Exploring Survey Data on Health Care
Prof. Pratap C. Mohanty
Department of Humanities and Social Sciences
Indian Institute of Technology, Roorkee

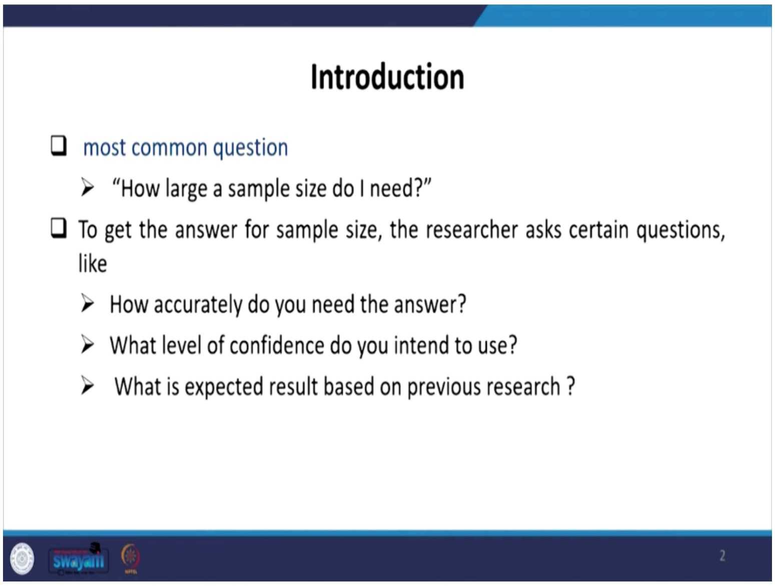
Lecture - 08
Sample Size Determination: Observational Study

Welcome students, this is our 8th lecture as part of the NPTEL-MOOC program on Handling Health Care Survey Data. This week we are trying to understand how to go to the field and collect data. In this regard, I have already discussed with you the requirement for the field like the questionnaire, the schedules, and some pre-piloting aspects of the survey.

In this lecture, we will be emphasizing the most important aspect of the field i.e., about sample size and that sample size is important not just in giving better statistical results, it is also important so far as the budget is concerned, so far as the environment is concerned.

So without delay let us start discussing about sample size determination. The most common question we often ask is “how large is our sample size?”, “What is that size do I need?”. To get the answer the researcher asks certain questions like, “how accurately do you need the answer?”.

(Refer Slide Time: 01:52)



Introduction

❑ most common question

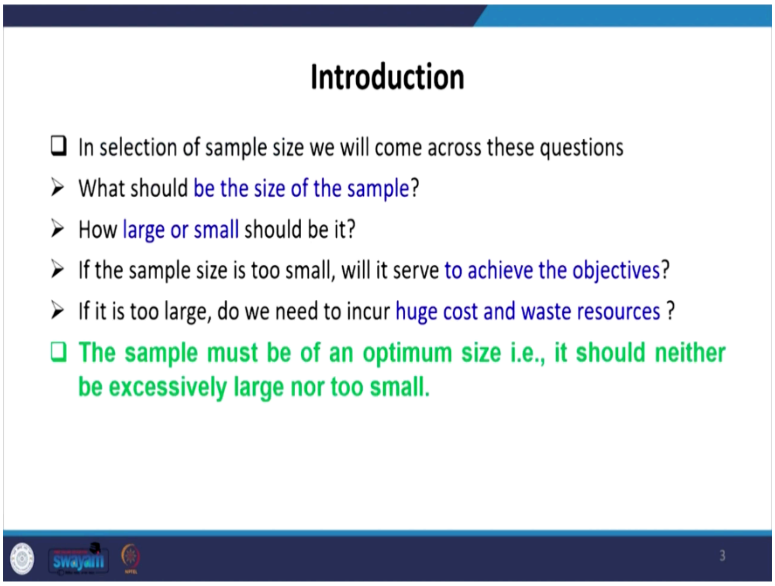
- “How large a sample size do I need?”

❑ To get the answer for sample size, the researcher asks certain questions, like

- How accurately do you need the answer?
- What level of confidence do you intend to use?
- What is expected result based on previous research ?

What level of confidence do you intend to use and what is the expected result based on previous research? So, with these three supporting questions, we will answer the first question about sample size.

(Refer Slide Time: 02:06)



Introduction

❑ In selection of sample size we will come across these questions

➢ What should be the size of the sample?

➢ How large or small should be it?

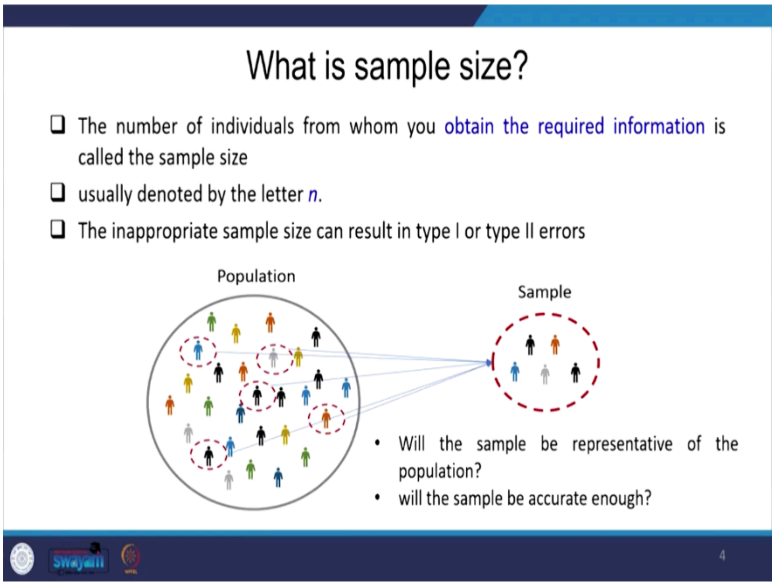
➢ If the sample size is too small, will it serve to achieve the objectives?

➢ If it is too large, do we need to incur huge cost and waste resources ?

❑ **The sample must be of an optimum size i.e., it should neither be excessively large nor too small.**

In the selection of sample size, we will come across the following questions; “what should be the size of the sample?”, “How large or small should it be?”, “If the sample size is too small, will it serve to achieve the objectives?”, “Is it too large, do we need to incur huge cost or resources?”. The sample must be an optimum size i.e. it should neither be excessively large nor too small.

(Refer Slide Time: 02:42)



What is sample size?

- ❑ The number of individuals from whom you obtain the required information is called the sample size
- ❑ usually denoted by the letter *n*.
- ❑ The inappropriate sample size can result in type I or type II errors

Population

Sample

- Will the sample be representative of the population?
- will the sample be accurate enough?

