want your estimate to be more precise then naturally you need to rather have a larger sample. If you are willing to rather give say + or - 20 %, then probably your sample size will be small as compared to + or - 10 %.

Then the next point is, I want + or - 10 % but how sure I want that my estimate is + or - 10 %? What is the amount of rather risk that I am willing to accept? Whether a 5 percent risk or a 10 percent risk. So, these are all some of the elements that are essential to compute the sample size and invariably these are all the elements that has to be rather given by the investigator to whoever is computing the sample size. The other three items that I have rather given are, you have to adjust what population size. Are you going to take your sample from a very large population? Or you are going to rather take from a small population? Because usually the sample size formulae assume that you are taking a sample from a very large population. If you are going to rather take your sample from a small population, you need to do some adjustment factor into it.

The next is, adjust for estimated design effect. See in my earlier lecture on sampling I talked to you about a clustered design effect, wherein you are going to rather select not individual subjects as your sample but you are going to rather select cluster of subjects as your sample. There could be a correlation between the subjects in the same cluster. So, in order to get over it you need to multiply your sample size by a factor called Design effect so that you have a larger sample which takes care of this correlation within the subjects in a cluster.

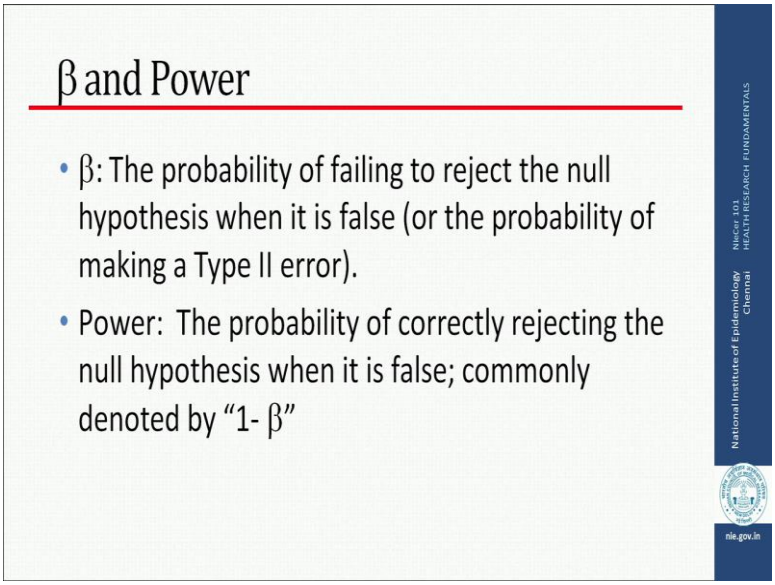
Then the last bullet point in this, it is adjust for expected response. You have to rather you know, you decide that you want to do 300 and if you just go and study 300 maybe you know 10 percent of them they did not turn up and you have only 230. In order to adjust for that you have some may be 10 percent extra as your sample size so that assuming a non-response, you still have sufficient sample to answer your question.

(Refer Slide Time: 07:31)



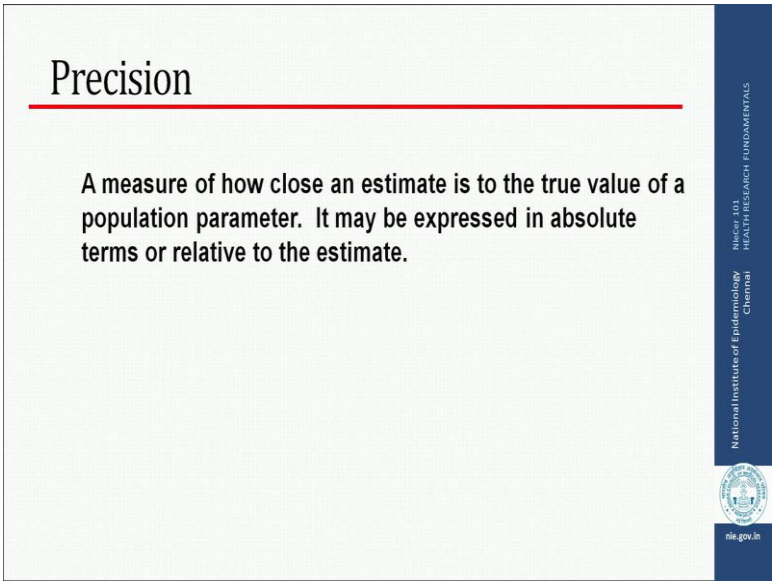
I am going to introduce you now to some concepts, which are essential to understand the computation of sample size. The first one is the α or the Type I error, the significance level of a test. What do you mean by that? It is the probability of rejecting the null hypothesis when actually it is true. In the statistical parlance, it is called Type I error. And the confidence level is the complement of that, that is 1 - α and that is naturally the probability that an estimate of a population parameter is within certain specified limits of the true value.

