MCHB/EPI Atlanta GA Conference

December 10 - 14, 2005

Sample Size Issues and Determination in Maternal and Child Health (Part I)

RUBEN SMITH: Okay, I think that's what we'll do. Okay, good. Okay, I think it's time to start the continue to the class. Okay, now, I am going to start. We introduce you to the sample size. So I'm going to start a little bit formula, some more detail about the sampling. So, in this part of the presentation, I am going to present--this is the outline for this—for section three. I want to talk about the sample cycle collation. A little in detail like introduction. Also, I'm going to present the information needed before determine the sample size. I'm also going to introduce some notations that we're going to use. I'm also going to introduce some basic study design that we are going to use in the course. Also, I am going to present the different method for determining the sample size that the method--we decide what method to use will depend on the objective that we have. So what we want to do, what kind of analysis we are going to do with the data that we recollect. So--also, I'm going to end with a summary. This will be a very short section. And always that the sample size collection should always should be based in the way that is data that is going to be analyzed. So when we are going to start to thinking about sample size calculation, we need to think about what we are--how we are going to analyze the data because this is a very important question. And also that we need to think that the sample size calculation is not like an exact procedure. It's like approximation with the sample size that we're

going to need. And the sample size calculation provide an approximate minimum number of subject that we choose sample. So if we're going to--that we talked before. You are going to think or respond right, it's going to be higher, we need to take in account the information and try to size—it's very useful to increase the samples size for the number--for the percentage or the no respondent that we're going to have. And also that decision on sample size are always a trade off between cost and precision. So there's no good when we have very high precision but we don't have the resources to have the (inaudible) sample size. So always we need to have a balance about what we can do with the money that we have.

Okay, also, I want to think about (inaudible) when we compute the sample size, we need to think about what is the question that we're going to answer. So what we need, what are the question that we're going to answer and also the sample size, we also need to know, what study design we're going to use. There are several sample--style design for a sample. One sample is study. I was interesting in one parameter. Only I was interesting in one population. Or I was interesting in two population. Let's say like two strata, for example. So also, we depend at the point of interest, what we're going to do. So although we need to think about how--what is the style design. Also I'm going to present some--introduce you about some of the study design. Another sample design, another study design could be like a (inaudible) study or a case control study. Or a survey. So what we're going to do. So what design we're going to use to answer our question of

interest? Also what is the parameter that we're going to use to answer the question? So all of you will have a question of interest. We need to think about parameters. So (inaudible) think we are going to use statistical method, what parameter we need to think? So the parameter could be the proportion or the mean or the ratio, for example, or a difference of proportion. Or it could be the relative risk. So always think about what parameter you are going to use when you--because the formula they are going to use will depend on the parameter, and also will depend on the study design that you have.

Also, what you are going to do with the data, are you going to use—are you going to do like a hypothesis test or are you going to do estimation? Because these questions you will need to know before to go to a book or and select what sample size formula you are going to use. Okay, these are the notation I am going to use for this section that I am going to teach today and tomorrow and maybe (inaudible). Okay. For the big N will be the population size, yeah, the little n will be the sample size. And what I am going to use theta like a population parameter like this could be like a—is a—a population parameter and will represent, for example, the proportion or could represent a mean. So this is like a general parameter. I even have like a little hat will be the sample estimate.

So (inaudible) because we are going to use for the formula. For P, this will be the way when I use like in the general term. But we have specify we are going to use proportion, I am going to use like the P for sample proportion and the pi for

population proportion. And the sample mean is the y. So in the training, we are going to study these. We are going to present cases for all these sample studies. We will do one-sample study, the two-sample study, or a parameter trial, yeah. Or we'll be the core study or the case control study, our (inaudible) survey, yeah? I am going to try to present here every one of these, like introduction for this one.

And when we have like a one-sample study, we are focused in the result of one group. That mean, only one population. Yeah? Okay. And the main outcome for the (inaudible) is in the theta, let's say, when theta is the parameter that we are interested. So this is the theta is the parameter represent, and the parameter we are going to use to answer the question for interest. So it could be like the proportion or could be the mean. And also the goal, we say before, could be estimation or hypothesis testing. So when we use estimation, I mean like we are going to trace like a confident interval. So we are not interested in a specific value of the parameter. So we don't know the value parameter, but we just want to know what is the possible range of value. So that is what I mean estimation.

But when we want to know if the parameter is equal or less or different the specific value, we use hypothesis testing. Yeah? So, here is the a graph wherein on this—I try to explain here the sample design for one sample study, yeah? So we have here like in this one sample study, we are going to assume that we have simple random sampling, yeah? For now. And the one simple random sampling mean that we have like a simple independent random sample is taken from a

population. So we are taking like a sample from a population. So we have the population and we select and subject an individual. For example is we want to estimate the proportion, so we need to classify this individual and that have the disease, who just have the disease, and compute the proportion—estimate the proportion. And we have the mean will be a similar case. Yeah, we have a population and we select a random sample of (inaudible). That's just to, like a little introduction for this.

I know that you are very familiar with this. For example, for one example estimation here will be like a medical officer, (inaudible) up the proportion of children who are obese in your school. He doesn't know anything about how will be the rate. You just want to know what are the possible range. So he will be interested into obtain like a confident interval. But in the second sample, it's like a hypothesis when the same medical officer wants to determine is the proportion of obese in the school is greater than 30 percent. So since we know we have a specific value, we are going to use hypothesis test. So, this is the possible two type of method that we are going to use. In chapter—I think it's chapter four will be all the methods for estimation. And with chapter five will be the method for testing hypothesis.

Okay, for the two sample studies, I include this in both in six sections because the statistical method will be the same for these two. One is you see the difference between both the two-sample study and the (inaudible) trial is that one

is (inaudible) study and the other is experimental trial. Experimental trial, I mean that the investigator control like in a treatment. So he decides who so you belong to group one, who one belong to group two. So the investigator controls who running this (inaudible). But in the two-sample study, (inaudible) study, we don't control, we just observe what we have, yeah? And the main outcome will be the difference between two population parameter. That could be the difference between two proportion or the difference between two means. Yeah? And the same, again, all will be (inaudible) the estimation or hypothesis testing.

Here, I tried to explain here what this two-sample problem is for the sample design. So we have two populations, yeah? In population one, we select a random sample of n-1 subject, yeah? And we compute the value of the parameter for that population. And in population two, also we select a random sample of n-2 subject, compute the parameter of interest where the parameter interest could be the proportion or could be the mean. But the important thing is how we select the sample. This is the important thing, yeah? And we decide that this is selected by simple random sampling. When we have, in chapter six, we are going to study like—there are two worldwide surveys, so we assume that this sample design is the sum of the random sample. The sample is not taken by random sample, random (inaudible), simple random sampling. Yeah. Okay. The thing is because we are going to assume in the first part of the training that is simple random sampling is because if we have another sample design, we just make adjustment for the sample size estimation. Yeah. Well, I think it's important

to understand first under the simple random sampling.