4

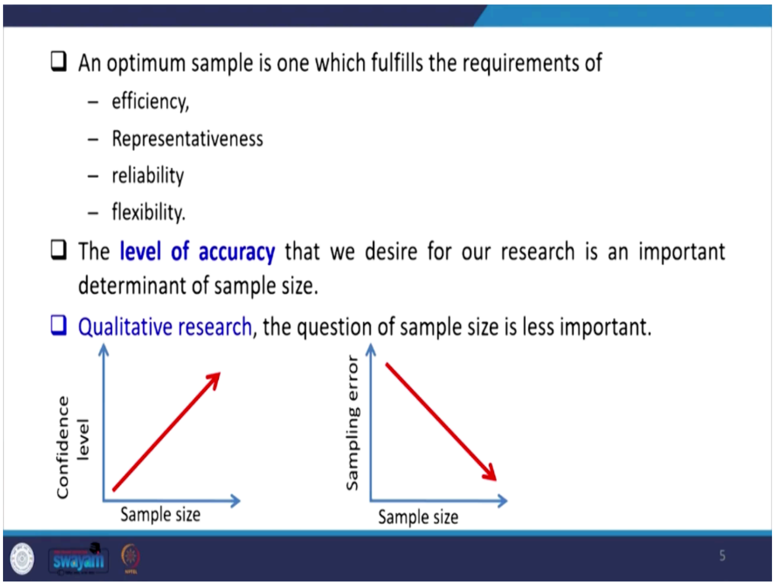
So, this is all about the introduction to sample size. Let us come to the discussion of what is the sample size? This is basically the number of individuals from whom we obtain the required information and this is usually denoted by the letter n.

The N we refer to population size. The size of our estimate brings some discussion about the population estimate from the sample i.e., from the n. n leads to certain assimilation, certain proximation for the large segment.

And that is why sample size matters. The inappropriate sample size can result in type I and type II errors. Like here we have given a population within a design and from there a sample is taken.

Some samples are selected, but this sample size is correct enough or not is to be discussed. Will the sample be representative of this population unit that we have just marked? Will the sample be accurate enough?

(Refer Slide Time: 04:07)



- ❑ An optimum sample is one which fulfills the requirements of
  - efficiency,
  - Representativeness
  - reliability
  - flexibility.
- ❑ The **level of accuracy** that we desire for our research is an important determinant of sample size.
- ❑ Qualitative research, the question of sample size is less important.

An optimum sample is one that fulfills the requirements of efficiency, representativeness, reliability, and flexibility. So, efficiency and other things I would discuss now. The level of accuracy that we desire for our sample or for our research is an important determinant of sample size.

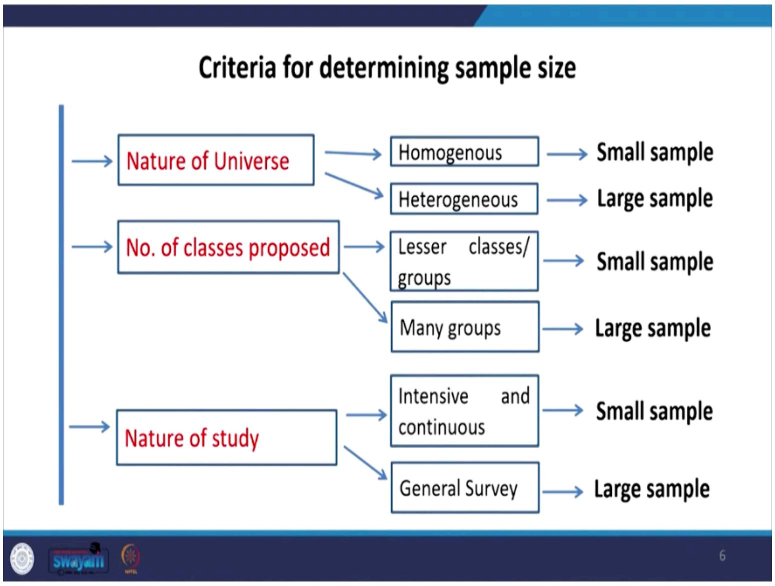
So, the level of accuracy gives better results. In qualitative research, the question of sample size is less important. Especially when we have a certain target of our observing qualitative information. In that case, the sample size may not be mattering much rather your different qualitative questions. The qualitative nature of observations are important.

Like in this chart, we have marked sampling error and confidence level. When your confidence level increases in your result, you can assure that your confidence level is very high for your testing. In that case, the sample size requirement is higher.

And when the sampling error or between the gap of the standard estimate with that of your sample statistics is higher, in that case, sample size also matters, but sample size reduces the sampling error.

So, it is inversely related. Some of the criteria to determine the sample size we have marked here like the nature of the universe.

(Refer Slide Time: 06:00)



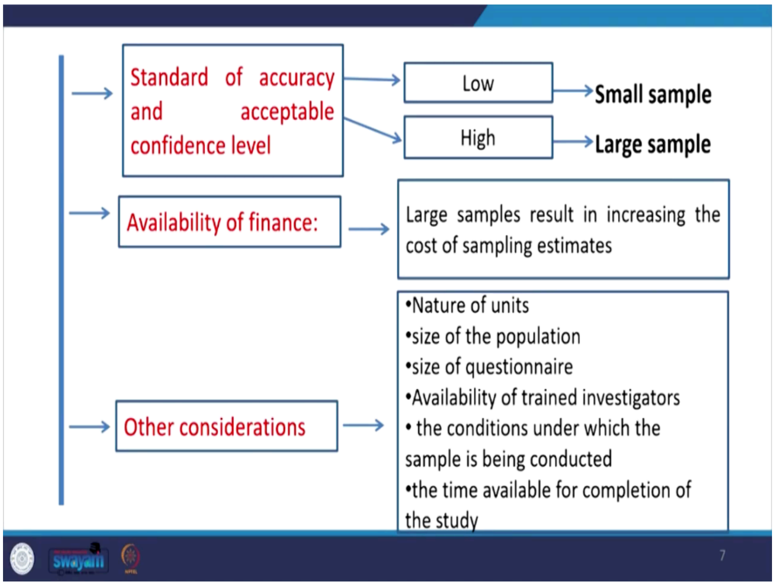
Criteria for determining sample size

What about your nature of universe, whether it is homogeneous or heterogeneous? When you are looking at your population, if the population unit itself is very homogeneous probably you need not require more sample size, small sample size would be justifying enough.

When it is highly heterogeneous you require a large sample size. Coming to the number of classes proposed when a lesser group or lesser class is compared to many groups, if there are many groups then many samples should be taken or the large sample should be taken, when it is of smaller groups or lesser groups, you are supposed to take less number of sample.

Coming to the nature of the study whether it is a very intensive or qualitative or general survey. In the case of a general survey, you are supposed to take a large sample, but when it is very intensive, very focused then you need to take so many samples. Then, coming to the standard of accuracy, acceptability and confidence level.