(Refer Slide Time: 08:23)



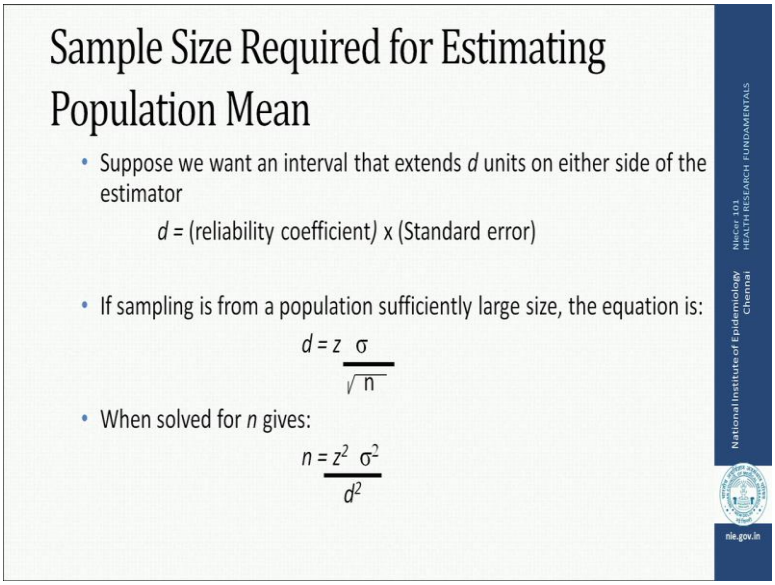
The next are β and Power. β is nothing but the probability of failing to reject the null hypothesis when actually it is false. So, if something is false you have to reject it, but you accept it and the probability of making this is called a Type II error. And the complement of that is 1 - β is commonly denoted as power and which is a correct decision and that is nothing but probability of correctly rejecting the null hypothesis when it is false.

(Refer Slide Time: 09:01)



Another important concept that goes into the computation of sample size is the Precision. By precision what you mean is? It is a measure of how close an estimate is to the true value of a population parameter. It may be expressed in absolute terms or relative to the estimate. They say + or - 10 % or + or - 10 % of the estimate.

(Refer Slide Time: 09:27)



Now, let us look into a scenario where you need to compute a sample size and your interest is to estimate the Mean of a Population. You have a sample, you have to compute a sample mean and you want to estimate the population mean from your sample mean. And what should be the sample size that you need? So, the general idea of the computation of sample size is, it is always a:

(reliability coefficient) X (standard error) is called d

and through d, we can estimate the sample size n. Say, the d is a formula given here is:

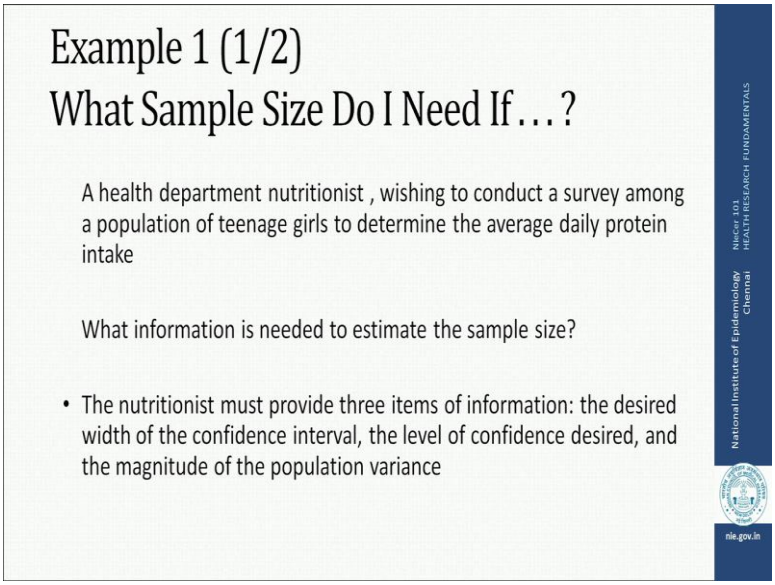
$z \times \sum / \sqrt{n}$.

How do we get that? We get that mainly using the concept of a sampling distribution.

What do you mean by sampling distribution? Suppose, I take several samples of the same size and each of the sample given estimate of the population mean and if I have a distribution of all those sample means, theoretically, it is proved that that distribution is at normal distribution and also the standard error which is the standard deviation of that state distribution is given by $\sum / \sqrt{n}$. So, what we normally we do it is? We try to rather have see the using the principles of normal distribution 2 $\sum$ limit of the sampling error 95 percent of the values they lie. The z in the formula is nothing but the standard normal deviate for a particular level of significance. And suppose, the idea of sample size is you fix that and then when you fix that you have only one unknown namely, the n in the denominator and if we can solve for the n, which is nothing but:

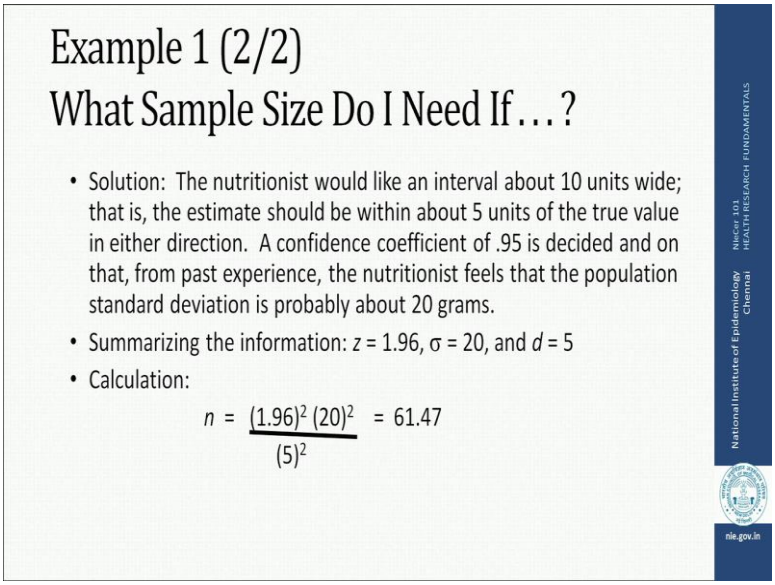
$n = z^2 \times \sigma^2 / d^2$, then you get an idea of what the n is.