So we have two samples from each population, and we just estimate the difference, yeah? And for the case I am going to teach, I am going to assume that this sample size in each population are the same. But the method are very similar for when we have different sample sizes—I mean different sample sizes for population one and population two. I think I explained in one of the section, one example when do we have different sample size. So the approach will be similar. Okay, for experimental trial, here we have the population, yeah? And we select n-1 plus n-2 subject, and we decide here, this is where I am going to display here. We decide which one at random go to group one. Let's say to the (inaudible) like the control. And the n-2 subject, we select a random, go to group two. So here is where the investigator control so that we decide. He assign at random who's going to group one, he's going to go to group two.

So, in these experimental trial, we follow the individual and see if, for example, he develop the disease or not develop the disease. Yeah, but these are full of individuals. And also we compute, yeah, the parameter interests for the group one, compute the parameter interests for group two, and compare. This could be—we use this information to do like a hypothesis test or treat it like a confident interval. For that reason, yeah, we need to—this is the sample design. Is there any question here? No? Okay. Yeah?

UNKNOWN SPEAKER: What would be an example where you would have group 1 and group 2 preferentially exposed or not exposed to something that could be a potential risk factor for disease but yet would still be ethical?

RUBEN SMITH: That could be ethical, yeah.

UNKNOWN SPEAKER: Well, what I mean—I mean, are you actually—the experimenter is randomly assigning the subjects to group one or group two and one would be exposed and the other isn't.

RUBEN SMITH: Yeah.

UNKNOWN SPEAKER: And some treatment or intervention that the—

RUBEN SMITH: Yeah.

UNKNOWN SPEAKER: All right. Could this just be treatment of disease like with a new treatment?

RUBEN SMITH: Yeah, yeah. Exactly. Yeah, yeah. Could be like a standard treatment, a new treatment. Or it could be like one receive the placebo and this one receive the new treatment. Yeah.

UNKNOWN SPEAKER: Thank you.

RUBEN SMITH: You're welcome. Okay, so now, we're up for the core study. The different between the core study and the experimental trial is the--the experimental--the core study is a observational study and the other one is not. So, again, we are focusing in result for two group. Yeah, and also we follow the two group--the two group over time. And here, we--typical--is a prospective study. Yeah. So, the exposure is measured before develop with the disease, yeah? The main outcome that we can analyze the different or to rate--or population rate different, or we compute (inaudible). So, we'll depend with the investigator, which one. Again, the goal is estimation. The goal will be estimation hypothesis test. Okay, here, this graph will be very similar to the one before. The only different that will be that we don't--we don't control. So we observe. So we have one population and we have with what--these are group that have not exposed and the group two there are being exposed. And we select any one individual from group and any two individual for group two. And we follow to see or know if they develop the disease or not. Yeah. So we compute the proportion of the disease of group one, and we compute the proportion of disease of group two and compare. So compute the parameter of interest. Yeah. And this can--we estimate. Also, okay, we--another study that we--we're going to work is the case control study. This a little--this is different because--also we are--this is our retrospective study. So, they need to understand--well, the different before the core study and the case control study, because in the--case control study, the

exposure is measured after the disease, yeah? So, we are interested--and also, the parameter of interest will be the outer-ratio because this is a retrospective study, yeah? And also, we are going to have the--the goal will be estimation or hypothesis test, yeah? We are going to try to explain the graph here. Okay, so we have a population of interest, and we are going to have like a control group and the case group. So, the control group doesn't have the disease. But the case group, this is the one that have the disease. And that's the reason how it's an observational study, yeah? So we select a one subject from the control group and then two subject from the case group, the case group--and again, we select--we check if we--this individual have--having exposed or not. There's a difference. So, we check if this is being exposed or no. So, after that we check, we compute the proportion of exposed among the control group and the proportion of exposed among the case group. So we had the proportion one, the proportion two. And with this information, we compute the outer-ratio. The--your rate--the outer-ratio is using this formula. Yeah. So, this a--everybody can see the different between this two, the core--the core study and the case control study, yeah? Okay. Now, the another study we are going to work is the--a survey study. In a survey study, we don't have any follow-up of individual. So you're--you'll just see what happen in a specific moment of time. This is the one, we have our survey. Yeah, and the general focus of survey will be the point estimation. So we got one interesting--we're going to obtain more like a point estimate or like I said before, like we're interesting to create like a confident interval, yeah, of the characteristic of the population. Usually--or no, in general--let's say, general hypothesis testing

is often a secondary objective. But when we compute sample size, usually we are more--we use formula to get confident interval in this study. And the main angle of interest could be the--could be that are now the proportion, could be the mean, or could be the different of proportion, that we can see this like a--this is like another parameter, yeah? About--we're going to see later. Or different of mean or the outer-ratio. But usually, when we use--when we--when we work, we want to estimate the sample size for survey, we're going to call in this training, it's more like a--when we have a goal of proportion, so we want to estimate like a proportion. So, we'll use the formula for proportion to get like a confident interval for the proportion. And for the reason I say that here, our the goal for survey is estimation. And the sampling method, what--the sampling method--usually is no simple random sampling but it could be. But it's no usual to have a simple random sampling on survey. Or it could be systemic also. It could be, but it's no general. But, usually, we use more complicated design like I said before. Like Brian was talking up before, a stratified sampling, a cluster sampling, or a complex sample design, the second--the complex sample design is any sampling that is no simple random sampling. So like, we can say that a stratify sampling and a cluster sampling is like a complex sample design. Yeah, because of--usually, complex sample design where we have a sort of combination of stratify and cluster, like the example that Brian was presenting before. Okay, here. So depending of what we want, we need to decide what method of the dominant--of the sample size we want to use. So it will depend--will depend on the interest that we want. So, we are interesting in--is parameter estimation, we'll use the

confident interval approach. Yeah. So we are interested in only one parameter. It could be like the mean, the proportion, or a proportion of parameter. But always thinking in that way that is a--it's like a (inaudible) like the data we're going to produce is like a confident interval. So we base our sample size in--determine the precision of the parameter. Or it could be this--also, in the interval approach, when we are more interesting in hypothesis testing, we'll use the statistical analysis approach. So--so we need one before to decide which one. What we want, we need to check what's technique we're going to use. So--yeah?

UNKNOWN SPEAKER: (Inaudible) this is a good point to say something here. Even--there are studies when you had both or even more, or you have four. You have a June--do you want to know only to estimate one proportion, one mean, but do you want one proportion or two proportions two means or also beyond finding the prevalent, you want also to do--test of hypothesis. So the question is--okay, now I have a larger problem to what do I do. You have to compute all of them and then the larger sample studies through. But then, this is a good point because he is clarifying exactly. Okay, you may have many options. Usually, it's more than one. We are going to go one by one, but you have to keep in mind, every time, in real life, you have more, so you have to do all and then the largest one is the one that--go with.