(Refer Slide Time: 07:26)



Standard of accuracy and acceptable confidence level

Low → Small sample

High → Large sample

Availability of finance:

Large samples result in increasing the cost of sampling estimates

Other considerations

- Nature of units
- size of the population
- size of questionnaire
- Availability of trained investigators
- the conditions under which the sample is being conducted
- the time available for completion of the study

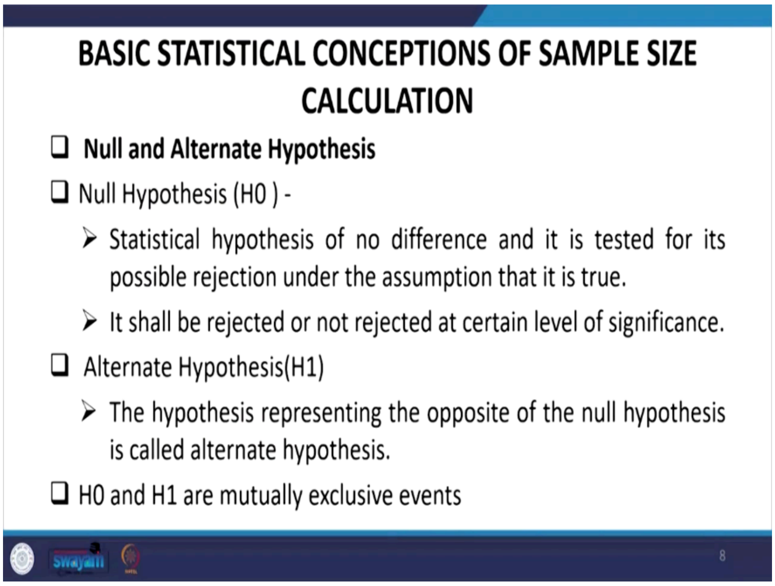
Whether it is high confidence level or low confidence level, if you target for high confidence level then, of course, you are supposed to take large size sample size as against the lesser one.

Like if you wanted to test a consignment and out of that how many products are good or bad, you have simply taken some size randomly.

If your acceptable confidence level is higher then, you are supposed to take a higher sample size. Regarding availability to finance if it is higher then, large samples result in increased cost of sampling estimates.

Or some other considerations are equally important like the nature of the unit you are supposed to take, size of the population, size of the questionnaire, availability of trained investigators, the conditions under which the sample is being conducted, the time available for completion of the study. These are also considered while taking the sample.

(Refer Slide Time: 08:45)



BASIC STATISTICAL CONCEPTIONS OF SAMPLE SIZE CALCULATION

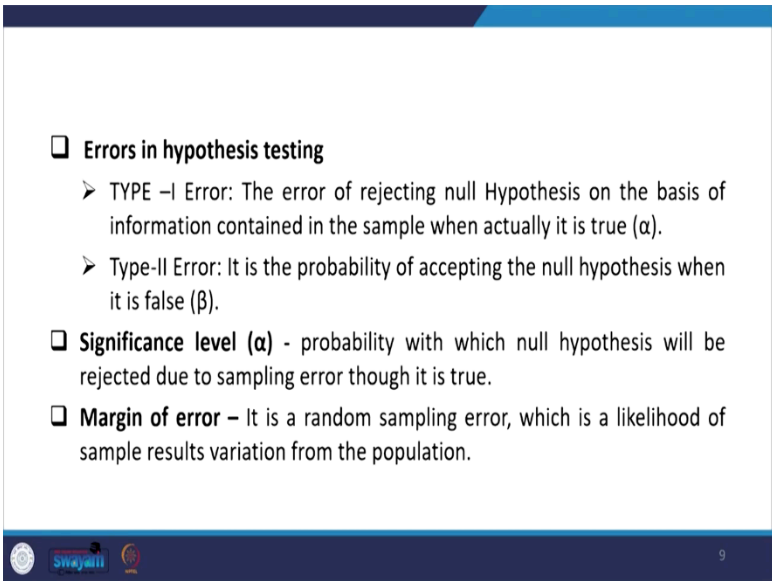
- ❑ **Null and Alternate Hypothesis**
- ❑ Null Hypothesis (H0) -
  - ➢ Statistical hypothesis of no difference and it is tested for its possible rejection under the assumption that it is true.
  - ➢ It shall be rejected or not rejected at certain level of significance.
- ❑ Alternate Hypothesis(H1)
  - ➢ The hypothesis representing the opposite of the null hypothesis is called alternate hypothesis.
- ❑ H0 and H1 are mutually exclusive events

Now, we are moving to the understanding of some basic statistical concepts related to sample size and its calculation. So, for sample size calculation why it is necessary and where is it necessary, is connected to hypothesis testing. So, as we all know that we have a null hypothesis and one alternative hypothesis. The null hypothesis basically is donated with H0 and H1 for the alternative one.

The null hypothesis is the statistical hypothesis of no difference and it is tested for possible rejection under the assumption that it is true. Though null hypothesis is considered as if there is no difference or it is always true. So, if your test is going to be rejected then accordingly, we can have some discussion.

Like in the null hypothesis, it can be rejected or not at a certain level of significance. As against the null hypothesis we have an alternative hypothesis. When the null hypothesis is rejected; that means, you are accepting your alternative hypothesis. The hypothesis representing the opposite of the null hypothesis is called the alternative one. So, HO and H1 are mutually exclusive events, and accordingly, we take testing.

(Refer Slide Time: 10:20)



- ❑ **Errors in hypothesis testing**
  - ➢ TYPE –I Error: The error of rejecting null Hypothesis on the basis of information contained in the sample when actually it is true (α).
  - ➢ Type-II Error: It is the probability of accepting the null hypothesis when it is false (β).
- ❑ **Significance level (α)** - probability with which null hypothesis will be rejected due to sampling error though it is true.
- ❑ **Margin of error –** It is a random sampling error, which is a likelihood of sample results variation from the population.

9

Now, there are some errors in the hypothesis testing. Those errors are called type I error or type II error. Type I error is basically when we are rejecting the null hypothesis when we consider that it is true to test any context you consider and we simply say that it is true.

For example, suppose you wanted to study the implications of COVID on the rural economy and their nutritional level. The first null hypothesis in this context might be that COVID has created no difference in the population, so far as its nutrition status is concerned.

So, the alternative will be yes it has created differences. So, the created difference may be positively maybe negatively. So, it has a two-tailed interpretation, a two-tailed direction.

So, one-tailed when we say that you wanted to test that COVID has positively affected the nutritional status. If that is your null hypothesis, then the alternative will be it is not affected or influenced.

So now, the error of basically the type one error we say in the error of rejecting null hypothesis on the basis of information collected from the sample and when it is actually true. So, you are actually rejecting the null hypothesis, when it is actually true then is called type I error. So, why it is called type one error? Type I error is denoted by α and type II error is denoted by β or 1- α.

So what is that type II error? Now, let us come to the first type I error, rejecting a null hypothesis when it is actually true. Like you said you wanted to test for a context, but your sample has given certain information on the basis that you are rejecting your population unit.

For example, one food inspector has come to a place and inspects vegetable quality in a supermarket or in a mandi. Now, the person has selected some samples. Based on the sample the inspector is finding that the vegetable quality is not good. These are infected qualities or there are certain problems with the vegetable.

But in fact, the total available vegetables in that market on average is good. In your sampling unit, the response you have taken might have been erroneous, which might have given certain wrong signals or wrong estimates. Based on that you have rejected your population that the entire mandi and its products are bad.