(Refer Slide Time: 11:47)



I think you know these concepts will get clearer, if you see an example. A health department nutritionist, he wishes to do a survey among population of teenage girls to determine the average daily protein intake. So, that is the research problem. And what information is needed to estimate the sample size? So the nutritionist must provide three items of information. The first one, the desired width of the confidence interval, the next one, is the level of confidence desire and you should rather give a rough magnitude of the population variance.

(Refer Slide Time: 12:36)

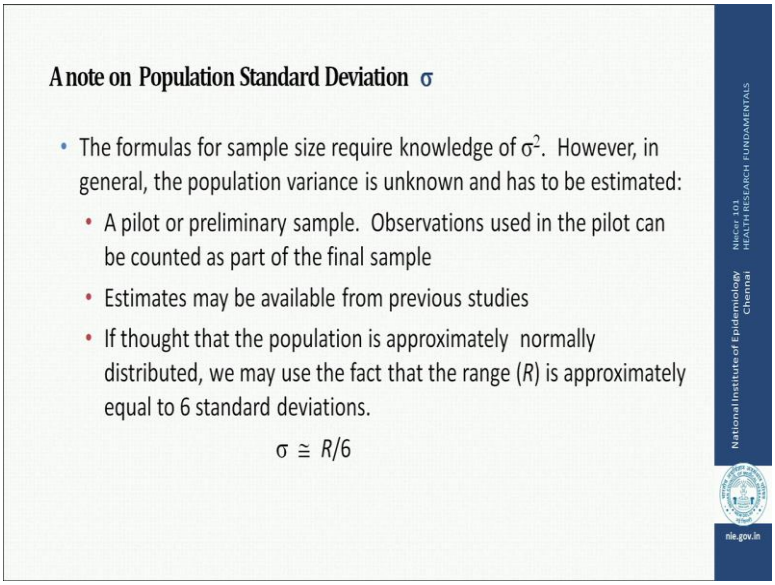


Assume, that he gives them all, say the nutrition feels that the 10 units this side and that side is what he is expecting, which means now 10 units on the whole, so 5 units this side and 5 units that side and the confidence coefficient of 95 percent is decided upon, and from his past experience from the literature review, the nutritionist feel that the population standard deviation is probably about 20 grams. Now, we have the information z is 1.96 because 95 percent confidence interval has got a corresponding z value of a normal distribution 1.96. ∑ is already given as 20 and then the desired length d is 5 units this area or that side. If you plug in all these value into the formula, it becomes:

$n = (1.93)^2 \times (20)^2 / 5^2 = 61.47.$

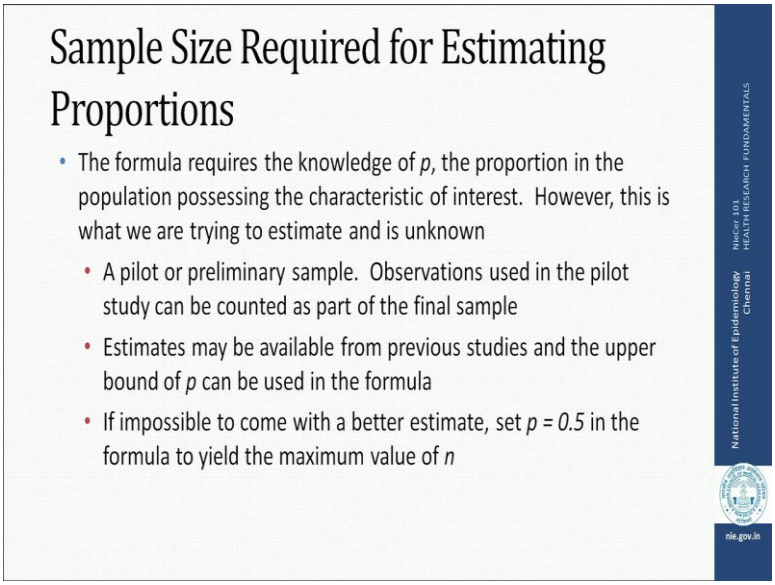
Which means you need to have at least 62 teenage girls in order to get an estimate of the mean protein intake and the estimate that you gave 95 percent of the time will be within 5 units this side or that side of the true mean population mean of protein intake.

(Refer Slide Time: 14:08)



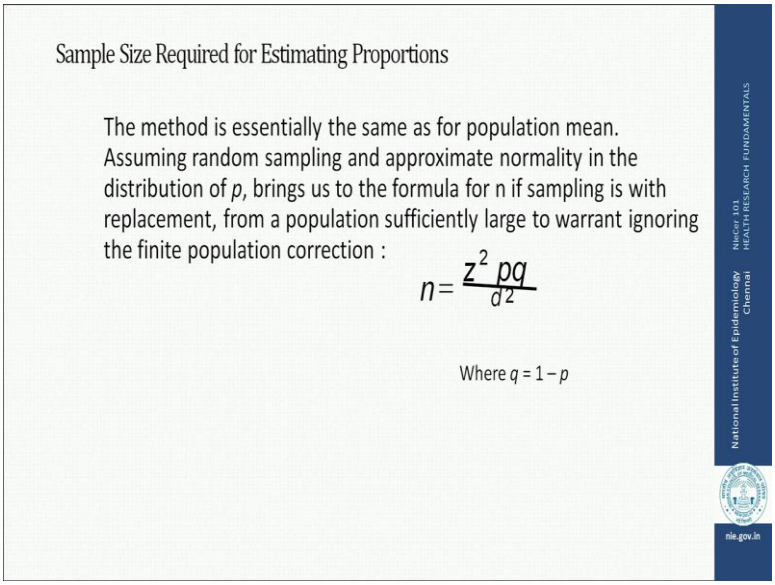
In this formula, we have ∑ that is the population variance. And in most scenarios you may not probably know a value of ∑ because this is study you are going to rather do ∑ may not be available and how to get a ∑ (σ)? One of the ways that you can get this variance is do a pilot survey, preliminary survey. Of course, you can even use this observation used in the pilot for your final sample tool. An estimate available from the pilot survey could be used or you can use an estimate which is available from previous studies, and suppose you know you have a large data available and you have the range of the data, assuming that it is normally distributed, you can get an approximate value of ∑ (σ) as the R/ 6. So, these are all ways of getting the value of ∑ in your formula.