RUBEN SMITH: Yeah, the same happened here when we have different variable and interest when we bring our survey, you select for all of them, or for the most

important and select the largest one. Okay. So, in summary. So, this is the one--I want to call your attention that this sample size formula depend or the statistical method--there are planned for data analysis, so that mean it's estimation or it's hypothesis test. Or also will depend--the formula will depend on the study design and also will depend on the parameter that we are estimating. So, always need to think about what we are going to--to have to decide in what formula we are going to use. I think I need to go for the next one. So--so, in chapter four, all we're going to see here is about this--the sample size determination using the confident interval approach. And the chapters five, they're going to--be teach for everyone will be about hypothesis test. So, for now in this chapter, we need to think that all interest is in--we want to obtain a confident interval, yeah? So, okay here, (inaudible) line for the chapter. So, in the--I am going to give like a introduction for this chapter. And also after that, we are going to study--we are going to specify the all formula for all this study design that we just--I just present. Okay, like here--like the introduction, yeah? So, again, the interest will be the--our parameters. So we're interest in our parameter and we want to obtain some--like a confident interval for that parameter. So, in the confident interval approach, we are focused is in precision. So, precision mean, this some to mean--precision mean in turn with the sampling error. So, because the focus in precision with the sample estimate that, let's say you had a--which is a measure by the wide or the confident interval. So when we are (inaudible), what we mean precision mean like we are producing this graph in the left one, we may here--like we have very high precision here. So with the--the red dot are very close to the target. The

target is the black--the black dot, black dot.

UNKNOWN SPEAKER: Wait a minute. How do you get the width of the confidence interval? Is that a, somewhat like a judgment call saying, “What is a wide confidence interval?”

RUBEN SMITH: Yeah, yeah.

UNKNOWN SPEAKER: So you’re--you know, is there a way to test--is it too wide? Or is that something that you look at based on, you know, what the mean might be, what you expect it might be or, you know--and maybe you’ll talk about that more while were looking through this.

RUBEN SMITH: Yeah, do you mean one--the--that mean--the result will depend of the investigator. What is the true, let’s say, significant different or the quality true significant different that is mean for the area of the research, for the investigator. So it’s not statistical definition. It’s more--like because for some area, like for example--like a different for example, them .5 maybe make sense, but maybe for other area, .25 is better. I mean--so it will depend of the investigator what the one--what is meaningful. Yeah. Then I will go.

UNKNOWN SPEAKER: One example. Let’s apply it to real life. Let’s put it in the context now that we are talking about the epidemic of the obesity. Let’s suppose

that we are all interested in gaining to weight, diet here,, and then we make sure our circumferences and then we will say, “Okay, for me, that”--yes. This is interesting because this is why that we are in this--each one can apply to yourself. So, if I said that is an inch difference, that the treatment that I take after--I don’t know--how many weeks. We can’t decide that to whatever. Is that inch meaningful? Or is two inches meaningful? Or is three inches meaningful? Or is four inches meaningful? So, whatever, it depends of course of the questions on what it is the population that I am talking about. So that is what he is playing at, it is subjective that is because it’s based in some of the problem and the population that we are talking. So, it is what is meaningful. That’s the key. What will be important? Am I happy if I have one percent or two percent? Or could be 10 percent enough? To say it did decrease by 10 percent whatever, would that be meaningful? Or in--in that type of example is what you have to think, what it will be the what? The width of the confidence seems about, it depends on this situation. But it is, if you put it in yourself, okay, either reduce the (inaudible) outcomes by 10 percent, would that be a huge reduction? Or I am thinking maybe, one percent will be enough. Okay? Something like that?

RUBEN SMITH: Yeah. And the important thing is that you have more high precision, you will need more high sample size. So, usually you will play a little bit. When you compute sample size, you need to play a little bit with this different precision to see which--and follow which makes sense for you but also if you have resources to do that kind of sample size. Also, you can compute the same

formula that we are going to present you. To compute the sample size, you can use before you do this, the experiment, compute the precision that you will have. So, this is--this is--this will be--this design. So, in this--in the first one, we have like a very high precision. I am right here on (inaudible) because we are in the target. So this is what we want. We want high precision and in the target. From--in the second one, you have like a--let's say here, no precision because you are you--your target is the black dot. But you--I don't know--the precision is very, very--you have more random error here than here, yeah? And here, you have precision but you are out the target. So, you don't want this. What you want is more this. So when--when you have this, for example, like a--in a study, for example, if you have a lot of--no respondent, maybe because when you compute the sample size, you—the assumption that you estimate or the (inaudible) for this parameter is unbiased. I mean in other words, this value turned to go to is equal to the parameter. But sometime when you have no respondent, you are not taking sample of all your subjects of interest. You're taking some or part of them.

So you can be biased, yeah, for this. So you want to use this (inaudible) for the reason not to (inaudible). To reduce the no response is very important because if you have a very high respondent, maybe you will have like a very biased and also no precision because--the no precision because you reduce the sample size. By then you can have high sample size, but the respondent are biased because some part of the population are not included in the sample. Okay, so this is the sampling error. So he said he had the precision is like a measure for

the sampling error because we are going to use the sampling error to compute the sample size. So the sampling error is introduced—do it to sampling. So because the sampling error, always we are going to have sampling error. We don't have sampling error, always we have census. If we have census, we have zero sampling error. What we want is to try to reduce the sampling error, yeah? For that reason, the sampling error is the difference between the sample estimate and the true value.

So this is what we wanted. So I want to explain here what the precision mean. So this is the difference. This difference is the difference between the precision. So the smaller this distance is, the more precision do you have and the smaller sampling error do you have. So let's try to talk about the confidence interval, a little more like more formula here. So these are the typical confidence interval for the parameter theta. So let's assume that the estimate, this is a little hard because it's like a estimate. You see the sample, yeah? (Inaudible) is normal, yeah? When the expectation is theta and the variance is this one because we're assuming that the variant is known. So we can use here the quantile from the C distribution, the normal distribution. If we assume we all know this value, we need to use here the (inaudible), yeah?

For that reason why, here we are going to assume that the variant is known because if the variant is not known, we need to use another formula. I mean in fact we need to use like a (inaudible) to do it because it's an alternative. But if we

assume that the variance is known, we can use computer formula directly by using a simple calculator. You are very familiar with this confident interval. So are you here? I just play a little bit with the formula. We can see here that here, I can write the confidence interval in this form. So remember here, this is the difference. This part is different between the sample estimate and the true value. So here we can see that the precision is this part. So we are going to use this information to compute the sample size, yeah? And so the precision, remember, is the difference between the sample estimate and the true value, yeah? So the sampling error is usually report as a plus or minus the precision, so that's the one that we used before. So the sampling error is plus or minus the precision. And the precision measure how close or estimates is to the population value.

So when we specify the precision, one was using D. I used X, you know? We are specifying this distance, or the distance between the estimate and the quantile multiplied by the standard error, yeah? Okay. So the precision, as I say before, is a major of how close our sample estimate is to the population value. So at the precision, there are different formula to specify the precision and the (inaudible) will specify the precision. It will specify the precision an absolute value, so we use specifying the distance between our estimate, our parameter. And we want that this difference be less than a specific value. Or another way will be (inaudible) to the true value. That means the difference in terms of the true value, the percentage. So there are two different formula dependent on how we have specified the precision. So the precision, again here, sampling error our precision

is like. So precision is related to the confidence interval. So the precision is equal to the quantile multiplied by the standard error. What I mean is the precision is the half weight of the confident interval where the c value, yeah, is that the value of the quantile for the normal distribution. And if we use alpha as 0.5, that mean that the quantile is 1.96.