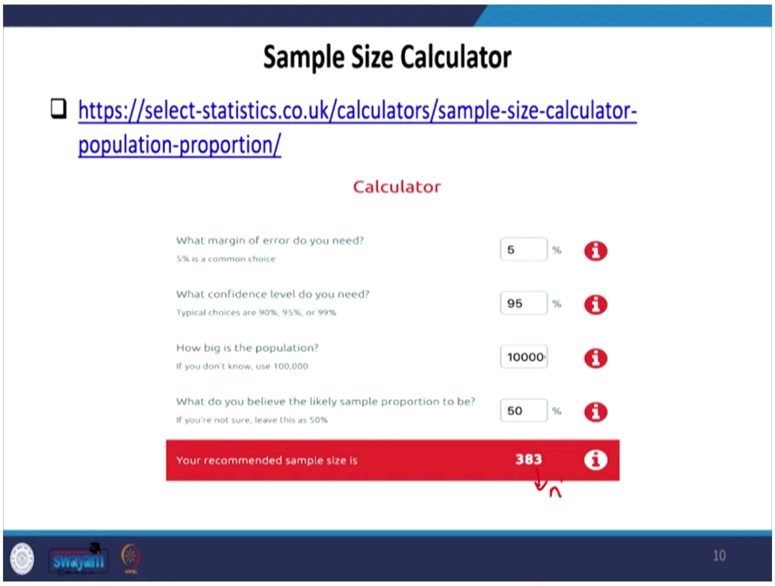
So, that might be due to sampling error or non-sampling error. We are going to discuss some of those things. But first of all, when we are rejecting the null hypothesis based on your sample estimates these are called type I error.

The opposite is type II error which is an error when you are accepting a null hypothesis when it is actually false. Your sample might have given the correct result, but your total product is not right, in that case, that is called β.

So, the significance level is also donated by the alpha which is a probability with which the null hypothesis will be rejected due to sampling error though it is true, that is called α error or significance level. So, the margin of error is a random sampling error which is a likelihood of sampling results variation from the population.

Basically, sampling error gives the standard deviation of the sampling distribution. So, the standard deviation of the sampling distribution is actually called sampling error. Now, what are the catch in understanding the sample size you can also click on this web link, you will be better guided.

(Refer Slide Time: 15:24)



Sample Size Calculator

❑ https://select-statistics.co.uk/calculators/sample-size-calculator-population-proportion/

Calculator

What margin of error do you need?
5% is a common choice
5 %

What confidence level do you need?
Typical choices are 90%, 95%, or 99%
95 %

How big is the population?
If you don't know, use 100,000
10000

What do you believe the likely sample proportion to be?
If you're not sure, leave this as 50%
50 %

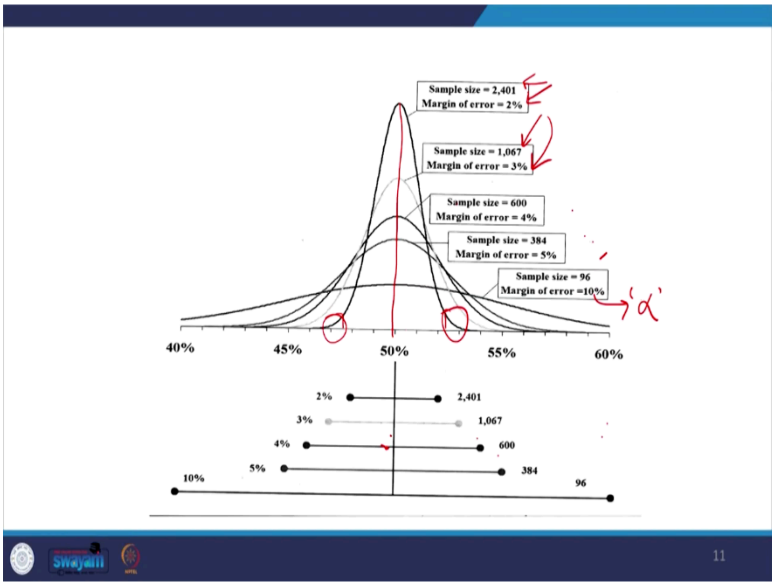
Your recommended sample size is 383

n

I will guide you so that you can able to have these aspects like what is your margin error, what is your confidence level, how big is the population, how do you believe the likely proportion to be included in the sample.

First of all, margin error if we set at 5 percent and your confidence level is at 95 percent and if your population size is 10000, then with the 50 percent probability of that sample to be included in your sample size is going to give you 383 number of samples. And that you can just test I will clarify those concept, confidence level, margin error and proportion everything in my successive slides.

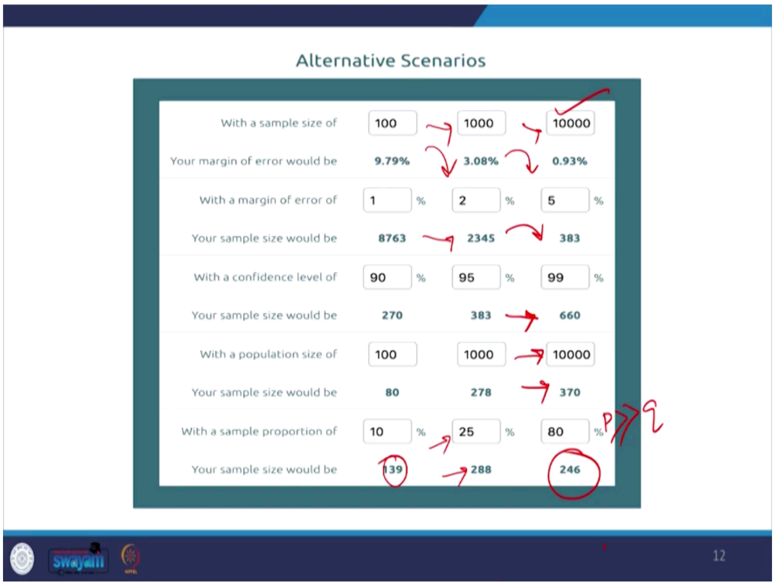
(Refer Slide Time: 16:27)



So, coming to these margin errors and the sample size determination. When the margin of error in your calculation is very less that means, you are supposed to take more sample size i.e., the sample size is higher. If your margin error is just 2 per cent you are supposed to take a higher sample size. When you are taking less size, you may compromise with a margin error. And when your sample size is increasing your distribution is becoming more leptokurtic.

It is symmetric normally distributed, but when the sample size is lesser, α regions are expected to be higher because the rejection region will be higher. But here when it is more of leptokurtic your margin error is very less.

(Refer Slide Time: 17:54)



Alternative Scenarios

With a sample size of	100	1000	10000
Your margin of error would be	9.79%	3.08%	0.93%
With a margin of error of	1 %	2 %	5 %
Your sample size would be	8763	2345	383
With a confidence level of	90 %	95 %	99 %
Your sample size would be	270	383	660
With a population size of	100	1000	10000
Your sample size would be	80	278	370
With a sample proportion of	10 %	25 %	80 %
Your sample size would be	139	288	246

So, we can find it out in our chart. Some alternative scenarios to calculate sample size are given. When the sample size is 100 to 1000 and then 10000, your margin error would be expected to fall. On the contrary, when your margin error is higher you may be supposed to take less sample size.