(Refer Slide Time: 15:17)



Suppose, we are going to estimate a Proportion not a mean, the formula is more or less similar but what you need to give is you must rather have knowledge of p that is the proportion of the characteristic or the factor of interest in the population. This also may not be, see you are going to do a study to estimate a proportion and invariably when I ask this question to the investigator he says, Sir, I am going to do a study to find it, how can I have an idea of p, can probably as I had rather mentioned earlier he can probably do a pilot study to get an idea of p or you can get from the literature, what could be the value of p and if it is impossible then the best thing is to estimate to get a value of p as 0.5, so that it is the maximum value of n.

(Refer Slide Time: 16:19)

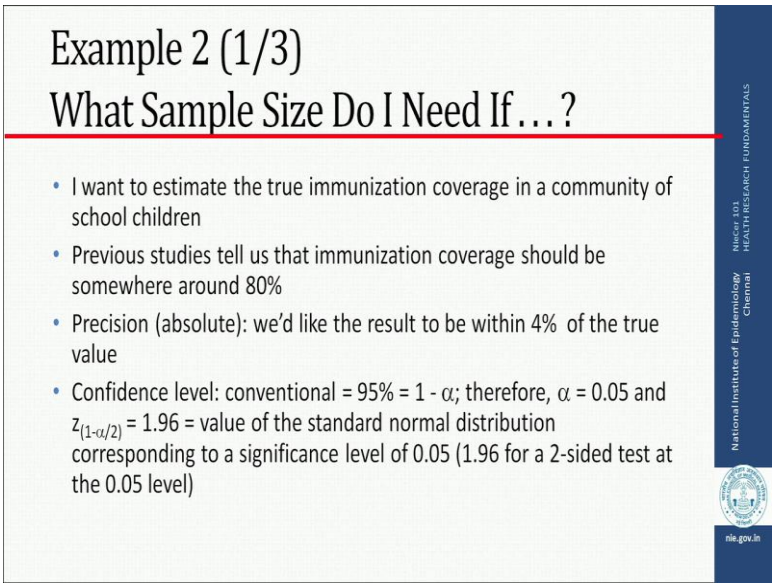


So the formula for that is very simple:

$n = Z^2 \times pq / d^2$

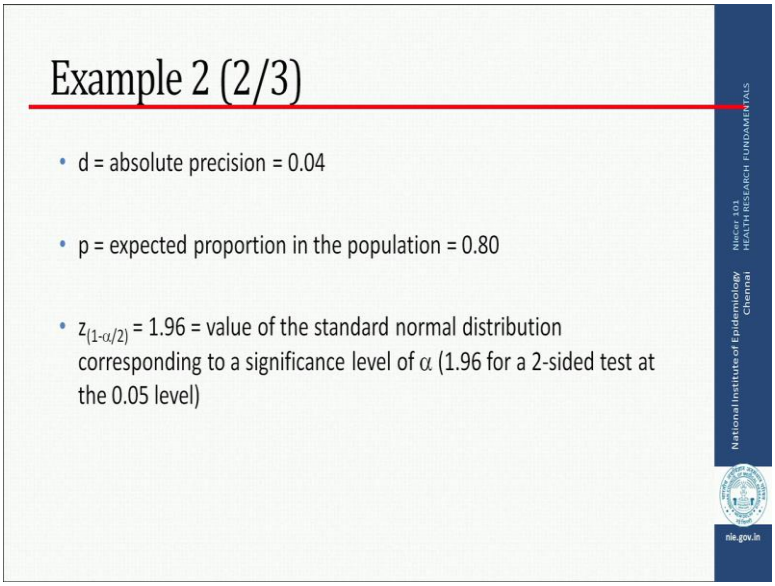
and here instead of your ∑ what you have is the pq, where q is nothing but 1 - p.