Yeah, this is very familiar for you. This is like the (inaudible), yeah. Okay, the precision, yeah? So we are talking about precision on 95 confidence level, we are talking about the precision is approximately two times the standard error. We are talking about precision, a 68 confidence level, we are talking about only one standard error. So another way to write, again, the confident interval is what? The sample estimate plus or minus the standard precision, yeah? Okay, so when we talk about a sample size, we still talk about confident level alpha. So the confident level alpha measure the probability that the true value, theta, is within the expected precision of the sample estimate. So I mean when we talk about a confident level, they talk about the probability of or sampling error is less than our precision. Yeah?

So the parameter that we need to compute the required sample size are what are the maximum sampling errors that can be (inaudible). So these are the parameter we usually get to compute sample size. So the larger your sampling error, the less precision the estimate is. Yeah? And also, another parameter we need to introduce to give or compute sample size is the probability like sampling

error is less than the precision. So usually it's set to 0.95, that means alpha 0.05. So what we are talking about when we get the alpha, we are talking about what is the probability or sampling error is less than (inaudible). Yeah?

UNKNOWN SPEAKER: I'm going to--

RUBEN SMITH: Yeah.

UNKNOWN SPEAKER: (Inaudible)...

RUBEN SMITH: Okay.

UNKNOWN SPEAKER: So we set it at 0.95, there's a five percent probability that the sample estimate is outside the confidence interval.

RUBEN SMITH: Okay.

UNKNOWN SPEAKER: I'm trying to put this into words (inaudible).

RUBEN SMITH: Yeah. And good. Okay, the mean is like when you say here, because you don't know the value of the parameter.

UNKNOWN SPEAKER: Right.

RUBEN SMITH: All you are going to talk here, okay.

UNKNOWN SPEAKER: But that's what I want to know. That when I get an estimate, I want to know how likely it is to be to represent the true value.

RUBEN SMITH: Okay.

UNKNOWN SPEAKER: Independent of bias.

RUBEN SMITH: Okay, let me, for example, if suppose like you only have take one sample, but suppose that you take 100 sample.

UNKNOWN SPEAKER: Yeah.

RUBEN SMITH: That mean, from the 100 sample, in 95 you will have--the parameter will be inside of the confident interval. But in five, the sample will not.

UNKNOWN SPEAKER: Right.

RUBEN SMITH: Yeah.

UNKNOWN SPEAKER: I'm trying to go back to this previous slide where—

RUBEN SMITH: What number?

UNKNOWN SPEAKER: This slide 4-6. So precision will be that estimate might your—the epsilon is—

RUBEN SMITH: What's the difference? What do you want?

UNKNOWN SPEAKER: Is the difference between the sample estimate and the true estimate.

UNKNOWN SPEAKER: (inaudible)

UNKNOWN SPEAKER: And we don't know the true estimate.

RUBEN SMITH: Yeah.

UNKNOWN SPEAKER: Like this one, right. So back here, then you got epsilon from—

RUBEN SMITH: What number?

UNKNOWN SPEAKER: (Inaudible) to 10? If you want to go to 10?

RUBEN SMITH: To 10?

UNKNOWN SPEAKER: Yeah. Okay.

UNKNOWN SPEAKER: Back to nine, I think. Nine.

RUBEN SMITH: Nine.

UNKNOWN SPEAKER: I know 10, so the probability that the error is greater than epsilon—

UNKNOWN SPEAKER: Now there's something that is lesser.

UNKNOWN SPEAKER: Yes, is less than (inaudible)—

RUBEN SMITH: I'm sorry (inaudible)—

UNKNOWN SPEAKER: (Inaudible)

UNKNOWN SPEAKER: Is five percent.

RUBEN SMITH: Okay, this one.

UNKNOWN SPEAKER: (inaudible)

UNKNOWN SPEAKER: So what do you mean by sampling error? And the probability is—

RUBEN SMITH: Oh, remember the sampling error is the difference between the true parameter that we don't know—

UNKNOWN SPEAKER: Yeah.

RUBEN SMITH: --and the estimate. So the thing is you want the difference—or maybe here that will help you.

UNKNOWN SPEAKER: Yeah.

RUBEN SMITH: So that difference—

UNKNOWN SPEAKER: Uh-huh.

RUBEN SMITH: --will be less than the precision. Let's say, for example, 0.05, yeah? So how about the mean, the difference between the true parameter, and your estimate will be less than 0.05. So that means that (inaudible) the true

parameter is 0.20.

UNKNOWN SPEAKER: Mm-hmm.

RUBEN SMITH: And you have set precision 0.05. That means your estimate will be like it will be between 0.15 and 0.25.

UNKNOWN SPEAKER: Mm-hmm.

RUBEN SMITH: So you will be happy with that range.

UNKNOWN SPEAKER: So how does that relate to the width and the confidence (inaudible)?

RUBEN SMITH: Oh, in this way, okay. Here is the confident interval. Remember this confident interval is here. This difference, you want not to be greater than 0.05. So when you have specified the precision, you are talking about is what is the value of this one. So you will be happy that the estimate and the true one will be less than w. Do you see why you can just specify the probability 0.95. So that is no guarantee will be like our probability less than one but will be 0.95, you use alpha 0.05.

UNKNOWN SPEAKER: Let me rephrase. Maybe I think I am understanding

some different direction. What you're asking is will the 95 percent confidence interval means that I have five percent error with that to find what I want? And the answer is no.

UNKNOWN SPEAKER: Mm-hmm.

UNKNOWN SPEAKER: We are talking about a 95 percent confidence interval, whatever type of confidence interval is that, you are assuming when you took the sample that there was an error in sampling just because you knew that you are not going to be able to tackle the whole population. You are just selecting a subset of the population, so you are doing an error just by not selecting the entire thing. Okay? Now with that confidence interval, either your confidence interval have or do not have the true? So it's yes or no. Did it have it or didn't it have it? But that doesn't mean that you are 95 percent certain that the true is there. Is that is incorrect to think? It's just the reasoning is when you took the sample, you assume the maximum error that I am going to allow. That I am going to permit to compute my estimate, I expect it to be this percentage.

RUBEN SMITH: You are taking one. You wish that the one that you taken is the one from the 95 that will have your parameter. It's possible that will be—that will not contained by you. I know you are hoping that that one is from the 95, from 100 that will contain your parameter. Do you know what I mean? Okay, is there another question? Did we answer your question, no?

UNKNOWN SPEAKER: Yes.