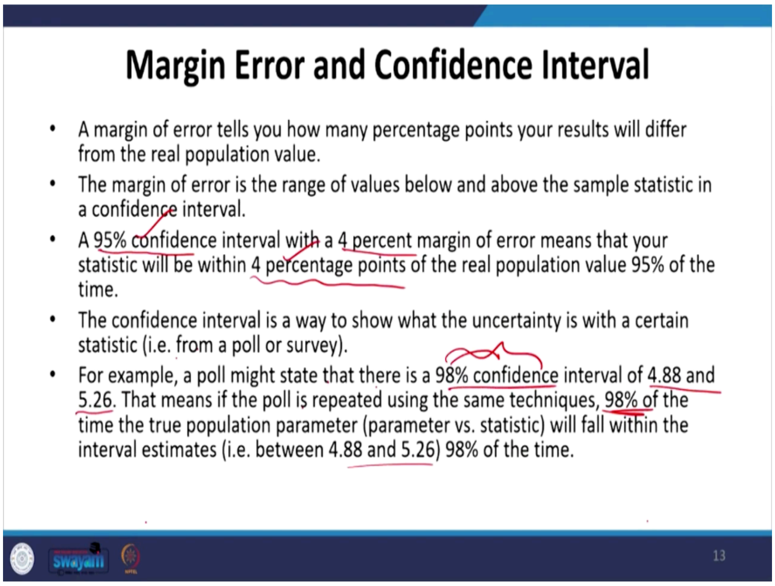
Similarly, some decisions are taken regarding confidence level, when your result is expected to be very confidently building or your result is quite confident then, you are supposed to take more sample size.

And when the population size is much higher you are also supposed to take more sample size. Now, similarly when your sample proportion is higher to some extent like with a sample proportion of 10 percent you are taking a certain n, but what is that you need not decide depending upon population size. When your proportion is 25 percent, you are supposed to take more.

When your sample proportion is becoming much higher then, the probability of success is higher i.e., p is higher and q is lesser, the probability of failure. In that case, you need not take more sample.

Some of the samples or the pure samples can also give you better result.

(Refer Slide Time: 19:53)



Margin Error and Confidence Interval

- A margin of error tells you how many percentage points your results will differ from the real population value.
- The margin of error is the range of values below and above the sample statistic in a confidence interval.
- A 95% confidence interval with a 4 percent margin of error means that your statistic will be within 4 percentage points of the real population value 95% of the time.
- The confidence interval is a way to show what the uncertainty is with a certain statistic (i.e. from a poll or survey).
- For example, a poll might state that there is a 98% confidence interval of 4.88 and 5.26. That means if the poll is repeated using the same techniques, 98% of the time the true population parameter (parameter vs. statistic) will fall within the interval estimates (i.e. between 4.88 and 5.26) 98% of the time.

13

Now we are clarifying some concepts like margin error and confidence interval. We researchers are often confused about these two concepts. Like margin error, it tells you how many percentage points your result will differ from the real population unit or population value.

The margin error is in fact the range of values below and above the sample statistic in a confidence interval.

So, from a confidence interval, it defines your margin error. As a 95 percent confidence interval with a 4 percent margin means, your statistic will be within 4 percentage points of the real population value 95 percent of the time.

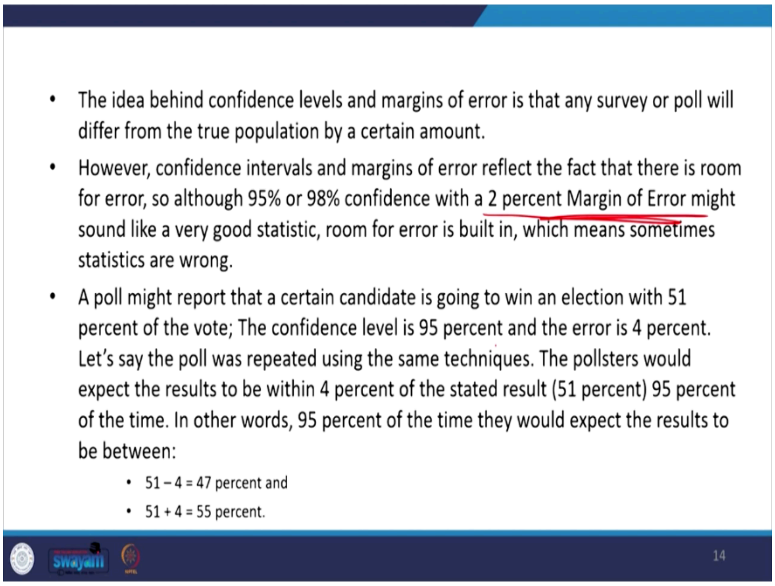
So, all 95 percent of the time, if you test it, your result from the mean value will be varying by 4 percent plus or minus. So, this is where confidence interval and margin error work out and gives the right guidance in testing. The confidence interval is a way to show what the uncertainty is with a certain statistic so; it gives a lower slab and upper slab in the poll or in a survey.

One example we have cited here, a poll might state that there is a 98 percent confidence interval of 4.88 and 5.26. 4.8 and 5.26 indicate your range of upper limit and lower limit because of the margin error.

That means if the poll is repeated using the same techniques, 98 percent of the time the true population parameter will fall within the limit of 4.88 to 5.26.

So that means, the error is very minimum and gives a range of your result. The idea behind confidence level and margins of error is that any survey or poll will differ from the true population by a certain amount.

(Refer Slide Time: 23:04)



- The idea behind confidence levels and margins of error is that any survey or poll will differ from the true population by a certain amount.
- However, confidence intervals and margins of error reflect the fact that there is room for error, so although 95% or 98% confidence with a 2 percent Margin of Error might sound like a very good statistic, room for error is built in, which means sometimes statistics are wrong.
- A poll might report that a certain candidate is going to win an election with 51 percent of the vote; The confidence level is 95 percent and the error is 4 percent. Let's say the poll was repeated using the same techniques. The pollsters would expect the results to be within 4 percent of the stated result (51 percent) 95 percent of the time. In other words, 95 percent of the time they would expect the results to be between:
  - 51 – 4 = 47 percent and
  - 51 + 4 = 55 percent.

However, the confidence interval and the margins of error reflect the fact that there is room for error, so although 95 percent or 98 percent confidence with a 2 percent margin of error might sound like a very good statistic, room for error is built, which means sometimes statistics are wrong.

So, this indicates, though 2 percent errors are mentioned and we are saying that this result is going to be almost on the time confidence. So, this always gives certain room for errors and that error is very petty. So that is why statistics give a better direction for taking decisions.