(Refer Slide Time: 16:38)



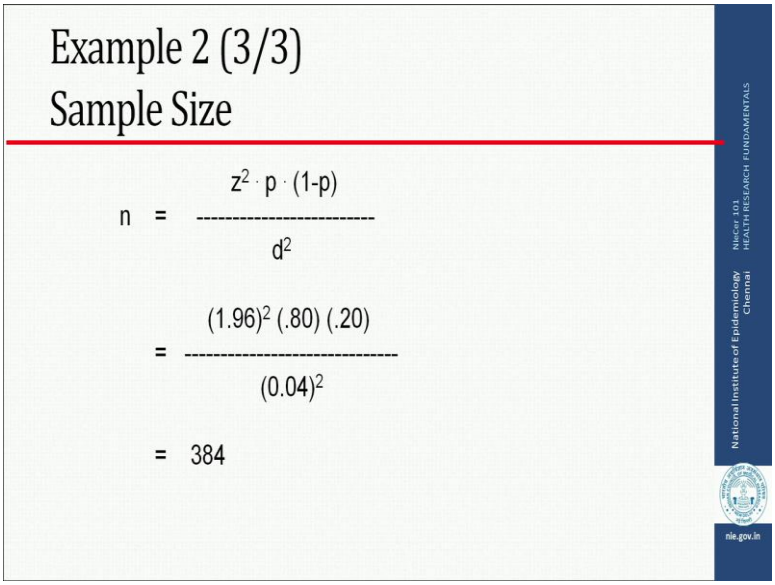
This also will be clearer if you see an example. Suppose, want to estimate the true immunization coverage in a community of school children. Previous studies tell us that the immunization coverage should be somewhere around 80 percent. Suppose the absolute precision, we would like the result to be within 4 percent of the true value. Then the confidence interval which is conventionally taken as 95 percent and 1 - α therefore, there are 5 percent α level, z α is 1.96. Then we have all the values that are needed for our calculation, d the absolute precision is 0.04, p the expected proportion of population is 0.8.

(Refer Slide Time: 17:20)



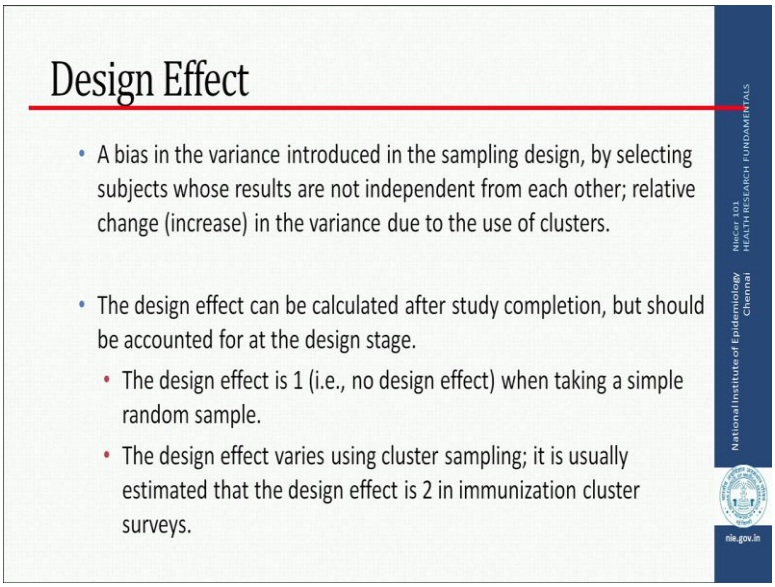
So naturally the q must be 0.2 then z α is 1.96. And when we plug them all into a formula what we get is 384.

(Refer Slide Time: 17:36)



So, you need 384 subjects to get an estimation of the immunization coverage, within 4 percentages this side or that side and you have 95 percent confidence that the true value lies in this particular interval.

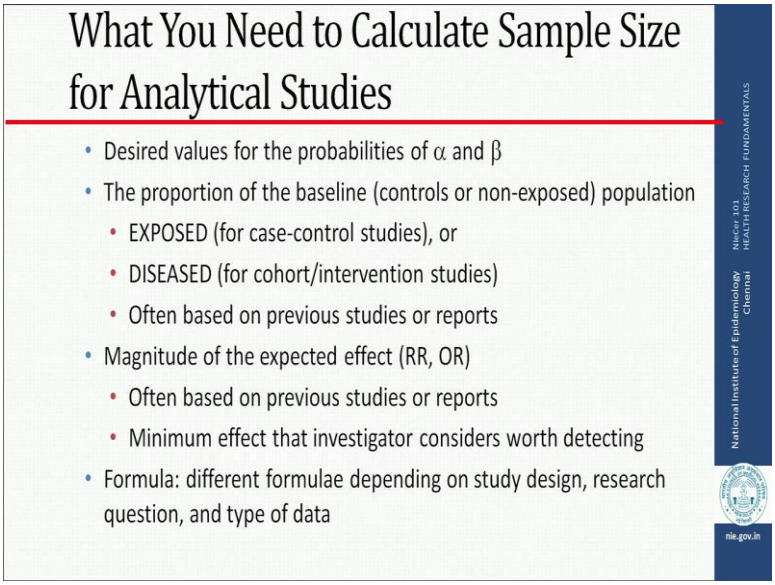
(Refer Slide Time: 18:00)



So, we had seen something I talked about the design effect earlier. The define design

effect is caused because of a bias in the variance introduced in the sampling design by selecting subjects, whose results are not independent from each other because in a cluster there may be you know, suppose a child is immunized in the first house hold there is a firmly a large chance that the child in the next house also is immunized, you cant say that it would be absolutely independent. In order to account for that you need to multiply your sample size by a factor called Design Effect.

