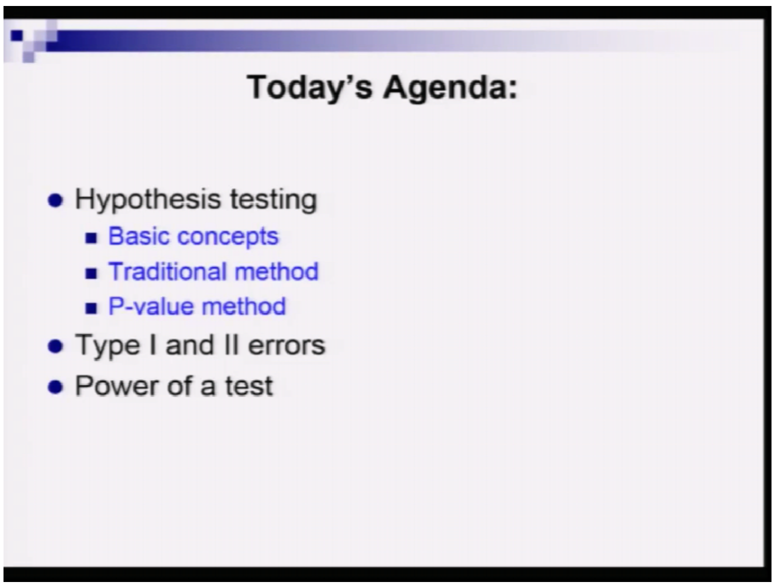
Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology Kanpur
Lecture 11
Hypothesis Testing (Part I)

Hello, friends. Welcome back to the lecture series on Applied Statistics and Econometrics. So today, we are going to start our discussion on the second part of statistical inference and that is called hypothesis testing.

(Refer Slide Time: 00:28)



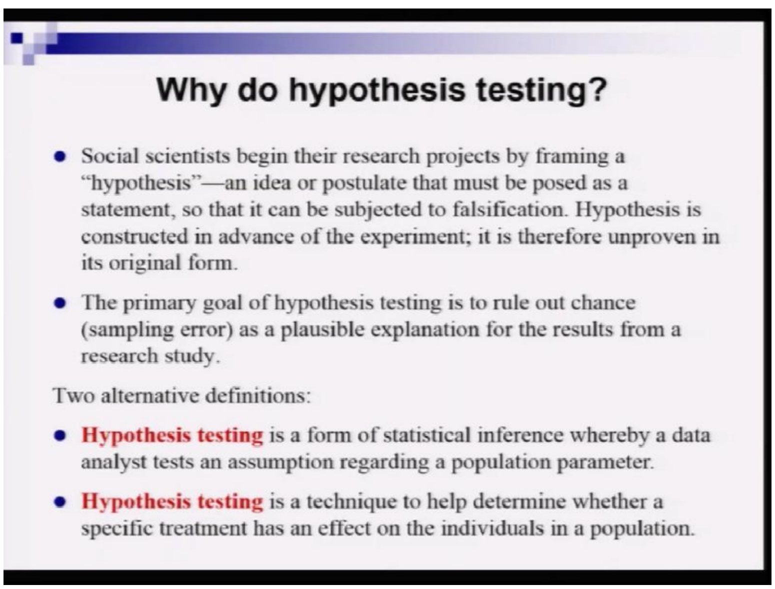
So, before we start discussing hypothesis testing theory, let us have a look at today's agenda items. So, we will talk about the basic concepts and I am going to show you some examples to make these concepts clear. Then, I am going to talk about 2 different techniques or methods for hypothesis testing and they are the traditional method and there is one P value method that we are going to talk about.

Then I am going to talk about type I and type II errors. And finally, I will end today's lecture with a very brief discussion on power of a test. So, today's discussion is going to be mostly theoretical in nature. And in today's lecture, I am going to talk about various steps of hypothesis testing procedure. So, that in the next lecture, we can cover more detailed tests and more specific tests.

So, let me start with some philosophical discussion on hypothesis testing. I will start with what do we mean by hypothesis? How they are important and play a role in socio-economic research? And then, I will define hypothesis testing in formal words and then, we will discuss the theoretical portion of hypothesis testing.

So, when we begin the scores, I have already told you that in socio-economic research and why socio-economic, in managerial and many other types of research areas, people or scholars start with a particular hypothesis in mind and then they actually try to collect data from reality and then, they try to gather evidence in favor or in oppose of their hypothesis with which they have started. So, hypothesis actually is heart and soul of in every socio-economic research. So, let us first define hypothesis and then see how that is related to the statistical theory of hypothesis testing.

(Refer Slide Time: 02:44)



So, what is a hypothesis? So, hypothesis is basically an idea or postulate that must be posed as a statement, so that it can be subjected to falsification. So, a hypothesis you remember here, it is very important point that hypothesis is constructed at the very beginning of an experiment. So, nothing has been done, no data has been collected, no equation has been retained, hypothesis comes at first. And then data collection takes place and then one can try to prove or disprove the hypothesis with which one started their research project.

So, the primary goal of hypothesis testing is to rule out the chance of sampling error as a plausible explanation for the results from a research study. So, what do we mean by that? So note that, we actually frame the hypothesis on our universe. So, hypothesis is a very general idea or general concept that is applicable to the population out there. But as statisticians or econometricians we have resource constraints.

We cannot reach out to the each and every element of the population in the study and collect data from them. So, we have to collect a small sample an adequate number of observations from the population and then based on that adequate sample, we actually try to draw some inferences about the population parameter. Now, that is basically the whole idea of statistical inference. Now, as there is sampling variation, of course, you get some information from sample 1, but if you draw another, sample 2 or sample 3, then you may not get back the same information or statistic value.