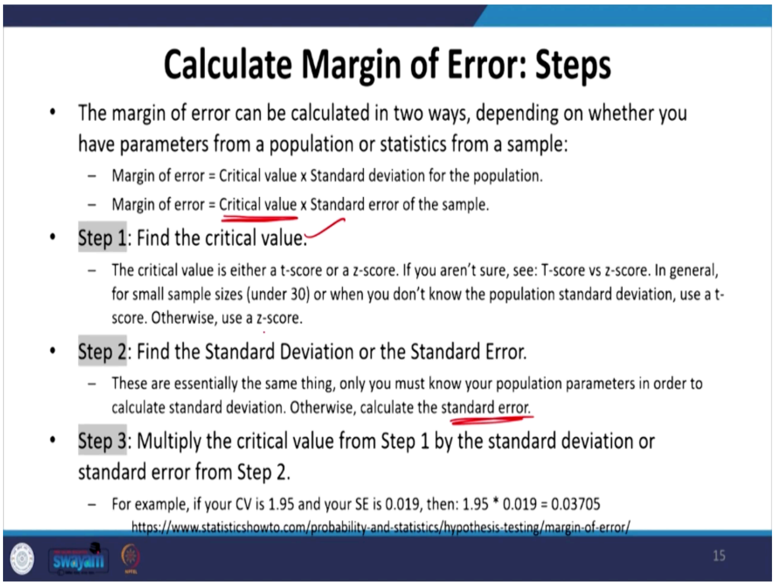
A poll might report that a certain candidate is going to win an election with 51 percent of the vote; the confidence level is 95 percent and the error is 4 percent; that means plus 51 plus 4 and 51 minus 4.

Let's say the poll was repeated using the same technique 95 times so that means, there will be an error of +4 or -4.

So, that will vary your result. When we are saying 51 percent vote a candidate is going to get that may get 55 percent or may get 47 percent. Now of course, not below that; that means, we can guess about the product or about the candidate in an election better.

Similarly, while calculating the margin of error there are some steps involved. The margin of error can be calculated in two ways, depending on whether you have parameters from a population or statistics from the sample. The margin of error is calculated with the critical values times standard deviation of the population if we are calculating for the population.

(Refer Slide Time: 25:28)



Calculate Margin of Error: Steps

- The margin of error can be calculated in two ways, depending on whether you have parameters from a population or statistics from a sample:
  - Margin of error = Critical value x Standard deviation for the population.
  - Margin of error = Critical value x Standard error of the sample.
- Step 1: Find the critical value.
  - The critical value is either a t-score or a z-score. If you aren't sure, see: T-score vs z-score. In general, for small sample sizes (under 30) or when you don't know the population standard deviation, use a t-score. Otherwise, use a z-score.
- Step 2: Find the Standard Deviation or the Standard Error.
  - These are essentially the same thing, only you must know your population parameters in order to calculate standard deviation. Otherwise, calculate the standard error.
- Step 3: Multiply the critical value from Step 1 by the standard deviation or standard error from Step 2.
  - For example, if your CV is 1.95 and your SE is 0.019, then: 1.95 * 0.019 = 0.03705

https://www.statisticshowto.com/probability-and-statistics/hypothesis-testing/margin-of-error/

Similarly, for the sample critical value and standard error. Second, the 1st step is how to find the critical value. Who gives the critical value? So, first of all, we say critical value for the sample case or for the population let us understand the critical value.

The critical value is either a T score or a Z score, Z tabulated value or the T-tabulated value. If you are not sure then we can see the T score table or the Z score table.

In general, for small sample sizes, usually less than 30 or when you do not know about the population standard deviation, then we apply the T score.

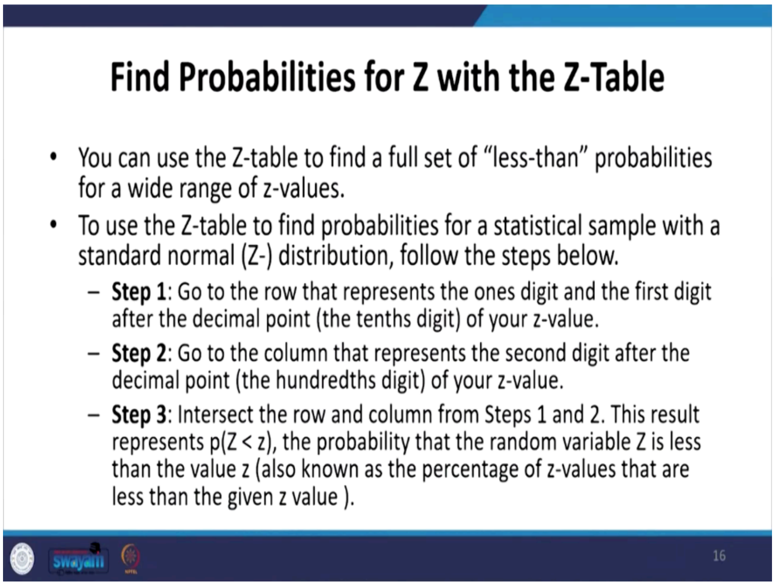
Otherwise, if it is correctly known then the Z score can be applied. The second step is to find out the standard deviation or the standard error. These are essentially the same thing, standard error is used for the sample and standard deviation for the population unit.

You only must know your population parameter in order to calculate standard deviation, otherwise, calculate the standard error from the sample. The third step is to multiply the critical value that critical value once we have determined here, we may multiply with the standard error that will give us the margin error.

For example, if the critical value is 1.95 and standard error is 0.019, then 1.95 times 0.019 boils downs to 0.03.

So, 3.7 is your standard error. So, further details you can follow this link I have appended to this page. The second one is how to find the probability value of Z with the Z table and the T value with the T table. I have put some references for your understanding like you can use the Z-table to find a full set of "less than" probabilities for a wide range of z values.

(Refer Slide Time: 28:21)



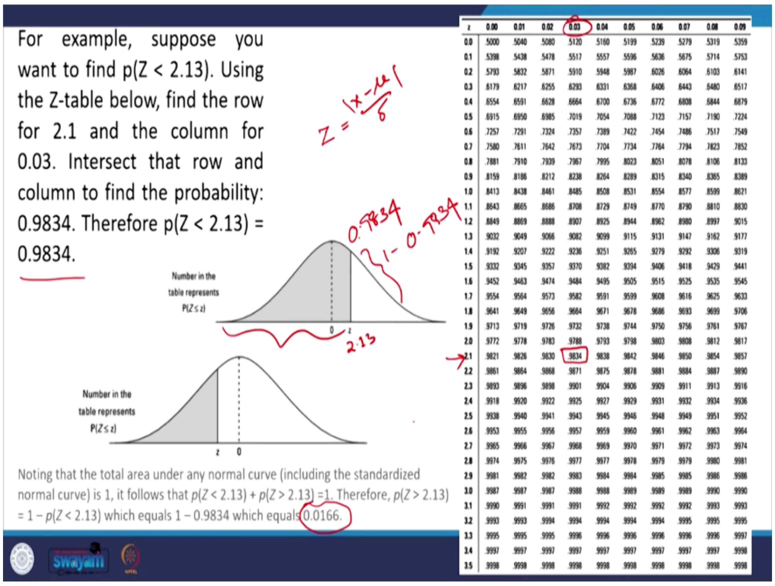
Find Probabilities for Z with the Z-Table

- You can use the Z-table to find a full set of "less-than" probabilities for a wide range of z-values.
- To use the Z-table to find probabilities for a statistical sample with a standard normal (Z-) distribution, follow the steps below.
  - **Step 1**: Go to the row that represents the ones digit and the first digit after the decimal point (the tenths digit) of your z-value.
  - **Step 2**: Go to the column that represents the second digit after the decimal point (the hundredths digit) of your z-value.
  - **Step 3**: Intersect the row and column from Steps 1 and 2. This result represents p(Z < z), the probability that the random variable Z is less than the value z (also known as the percentage of z-values that are less than the given z value).