(Refer Slide Time: 18:45)

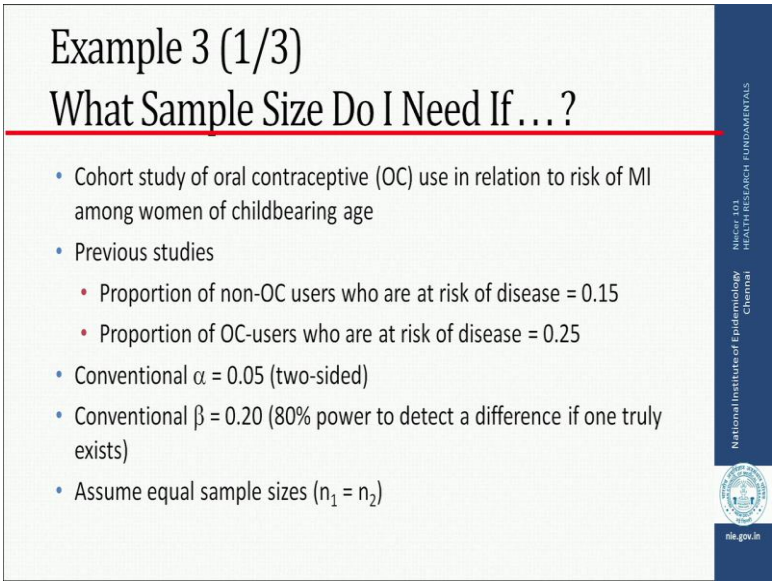


In similar sort of scenario, suppose we are going to rather do a case-control study or a cohort study and how you should go about in estimating the sample size. What you need is the desired value of the probabilities of α and β and the proportion of base line or there is a control or non expose population. In the case of case control studies, the proportion of exposed or in the case of cohort studies, the proportion of disease and you need to rather have some idea of them. These are all often based on previous studies or reports and also you should have some sort of an idea of the magnitude of the expected effect that is the magnitude of relative risk or odds ratio.

This again is based on a previous studies or reports. And what is the minimum effect the investigator considers? What is detecting? These are all some of the information once it provided then there are very easy formula available, the different formula depending on

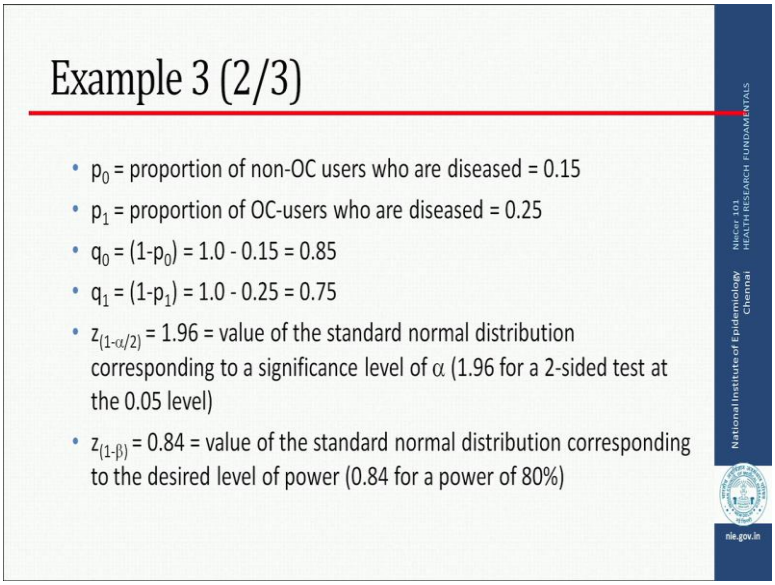
the study design, research question and the type of data.

(Refer Slide Time: 20:03)



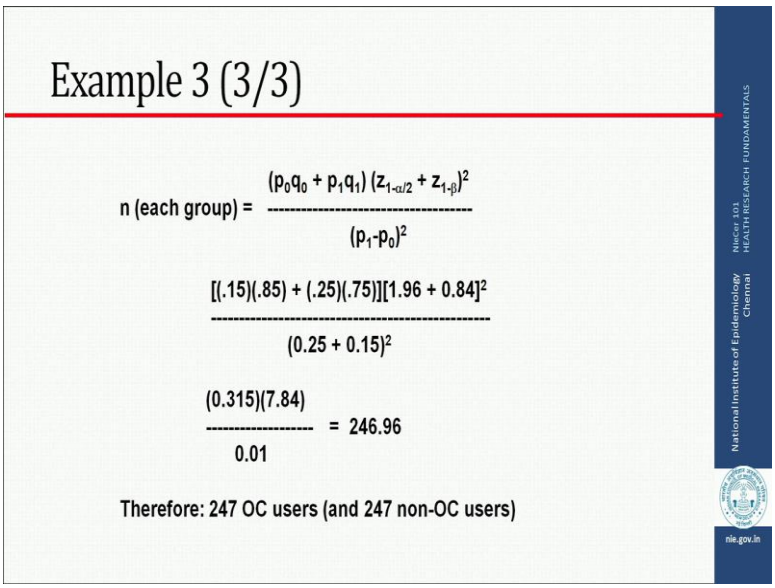
Now, let us take few examples and this example should rather give you some idea of, how the sample size is computed in different scenarios? Take for example, a cohort study of oral contraceptive use in relation to the risk of myocardial infarction among women of childbearing age. Previous studies have indicated that the proportion of non-OC users who are at risk of disease is 0.15 that is 15 percent of non-OC users' women of childbearing age have a risk of myocardial infarction. So, proportion of OC users who are at risk of disease is 0.25 and the conventional α is 0.05. Suppose, the β is taken as 20, 0.20 that is you want 80 percent power to detect the difference of it truly exists and assume that you are going to rather have equal sample sizes for your users and non-users.