RUBEN SMITH: Okay. Okay. Okay, we are here. Oh okay. So that means, again, when we specify that one, we are specifying that the sampling error, this thing is more than what we want. So more precision, smaller sampling error, yeah? So also another parameter that we need to give to the formula is estimated by the ability with respect to the characteristic of interest. So we need to know, I need to know, or at least, to have a guess on the variance, maybe from a previous study or from reference, but we need that parameter. For example, for the proportion, that means, for example, we need to specify that what is the proportion that we expect. For example, for the maximum variability, we cannot say for the proportion is when we have like 0.5. So it's more variable than we have population to have like a proportion of 0.90 because here is less variable because we have in the group one, for example, for this one, for the proportion 0.90, 90 percent are in group one and 10 percent are in group two. So these population are more similar for the group one. Yeah? And the other one. So for the means, we want to estimate the mean, we need to specify the variance. So another parameter that we can use to estimate the population size if we have it. Yeah, so we have the population size. We can use it. Yeah? So we have these four parameter that we need to compute the sample size. In case of the confident interval approach, in the hypothesis testing we need to specify the power. Remember this case is only for the sample size for the confident interval

approach. Okay, let’s say this is a sample as I try to help here. So let’s suppose one example that we want to estimate the population proportion, yeah? This sample size, I’ll repeat here, depend on the sampling error by representing for the epsilon, yeah? And the alpha level, yeah, for alpha, and the anticipate population proportion, and also the population size. So these are the four parameter that we need for this example, for this conclusion, for computing the proportion. So in this graph, I tried to represent the relationship between the sample size and the sampling error by the proportion. For example, for this and this graph, in the y, x we have the sample size, yeah? And the x axis, we have the sampling error, yeah? So this is computed by using alpha of 0.05 and a population size of 50,000, for example. So here we can see. This graph would show how the sampling error decrease, the sample size increase, yeah? So this is what we’ve been talking before. But we can see here in the graph, yeah? And also, we can see here in this same graph how if the variability increase, the sample size increase. So you can see here that if we have more—the proportion most dissimilarity of the population, the sample size will increase. For example, for 50 percent of the proportion, we can see always in the top because here, the population is more variable. That means this is more different. So here in this graph, we can see that the information. Also in this graph, we can see the relationship between the sample size and the sample error, but also by the alpha level. Okay, here we can see that the sample size, again, can expect the sampling error. The sampling error decrease, the sample size increase, but also here this is the new thing, the new one in the graph. How the alpha level

decrease, the sample size increase, where we can see that. If for lower alpha level, the sample size are more higher. Yeah? And this is for the proportion of 15 percent. Now in this graph, I try to explain here the relationship between the sample size and the population size. So here we can see that how the population size increase the sample size also increase. Here we can see this, yeah. I also, again, have considered here the sample size will be bigger when we have the proportion. The true proportion, we assume, that's 50 percent. It's the largest one, yeah? This one will be a continuation. See, this is the same graph. I now I tried to increase more the population size. So here, we can see. I plot here. I graph here the population size from one from one million. So we can see that the population size always increase, how the population size increase. But here we can see for the case of the proportion that the sample size doesn't increase much after this population size is more than 100,000. Yeah? So we can see that almost it's constant. So doesn't matter how big is the population is, the sample size doesn't increase. So that is the reason in many national survey or state survey, the sample size for the proportion is based in about 1,100. So it doesn't make sense to increase more because you are no—you don't gain precision. Yeah? And this was computed for the sampling error of three percent. So the lessoning--I'll trying to explain here again. So that means that the different for each sample size, we have been happy when the difference between the sample proportion and the true proportion is equal to three percent or less than three percent. So we don't expect that will be bigger than that because this was the one we use. Okay, this is our exercise, so these are for you. So the exercise, so

what happened to the required sample size are the sample error increase, holding the other parameter constant. So what will be the answer?

UNKNOWN SPEAKER: (Inaudible).

RUBEN SMITH: So what that is, you compare your sample size with error. So is this sample size error increase? So you'll have more sampling error, what happened with the sample size? Do you expect to have less sample size or more sample size?

UNKNOWN SPEAKER: (Inaudible)

RUBEN SMITH: So did you—

UNKNOWN SPEAKER: (inaudible)

RUBEN SMITH: Huh?

UNKNOWN SPEAKER: A or A or (inaudible)—

RUBEN SMITH: A, A (inaudible). So this is the answer, that is the solution (inaudible) A, B or C.

UNKNOWN SPEAKER: (inaudible)

RUBEN SMITH: B? Okay. You are right. So it will be B. Yeah? So well, now the other one. So what happened to the required sample size as the level alpha decrease, holding the other parameter constant? So the answer will be A, the required sample size increase, B, the required sample size decrease, or there is no association between the required sample size and sampling error.

UNKNOWN SPEAKER: (Inaudible) A.

RUBEN SMITH: A? Yeah. It will be A. Very good. So you are awake. Okay, so now I am going to go more detail. Try to explain the study, the sample size formula where we had the one sample situation. So we are interesting in estimate the population parameter. In this case, will be the proportion. So we are interesting to estimate the population proportion. So there are two types of formula. I am going to try (inaudible) in this class. So why when we specify this (inaudible) position, I mean that the difference between the sample estimate and the true one is less or equal to the precision. Or we specify at the other position, I mean that the difference is (inaudible) or the percentage of the true value. So yeah?

So in this one, we want it like a point difference. I mean like the difference will be like difference between these two in points. By this one, we'll be like a point

percentage. In the second one, the difference will be like in percentage with respect to the true parameter. I am going to try to explain this very big example. So in this one and the other case that we are going to start is also the population mean. We specify the difference between the sample and the true parameter, the true mean. So we are going to start the case one, the (inaudible) and the (inaudible). I mean the variant is no and the variant is a no. (inaudible) Okay, the first case is debating the proportion, (inaudible) precision. So I'm try to introduce the section with example.

So we want to estimate the proportion in a large, let's say, infinite population with 10 percent point of the true proportion with 95 percent confidence. So that means is it doesn't mean that we want is that the difference between the sample proportion and the true proportion is less than 0.10 percent, being the difference between 0.10. This is what we called specify precision. And we are using alpha level of 0.05. So this is what we want. So this is what we mean, maybe that will answer your question. So what we want our sample size such at that the difference between the sample and the true one will be between 10 and minus 10. So would you be happy with that? So you don't want your estimate to be greater than one, than 0.10. So that means in terms of absolute value, that means that the difference between the sample and the true proportion will be less than 0.10 equal to 0.95 percent. So there is no 100 percent guarantee, but you have 95 percent guarantee that your sample estimate will be no greater than 0.10 as a true value.

So what are the assumptions for this case? The assumption will be that a simple random sample is taken from a large population. This is important, like I say here. When I mean large population, I mean that is, let's say, larger than 20,000, let's say, because after that you will know you will not see much difference. But for now, I think I'm sorry I'm not going to assume, but the sample size are very large. So we are going to assume that the population is very large that we don't need to make a—we don't need to use the sample size, the sample of the population size.

If we need, we can make some adjustment. That (inaudible) that one we call the final population correction where I'm going to present in chapter--in section six that we start to use final population correction and when we use for survey. And we are going to present that case. So in this situation, we have that each sample subject can be classified according to two categories. I mean that the event is present or the event is not present. So I mean that have the disease or doesn't have the disease. So here, the statistical, mathematical (inaudible) point. So we assume that each individual have a distribution aspect normally. These are the assumptions. And we--and the parameter--the probability to observe the event is pi. And the probability of not observe the event will be one minus pi. So this is what--this is what we have. So the sampling distribution, or the sample proportion, is approximately normal. Yeah?