1st step we need to go to the row that represents the one digit and the first digit after the decimal point like the tenths digit of your z value. I will show it to you by example.

So, then we go to the column that represents this second digit after the decimal point of your z value. The third step is to intersect the row and column from 1 and 2 this result represents the probability of that z value with that limit.

(Refer Slide Time: 29:24)



For example, suppose you want to find p(Z < 2.13). Using the Z-table below, find the row for 2.1 and the column for 0.03. Intersect that row and column to find the probability: 0.9834. Therefore p(Z < 2.13) = 0.9834.

Z = |x-μ|/σ
0.9834
1- 0.9834
Number in the table represents P(Z ≤ z)
0
z
2.13
Number in the table represents P(Z ≤ z)
z
0

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998

Noting that the total area under any normal curve (including the standardized normal curve) is 1, it follows that $p(Z < 2.13) + p(Z > 2.13) = 1$. Therefore, $p(Z > 2.13) = 1 - p(Z < 2.13)$ which equals 1 – 0.9834 which equals 0.0166.

And rest you can follow from this and I am just going to show it here like on this table the one-sample Z table, but this Z table is positively skewed. One is given the z value which is higher than 0.

So, if you are interested in finding out the probability of the z value at 2.13 using the Z table value, we can find out like in the first table you need to check with 2.1.

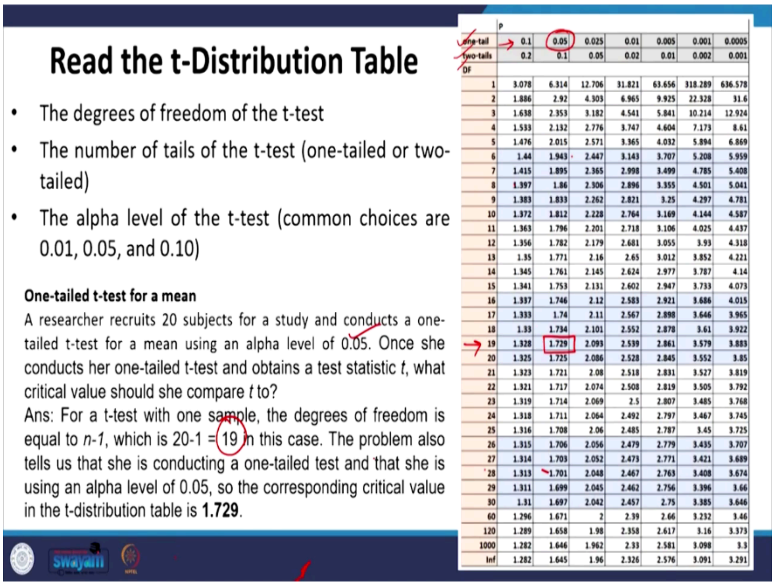
From the row where is a 2.1, we can mark it first here, then the second digit after decimal that is 0.03, we can mark it 0.03 over here and now with the intersection we get a value this is going to give us 0.9834. So, basically, it says the z value with 2.13 what is their probability value over this region.

So, in this case, we have derived the tabulated value 0.9834. So, now, noting that the total area under any normal curve including the standard normal curve is one. Here we have defined the Z value as the standard x.

So, now it follows with the z value less than 2.13 the probability value is 0.9834. The cumulative probability value of the rest of the areas will be 1- 0.9834.

So, this should be added to 1. Therefore, the probability of Z greater than that of 2.13 should be 1-p of Z of less than 2.13.

(Refer Slide Time: 32:14)



Read the t-Distribution Table

- The degrees of freedom of the t-test
- The number of tails of the t-test (one-tailed or two-tailed)
- The alpha level of the t-test (common choices are 0.01, 0.05, and 0.10)

One-tailed t-test for a mean

A researcher recruits 20 subjects for a study and conducts a one-tailed t-test for a mean using an alpha level of 0.05. Once she conducts her one-tailed t-test and obtains a test statistic *t*, what critical value should she compare *t* to?

Ans: For a t-test with one sample, the degrees of freedom is equal to *n-1*, which is 20-1 = 19 in this case. The problem also tells us that she is conducting a one-tailed test and that she is using an alpha level of 0.05, so the corresponding critical value in the t-distribution table is **1.729**.

	P						
one-tail	0.1	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	0.2	0.1	0.05	0.02	0.01	0.002	0.001
DF							
1	3.078	6.314	12.706	31.821	63.656	318.289	636.578
2	1.886	2.92	4.303	6.965	9.925	22.328	31.6
3	1.638	2.353	3.182	4.541	5.841	10.214	12.924
4	1.533	2.132	2.776	3.747	4.604	7.173	8.61
5	1.476	2.015	2.571	3.365	4.032	5.894	6.869
6	1.44	1.943	2.447	3.143	3.707	5.208	5.959
7	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	1.397	1.86	2.306	2.896	3.355	4.501	5.041
9	1.383	1.833	2.262	2.821	3.25	4.297	4.781
10	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	1.356	1.782	2.179	2.681	3.055	3.93	4.318
13	1.35	1.771	2.16	2.65	3.012	3.852	4.221
14	1.345	1.761	2.145	2.624	2.977	3.787	4.14
15	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	1.337	1.746	2.12	2.583	2.921	3.686	4.015
17	1.333	1.74	2.11	2.567	2.898	3.646	3.965
18	1.33	1.734	2.101	2.552	2.878	3.61	3.922
19	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	1.325	1.725	2.086	2.528	2.845	3.552	3.85
21	1.323	1.721	2.08	2.518	2.831	3.527	3.819
22	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	1.319	1.714	2.069	2.5	2.807	3.485	3.768
24	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	1.316	1.708	2.06	2.485	2.787	3.45	3.725
26	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	1.311	1.699	2.045	2.462	2.756	3.396	3.66
30	1.31	1.697	2.042	2.457	2.75	3.385	3.646
60	1.296	1.671	2	2.39	2.66	3.232	3.46
120	1.289	1.658	1.98	2.358	2.617	3.16	3.373
1000	1.282	1.646	1.962	2.33	2.581	3.098	3.3
Inf	1.282	1.645	1.96	2.326	2.576	3.091	3.291

Now, let us move to the T distribution. T table I have presented here for your reference. Usually, the researcher who just started reading the statistics at the undergraduate level might get confused in reading the tabulated values.