(Refer Slide Time: 21:18)



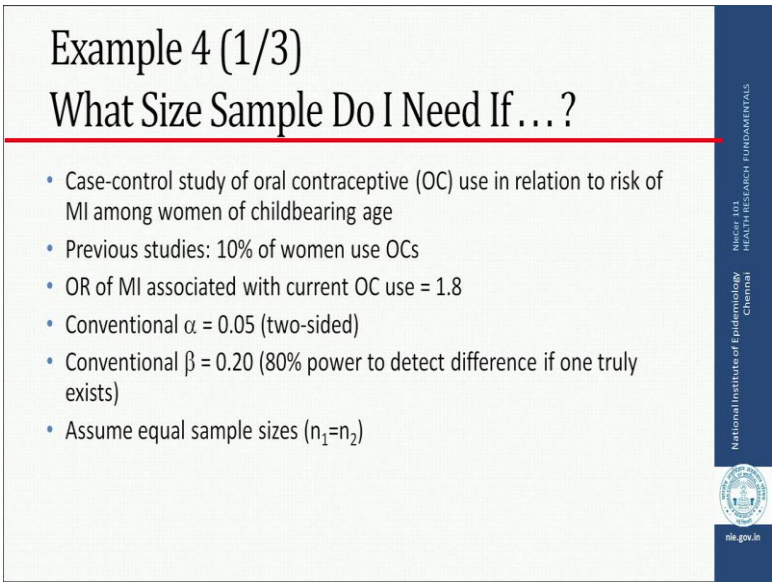
The formula is obtained using these following parameters you know p_0 which is nothing but proportion of non-OC users who are diseased, which is given as 0.15. P1 is proportion of OC users who are diseased, which is given as 0.25 and your q_0 which is a complement of p_0 is 0.85. Your q1 is a complement of p1 which is 0.75. Z α is 1.96 we saw in the last example. Z β is 0.84.

(Refer Slide Time: 21:53)



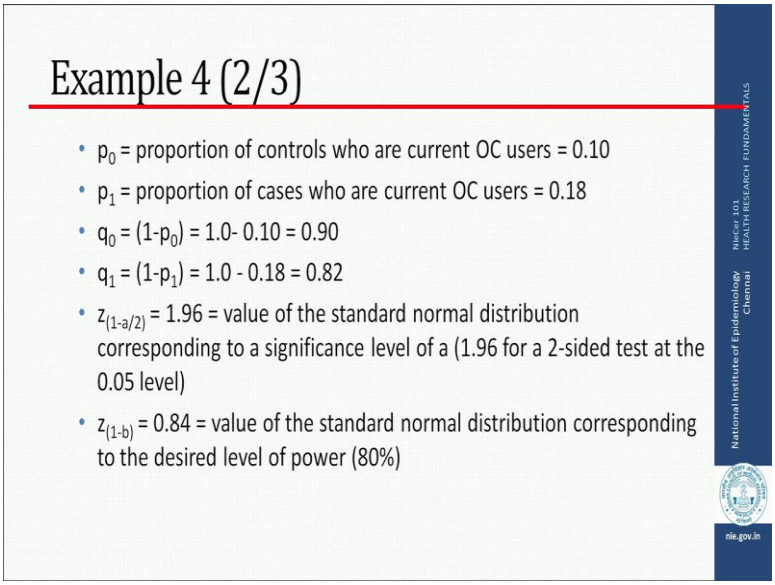
We have all these values which can be plugged into the formula, which gives n is equal to 216.96 or 247. So we need to have 247 OC users and 247 non-OC users, follow them over a period to get a desired result.

(Refer Slide Time: 22:20)



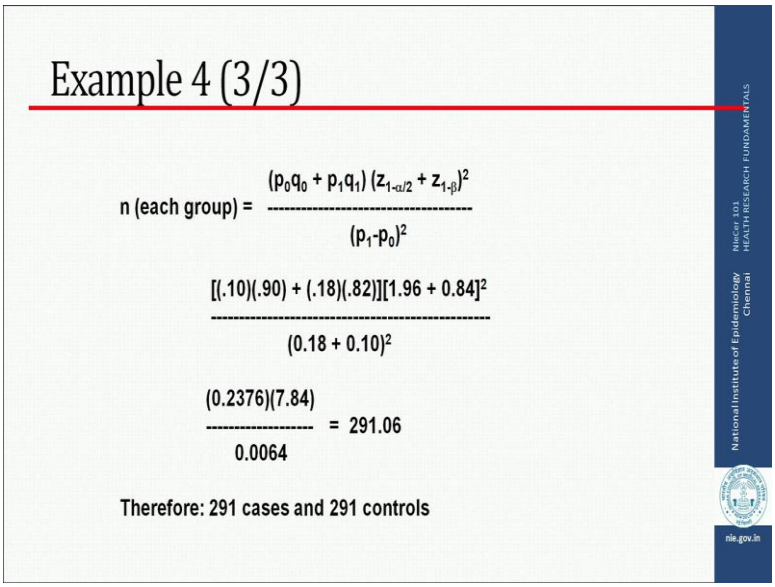
Now, let us take an example of a case-control design. How do you go about? In the case control study of oral contraceptives use in relation to the risk of myocardial infarction among woman of childbearing age. Previous studies says 10 percent of woman use OCs and OR of MI associated with current OC use is 1.8. Then the other thing as conventional α as 0.05, conventional β is 0.20 and assuming equal size for case in control.

(Refer Slide Time: 22:47)



And see you have all these parameters p_0 is equal to the proportion of controls who are current OC users which is 0.1. P 1 is equal to proportion of cases, who are current OC users and that is 0.18. q_0 is 0.9 and q1 is 0.82. z α 1.96, z β is 0.84.