I call it approximately normal if it’s (inaudible) large. So this formula that we are going to present, assume, yeah? That we have a large sample size, yeah? Okay. And this is normal and that the normal the mean will be equal to the proportion and the variance will be this one. So you are very familiar with this, yeah? Yeah? Okay, so the confident interval for the proportion is this one. So this is why did you know. Yeah? And the c is the quantile for the normal distribution. Also in the note in chapter eight, there is a table for the normal quantile. So here, because remember, we are going—this is why where you call this the confident interval approach because we are going to use the confident interval to compute the sample size. So we have the confident interval. And here is the one we call the precision. So we want the precision is equal to this one and only (inaudible). So you can see here, the sample size only depends of (inaudible) factor. That’s independent for this one.

So for that reason, the precision is the estimate to be—I’m sorry—the precision, this definition of the precision or estimate are level alpha. So that means that we want the precision and (inaudible) of the number of percentage point, yeah? So that means that the difference between the two--I am repeating this one here. All right, this is repeated. Okay, this is the confident interval. And here, the confident interval (inaudible). Okay, there is. Okay, yeah. So here is the precision, so we need only to solve for the sample size. So if we solve for the sample size, the width we obtain, the formula for the sample size. We can see a sample size is this one. It’s a function of the quantile (inaudible) distribution, is a function with

the precision and is a function of the proportion.

So to use the formula, we need to assume some value for the proportion. This is maybe we think well, how are we going to know this one (inaudible) we want to estimate. But I think that we need some guess to use the formula. So here is the number of the formula where I am going to present. So here is the example. Oh, when I write here, when you find here, what you see here, the red in the top, it's because I made some corrections with this light. I think I modified this one. So, in this example, yeah, a medical officer serves to estimate the proportional children who are obese in that area. Yeah. So the question is how many children must be studied and the resulting estimate is to fall within 10 percent point of the true proportion with 95 percent confidence.

So, the investigator, these are questions that the investigator need to know. This is no good questions for the statistician. These—yeah, so you—the investigator need to know why is this true, I mean, what is the proportion that he is suspecting in the population. I know maybe sometimes it's no easy question but is something that we need. So the proportion for this case, meant for previous tally, for example, but the proportion of obese children, is expected to be about 30 percent or close to 30 percent?

So, the solution for this example is, okay, the anticipate population proportion is .30, just so the precision, as I've said before, is .10. That mean that I was specify

this equal to .10. That means that we're happy. We're estimate that will be between 20 percent and 40 percent because this is 10 percent below the 30 and 10 percent above 30. So, this is—this one, we specify this, we ask to tell him the formula while (inaudible) what are the possible value for all estimate in this event. So, with the probability of 95 percent, so we use the formula (inaudible) the formula here. So for the C value will be for (inaudible) equal to .05 will be 1.96, yeah, divide by the precision that is .10, square this one and the expected proportion is—oh, I'm sorry. The expected proportion is .30, multiply by 1 minus .30, will be 80.66. So that mean that a sample size of 81 children will be needed for this example. Does anybody have any question here? No?

Okay, now, suppose that we have the same example. But the thing is we don't have the same problem before but we—it's not possible to make any assumption regarding the proportion of these children. So we don't know what is the possible value of this. So, in that case, we need this relationship. So, this is—the relationship between the sample size and the population proportion. So, here, we can see here, in the X, we can see here the all possible value for the proportion but in between 0 and 100 percent, and this is the sample size. So here we can see that the sample size increase how the population proportion increase but when is point—when this 50 percent is 30 degrees. So that mean that the maximum sample size is observed when the proportion is 50 percent. So, it is not possible to estimate it, the proportion from previous tally. The safest choice is to use pi equal to 0.5, since the sample size will be the largest.

So, if we use this information and the same done before, only what we need to change the value of pi is 50 percent, the same precision, that mean now, now we are talking about that we are going to be happy, yeah, if the precision will be between 40 percent and 60 percent. Yeah. And the confident level will be (inaudible) that mean that the sample size will be larger. That will be 97. So a sample size of 97 children will be needed.

So this is a exercise for you. So a local health department want to estimate the prevalence for tuberculosis among children in the five-year and this is (inaudible) in one locality. The question is how many children should be included in the sample so that the prevalence may be estimated to be five percent of the true value, with 95 percent confidence? Is this (inaudible) that the true rate is a likely to be greater than 20 percent. So what would be the solution for this problem? Will be A. 245, B. 240, or C. 246? Huh?

UNKNOWN SPEAKER: It’s greater than 240.

RUBEN SMITH: Greater than 240? Because?

UNKNOWN SPEAKER: Because all of them are greater than 240.

RUBEN SMITH: All right.

UNKNOWN SPEAKER: You want the calculation.

RUBEN SMITH: Yeah, unlikely. Let's see. Let's see. So what do you think? I agree with Adriana? Huh? What will it be?

UNKNOWN SPEAKER: Two forty-six.

RUBEN SMITH: Two forty-six? Okay. Oh, I'm sorry. Okay, the answer is 246. So here is the solution. So the anticipated population proportion will be 20 percent, just so the precision is .05. So the—I'm sorry, the precision would be, yeah, .05. The confident level will be .05. We use the formula, yeah, before. It will give us 246. So a sample size of 246 children will be included in this sample.

Remember, something is important here. So, here, this is the sample size that we—we compute sample size. This is the sample size for respondent. So what happened with no respondent? So what (inaudible), we expect, like for example, this is assuming that everybody is going to answer. So, if people and the respondent, the precision we saw before in the graph, it would go down. So one way to do is to (inaudible) do is to increase the sample size for a percentage that we expect that will be respondent.

UNKNOWN SPEAKER: Well, actually, they (inaudible).

RUBEN SMITH: Yeah, we're done with the (inaudible) but for examples some people used 10 percent. But there are no rules because in your case, you have, like, you say 50 percent of no response. So it will depend on the problem of the specific area.

UNKNOWN SPEAKER: Let's go to the real life. Let's suppose that we're going to the (inaudible) in rural counties. So we know that you have to go into the field and this is a reportable disease, everybody is (inaudible) and they figure out they have to go, they have to go get some treatment, right? But you have to go to the family members. So the question is how do you test for tuberculosis? And the children, you have to make the test and then, would the parents agree to allow you to make the test to the children? What if the response rate in each one of your states, you know your people? So what is the percentage, that's where you have to start thinking. Okay, maybe yy would say the people is rarely, rarely the—as soon as you get there and you explain what is the situation, they say, "Yes, I want to be tested." I would have—the response rate is probably 70 percent, 80 percent, 90 percent. Sometimes, the price, they don't want you (inaudible) when did you go? Did you go during the weekends? Or you're going to plan these during the evenings? When are you going to make those tests to try to pick up the children after 4:00 p.m.? That's thinking something. So, what is the proportion of people that I'm going to find at 4:00 p.m. during the weekdays to try to make it feasible? They have to get to 246 after all these tests. So in each one

of your (inaudible) you would have a different percentage and that what you're saying, "Okay, forget it." If I don't have time—houses—if I find four people in the house at 4:30 p.m., let's suppose, that would be good. So I have to adjust and overestimate by whatever percentage that I am just thinking. Does it make sense?