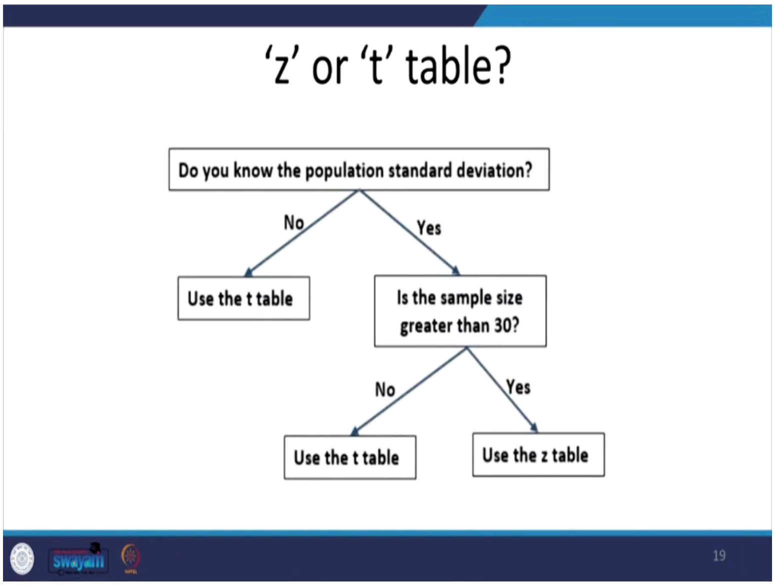
So, reading the T distribution we require three things. The first one is degrees of freedom, then the number of tails of the t-test whether it is a one-tailed test or two-tailed test and the third one is the critical limit, the alpha level. The critical limit that is at 1 percent or 5 percent or 10 percent.

So, that has to be mentioned in this example we have given in the table the figures are given one-tailed, two-tailed and degrees of freedom. First of all, like in a study, a researcher requests 20 subjects and conducts a one-tailed t-test which is clearly mentioned.

Once she conducts her one-tailed t-test and obtains a test statistic t, what critical value should she compare t ? So, what is the critical value or the tabulated value of t?

So, the degree of freedom is actually 19 in this case which is 20 minus 1. So, we are straightway going to check 19 here on the row. The problem also tells us that she is conducting a one-tailed test. So, we produce a one-tailed test we have to compare these figures these p limit is the alpha level and it is said that it has to be at a 5 per cent level.

(Refer Slide Time: 34:43)



‘z’ or ‘t’ table?

Do you know the population standard deviation?

No

Use the t table

Yes

Is the sample size greater than 30?

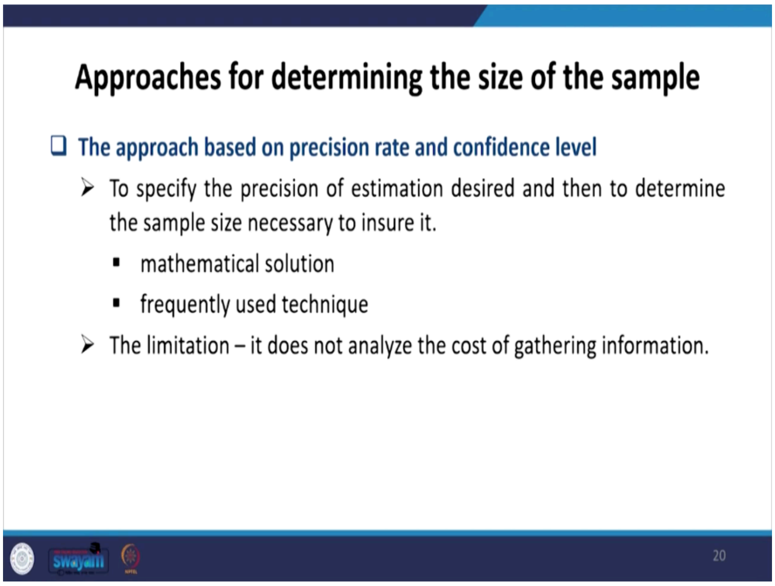
No

Use the t table

Yes

Use the z table

(Refer Slide Time: 34:45)



Approaches for determining the size of the sample

- ❑ The approach based on precision rate and confidence level
  - ➢ To specify the precision of estimation desired and then to determine the sample size necessary to insure it.
    - ▪ mathematical solution
    - ▪ frequently used technique
  - ➢ The limitation – it does not analyze the cost of gathering information.

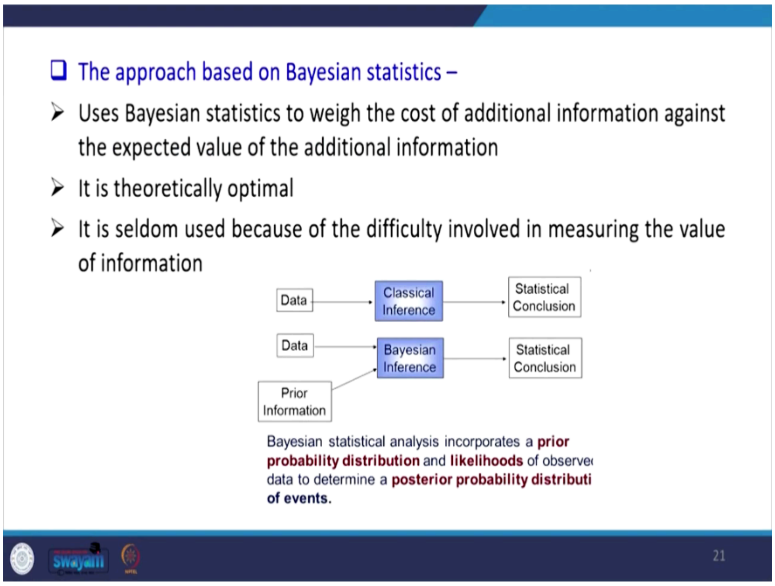
Now, the decision we usually take, once your estimated statistic is higher than that of the tabulated value; that means, we are not accepting the null hypothesis. That means, we are rejecting the null hypothesis; that means, that variable is significantly deferring to the assumption.

So, accordingly we take decisions. So, the final call on this is whether t-test or z-test. So if we know the population, deviation stuff i.e., population standard deviation then; obviously we will go for the z-test and if we do not know then we are supposed to go for t-test.

Once we know our standard deviation then the second call is what is the sample size if it is greater than 30 then we will use the Z table else it is of T table.

Now, there are approaches for determining the size of the sample. The approach based on precision rate and confidence level. To specify the precision of estimation desired and then to determine the sample size necessary to insure it. We require mathematical solutions and frequently used techniques. The limitation is, it does not analyze the cost of gathering information.

(Refer Slide Time: 36:13)



□ The approach based on Bayesian statistics –

- Uses Bayesian statistics to weigh the cost of additional information against the expected value of the additional information
- It is theoretically optimal
- It is seldom used because of the difficulty involved in measuring the value of information

Bayesian statistical analysis incorporates a **prior probability distribution** and **likelihoods** of observe data to determine a **posterior probability distributi of events.**

21

So, some of the approaches are based on Bayesian statistics. I am not going to the depth of it, but just to give you certain minimum guidance because of the fact that we are not going to apply it. So, some approaches use Bayesian statistics to weigh the cost of additional information against the expected value of the additional information.

And it is theoretically optimal. It is seldom used because of the difficulty involved in measuring the value of information. So, it goes by understanding prior probability distribution, likelihoods of the observed data and posterior probability distribution of the events.

So, that is why this is hardly used. I think I should stop here. I will start with the determination through the approach based on precision rate and confidence level for the next class. Let me stop here, I will continue from the next class and clarify the exact sample size.