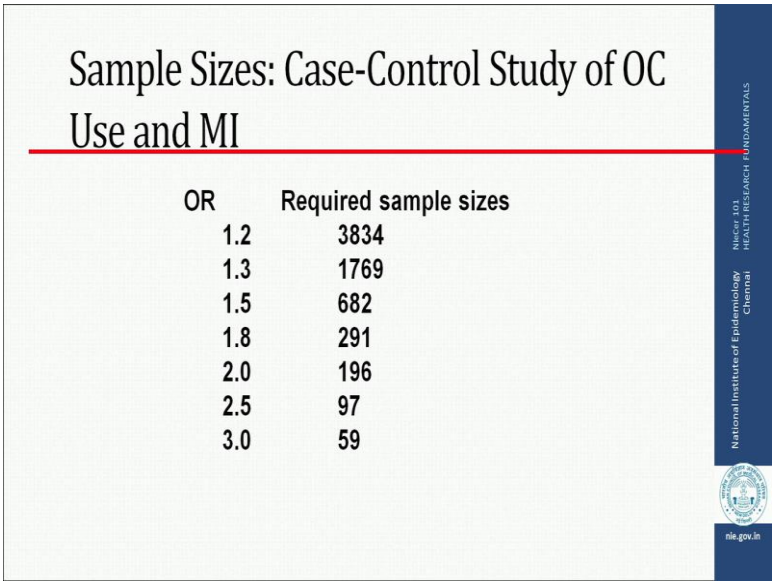
(Refer Slide Time: 23:27)



If you plug them on in a formula then you get 291.06 which indicate that you need to

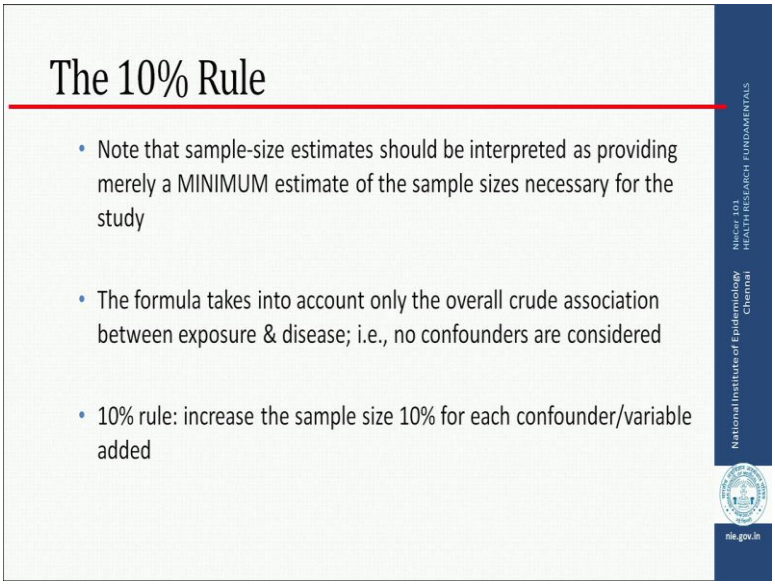
rather have 291 or 292 cases and 292 controls in order to get an estimate of your OR.

(Refer Slide Time: 23:43)



Now, this slide gives you the required sample size for various OR. Say for example, you want to detect an OR of 1.2 then you need a sample size of 3834. Whereas you have to estimate an OR of 3, it is enough you have 59 in each group. So, what it means it is, if you want to detect a very small difference then you need to have a large sample to identify that small difference. If you want to detect the large difference then it is enough you have a small sample, then you know you will be able to get an estimate of your OR which is 3 or more.

(Refer Slide Time: 24:32)



Then see in any of this analytical studies, when you are looking for an association of one variable, there could be a third factor which could be effecting the values of this association, which is in the epidemiological parlance, we call that as confounders. There could be one variable or two variables, which are confounders in a particular association. The general rule is, if you have some confounders in your studies you hike your sample size by 10 percent for every confounder variable that you have.

(Refer Slide Time: 25:09)

Now, having seen different scenarios, where the sample size is computed and then the concepts behind it, I am going to introduce you to 2 softwares, which are free softwares in the open source, which can be very easily used to compute sample size of different study designs. One is called OpenEpi and this software is supported by the CDC Atlanta and the website for that is www.openepi.com and it is very, very simple software to use. There is a tutorial in for this particular software which gives you some examples and depending on what sort of a study design that you have, you can plug in the values that the software ask and you will get the desire sample size.

Another one, which is called PS that is Power and Sample size calculation, this is by the Department of Bio statistics Vanderbilt University, this is also an open source software. This is also fairly user friendly software where you can rather compute to your sample sizes. So, wherever you do an investigation and when you compute the software, you give the software in your reference saying that you use this particular software and these are all the assumptions or these are all the values that you had rather plugged-in in this software, so that this is my sample size. This should be reflected in your method section.

So, to recapitulate the module, sample size there is no magic number as sample size available. Sample sizes have to be computed using various parameters that are supplied

by the investigator. Investigator may have some idea of it ready made, if he does not have those ideas you may have to probably do a pilot study to get this kind of an idea. And then you know it depends on, how much risk that you are willing to take? How much precision that you want on your estimates? And this is usually there is no fixed number we can always negotiate depending on the resources that are available in terms of money and time. Suppose, you know I say you need to rather have 300 and you do not have that much resource and you do not have that much time to do, you can always rather they know reduce the sample size, but you should know what price that you are going to pay for reducing you may have to compromise on the precision or the risk that you will be taking on this sort of estimates.

Thank you so much.