RUBEN SMITH: Thank you. Okay, now, we are going to another point. So we now--we still want to estimate the proportion but we want to use a specific relative precision. These are different. Before, we were only measuring the sampling ruling between the different between the two proportion, but now we want the different in terms--or the true parameters that we don't know, but it's possible to compute that formula. So the relative precision is—the precision is the percentage of the estimated value.

So, in this example, estimated proportion in a large—I mean, in (inaudible) population, with 10 percent, the 10 percent of the true proportion, that mean not—no 10 percent on points, no—10 percentage of the true value, with 95 percent confident. Okay, that mean that the difference, I mean, that the different between the estimate and the true one will be less or equal to .10 percent of the true one. (Inaudible) and maybe the example will be much better. But we think—the different between 10 percent of the true parameters. So, in this case, it will be 10 percent with respect of the true value. That mean, for example, if the—I need a calculator here. So let's say they have sample with a different. So the

assumption will be the same. So, here, there is a typo here. I correct it. I think (inaudible) larger. So, here, the assumption are the same than before. The (inaudible) the proportion so we don't need to see the real. So on the sample distribution also would be the same. So nothing new here. But now, the different would be here. Let's see here. So the confident interval is—you're familiar with this one. But now, you wonder—remember, the formula that you want to have—you want to divide the confident interval divided by theta, by 30 times divide the proportion so that you have the term, you want to remember that this is the precision. I mean, the precision for this one. So, this one.

So, if you divide by pi, you would have the parameter that you're interested and that mean that now, the precision is different than before because you have one minus P, divide P. So, because I divide in both side, I divide by P here and divide by pi here, I have now the precision with this one. So, the precision—this is the precision is pressed as a percentage of the pi. Let me state the example here. Okay, this will be the formula for the, for the precision and use (inaudible) for N is the same as before and now the formula is this one. It's different because, remember, this part is different. I think I (inaudible) here. Before we had, like—before, we have here, pi multiply one minus pi. Now we have, one minus pi, divided by pi. Okay, here, we can see the relationship here. The relationship between the sample size and the precision is different than before because, now, we can see here, in the graph, that the—how the proportion decrease, the sample size increase. Before was different. Before, the sample size increase,

how the proportion was getting closer to 0.5. But now, it doesn't matter. Now, only this is—the proportion decrease, the sample size increase. So it's a little different here. So, the secret to (inaudible) this thing in this example. The same example done before, but now, remember, before we have, I think it was 10 percent point, now we have 10 percent different from the true value. Now, let's say that the proportion of obese children is expected to be 30 percent, so I would read the question again. So the question is how many children must be studied if the resulting estimate is to fall within 10 percent of the true proportion, with 95 percent confident?

So here is the different here. So, this is the solution. So the anticipated population proportion is .30, but now the (inaudible) precision is .10. So now, we are going to be happy. It is 10 percent below this value and 10 percent above this value. So it's different than before. So now we are going to be happy if it's 24—between 20 percent and 33 percent. So, and the confident level is 95 percent. So, the sample size, using the formula, will be now 197. So, the sample size is much, much bigger because we have more precision. Before (inaudible) sample size, 81? Eighty-one. So, because here is different so we are here—we want here more precision here. Is there any questions here? No question. Okay, listen, this is a homework for you. So here, the investigator said to estimate the proportion of a woman in a population who seek prenatal care within her first trimester. The question is how many women should be included in the sample so that the proportion may be estimated with five percent of the true value, with 95

percent confidence? So, it is estimated by a proportion of woman who receives such a care would be between 25 and 30, and now I'm giving, like, a range. We now that this in this range. So, what would be the answer? Will it be A or will it be B? Try to see the graph I showed before to see what it will be. The sample size is compute one is for 25 percent, another is compute by 30 percent. So, what would be the biggest sample size?

UNKNOWN SPEAKER: (Inaudible) because if the portion of women is equal to 25 percent, (inaudible) but if the proportion is apropriate, then the smaller size would be (inaudible) but you don't know, so you should go for the largest.

RUBEN SMITH: Yeah, that's right. Thank you. Yeah, this is—we always go to the largest one. So here, the solution. So, remember, the anticipated population proportion is, we said before, between 25 percent and 30 percent. And the relative precision is five percent. So, if we assume 25 percent or estimate will be between 24 percent and 26 percent is—we assume 30 percent or estimate will be between 28 and 32. So this is wider, yeah. So if we use confident level—I'm sorry. If we use confident level, 95 percent, yeah, alpha (inaudible) .05, the C value would be 196. (Inaudible).

So, if we use this, the formula, yeah, for proportion, we'll have 25 percent, we have 4,310 for P equal 30 percent, the sample size will be 3,485. So, we're selecting the biggest one, so the sample size will be 400--4,610 will now be

needed to satisfy this object. There is any question here?

Okay, now we have—now, okay you are done with the proportion for one sample study. Now, we're going to go—we want to estimate the population mean, so the parameter of interest now is a population mean. That mean that the response is continuous. Yeah? So, in this example, the objective is estimating mean in a large population also. We have 10 unit point of the true mean, yeah. So with 95 percent confident. So again, so remember what we saw before now is all estimate minus the population parameter that's nill will be less than 10. Remember, the same logic before, so now we change--now, instead of proportion, we change to the mean. So, here, so now the outcome of interest is continuous. And again, the sample is taking—using a simple random sampling here and the assumption is the sample is taking from a normal distribution with a unknown mean and the variant is sigma squared. We are going to start the case where sigma is the--see the variant is unknown and is also known? You would see that there is no different. So here, the statistical assumption about the normality. So, remember, we assume that every subject is independent for everybody, yeah.

So, the sample distribution or the sample mean is normal, yeah, with the knew--with the expectation will be knew by the variance, will be sigma squared divided by N. And this is the variance of the (inaudible). So, if we don't have, like, a (inaudible)—we don't know about the—the population is coming from a normal

distribution but we know that for the (inaudible) that the sampling distribution always, or the mean will be normal if the sample size is large. How large? Well, it would depend on the variable, but usually, it would take as 30.

So, again, here, I am going to assume that the sigma square is known. So we know this variance. So this—so we're going to study is divide this case, we are going to assume that this quantity is coming from the normal distribution because this assumption, yeah. So, like we study before, so the probability or the different between the parameter and estimate is less than the precision. It's the same than before. So it's nothing new. Only that I am changing--this is the standard error for the mean here and this is different between proportion, between the parameter and the estimate. So this is given by the precision, yeah, and like before. So now, I am able to solve this equation and yet, the formula for the sample size. So the sample size formula is this one.

Remember, it's very similar than before. Before, we have like a—for the different between two proportion, the proportion estimate (inaudible), we have P multiply by one minus P, because this one is the variance of the variable. Now, is the variant of the variable that is continuous. So here, in the example, so, okay, we want to estimate (inaudible), we want to estimate the average price, the average price (inaudible) that 20 pills of the common used vitamin. The estimate is required to be within $0.10 of the true value, with 95 percent confidence. So the study size, a random sample of pharmacy is to be selective. So, basing a small

pilot study, the standard deviation of the price can be estimated like a—as a 85 percent, they mean that. The sigma is equal to .85 in dollars term. Yeah. So the question is how many pharmacy should be selected? So the solution, the same like before, we use the formula. So we have the precision, is .10 in dollars and we want that the different between the sample mean and the true mean will be less than .10 cent—I'm sorry, .10 dollars. So the variance is known, is .85 square. So we're using, again, confident level of 95 percent. I'm using the formula, the sample size will be 278 pharmacy should be selected. So this is you plot the number in the formula.

So (inaudible) is the variance is unknown? So, before, so if the variance is unknown, (inaudible) some changes in the confident interval. Because the confident interval now, we can use the normal table or the normal quantile because these parameters are known. So we can see now that the quantile from—we have to use the quantile for the key distribution. What is the problem here? The quantile for the distribution half end there. So half the sample size. So it's not than straightforward than before. So it's very similar. Now, instead that we have this, the C distribution, the normal distribution, we have like a (inaudible) because—so we know—oh, sorry. So we don't know the N and also we don't know the variant, so we need an estimate, an estimation of the sample variance. Yeah. And we must take into account in the sample size collection the distribution of the sample variance. Yeah?

So, now, it’s a little more complicated here because the minimum sample size n need to achieve a specified precision, so we need like a probability. So we need another parameter more. So we need to specify that what is the probability to observe this precision? For the reason, if we are going to assume that is that the variant is unknown, we need to use a special solver to compute this sample size. So, because the solver need to use like a—need to use like a tentative procedure because we can see now the sample size is here and here. (Inaudible) the sample size is here and here. So, or there. So, we need to compute using a solver and the parameter needed—okay, this is the parameter needed to compute sample size. Again, the sampling error or the precision, the alpha level, is specifying the difference between the sample estimate and the true parameter. We need also an estimate of the variance, but we need another parameter. We need the probability to obtaining a confident interval. The confident interval (inaudible) less than the precision. So there is another parameter we need to specify, because we have the uncertainty of the variance.

So here is one example, yeah. This is the same example than before, but now we are going to assume that the variant is unknown. To do that, so we need now, this is the solution. So, again, the same thing. The precision is .10. So we want the different between the sample mean and the true mean to be less than .10. The variance is the same. But now, we need to specify what the probability that the precision is less than—or different—the precision is obtained. So we need to specify this parameter, this one here. So we need to specify this probability. So,

in this case, I am going to use example that we—I produce when using the probability .05. That mean, that doesn't matter. I say, .05 is like a—is the same, almost the same than assuming that the variant is known. But, also, the .75 or .95, that the samples I have to—I have to hold (inaudible) for this quality, with probability bigger than Y minus (inaudible). So for using that, I need—we need to use a special solver. I am going to present example using SAS. So SAS can give you some precise calculation for the proportion and for the means only. Yeah, for the--not for proportion. Only for the means, I think, yeah. So for this example, okay, you can use this sentence--this is the problem. So here, I specify this (inaudible) is specify a (inaudible) that the precision is due by the confident interval. The thing is because a power is--they have more--they can do more--(inaudible) when you have hypothesis testing, that when you have confident interval. So here, I specify the precision; here I specify here the precision. Here, I specify the standard deviation, and here, I specify the half--the probability Y. (Inaudible) .75 and .95 for this parameter, alpha .05--and this another, let's say, sentence that we need to specify in the problem. This is probability (inaudible) this is in condition that the parameter will be within the confident interval. So for you, maybe it's better to know this one because you can have this unconditional or conditional. Unconditional is that you are conditional that the parameter are inside the confident interval. So I would recommend you to use unconditional. So here, in this problem, you can specify--you can require the computed sample size. So I--you--SAS require to--you use it like--(inaudible). Request a solution for sample size. So you do that--I'm sorry. You do that, this is the output from SAS.

So here, SAS specify, here in this column, you can see in this one, you can see the--specify one minus gamma, so the probability. Here is the--because it's a (inaudible) you will obtain the same one that you specify. You will--if I produce the very approximate one. So on the sample size, for example, you assume .5 will be 280. So almost the same that you produce one. You assume that the variance is known. Here. So here, you specify .75, you will observe 296, and usually, we specify like a .95 or .90 for this probability to hold. Will be about 318. So for this--this is the way the SAS percent (inaudible) produce, let's see what happen there--oh, here, okay.

Another example. Okay, this is the same example. But now, I am changing the variance. So I am holding the same with one minus gamma equal to .95, but I am going to change the variance so that you can see some capability there that the procedure have. So here I change the standard deviation, I keep the precision equal to .10, the one minus gamma equal to .95, the alpha level .05. But here, now, I want to see some plot. So the plot of these--so the plot is the--I want the solution--I want to, like, graph what I will have on the y axis the sample size matched with the precision. So here, we can see here. Oh, here, okay. Here, this part. I'll try to go more detail here. Here, in the X, I want to see that the range of the precision is change from .05 to .20. I want to reference--reference in line for the X axis precision equal to .1. Here, I am requiring to the program to use different symbols for me to see more--more easy to see the graph. So here, the simple size, we can see here again that the sample size, the variance increase,

the sample size increase. Here we can see for example, for the variance equal to .70, the sample size will be 222. For a variance equal to .85, the sample size will be 380. So the sample--there's nothing new here. So if the variance increase, the sample size increase. But we have now the SAS had the capability that we can see some graph, we can see the result visually. So for example here, the graph show that the--is divide the increase, yeah, the sample size increase. Here, okay, here in the Y, we have the sample size. This is the sample size. Here is the precision, in the X. So we can see that if the precision increased, the sample size decreased. Yeah. So here we see here is the variance increased, the sample size increased. Remember, here--okay here we see the label here. For .7, this is the graph for .7, this is the graph for the variance equal to .85, and this is the variance for--equal to one. So we can see here the--how the variance increase, the sample size increase. It's getting late. So I think it's maybe it's a good time to stop. Yeah. So what is the next one?

UNKNOWN SPEAKER: (Inaudible) difference of proportion.

RUBEN SMITH: Okay. Okay. So I think it's a good time to stop here. Yeah, because a lot of--huh. Yeah. I agree with you. Okay. Is there any question? I don't know. I will be happy to answer. But I think it's too late for question. So tomorrow, we are going to see again 8:30, the same room I think, yeah. Okay. Okay, time.

UNKNOWN SPEAKER: Now, for tomorrow, the exercise that I have for you is for me to think (inaudible) your problem, so what I'm hoping is that we can work on your problems and what do you have in your own states and so on and so forth, so you can think of those and have it ready tomorrow, what you would like to answer. That would be great, okay? Thank you. See you tomorrow.