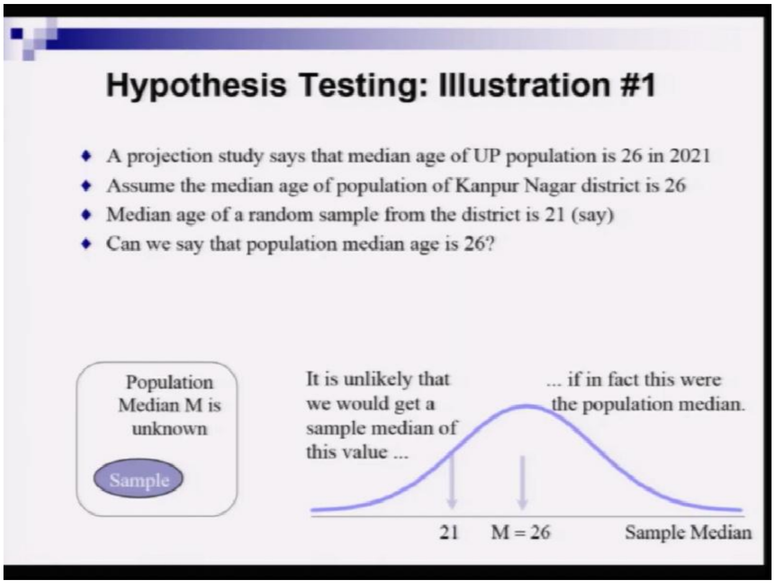
So, in that case, if you conclude something based on one particular sample based observations or learnings then there is a chance that you actually are making a wrong decision because there is sampling error involved. So, there is a possibility that if you take a different sample or you have been dealing with a different sample, you may have actually found out sufficient evidence in favor of your hypothesis.

So now we are going to talk about two alternative definitions and I am going to show you illustrations for each type. So, that these two definitions become very clear in your mind. They are not competing definitions, these are basically doing the same thing. But they are just you can say that these are 2 different perspectives or viewpoints at hypothesis testing from two different directions.

So, let us start with the direction with which we are comfortable with as of now, because we have been discussing this thing again and again. So, by following that statistical inference discussions, we can say that hypothesis testing is a form of statistical inference, whereby a data analyst tests and assumption regarding our population parameter. And this thing will be much clearer in the next slide when I am going to show you an illustration.

And the second definition one can put forward is from an impact evaluation point of view. So, from that perspective, I can say that hypothesis testing is a statistical technique that helps to determine whether a specific treatment has an effect on the individuals in a population or not. So, this point also will be clearer to you from the illustration that I am going to show soon.

(Refer Slide Time: 06:53)



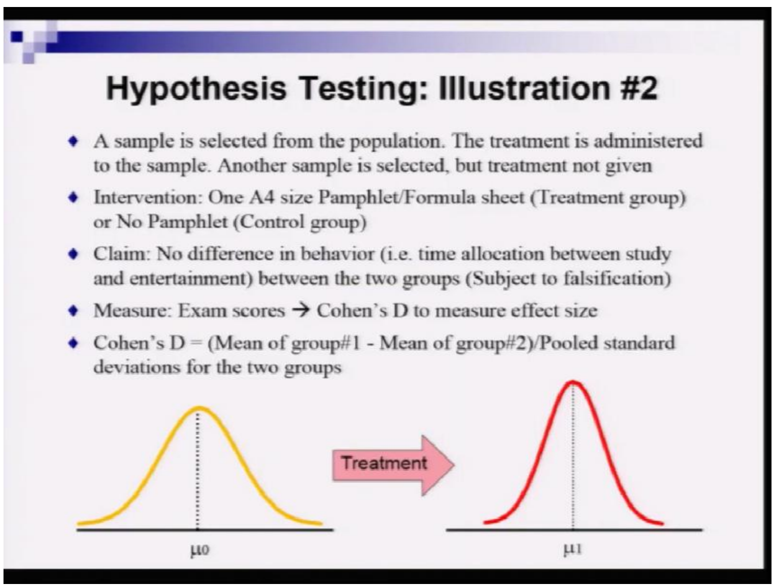
So, let us start with illustration number 1, where I am going to give you some idea or example for the definition 1 that I shared with you. So, suppose, there is a demographic projection study, which says that the median age of UP population is somewhere around 26 in 2021. Now, you assume that the median age of population of Kanpur Nagar district is 26, you make an assumption based on what demographers are saying, but you need to test that. So, maybe you can collect a random sample of the residents of this particular district and you found the median age to be 21. Of course, it is a hypothetical number.

Now, what can we say about the population median age? Is it 26? Is it less than 26? Is it above 26? So, what is it? Can we make some kind of conclusion based on these two numbers in hand? So, the diagram that I am showing you at the bottom of the slide then it becomes much more clearer. So, we start with big population universe, where the median values actually unknown and then we collect a small sample from that population.

And then median is also a sample statistic, like the mean and variance and proportion. So, median is also a sample statistic, so of course, there will be a sampling distribution for the sample median. And suppose, we have we know that sampling distribution with us. So, we know that the midpoint will be at 26, if the demographer's projection is correct. But if it is correct, then basically what we have obtained from our sample, which is 21 that is basically way apart from or way of from the number 26.

So, it is highly unlikely that we would get a sample median of a value say 21. If indeed, the population median is actually 26. But these are all done in a subjective manner. By eyeballing and making some subjective judgments. But what can we do statistically? So, that is basically the subject matter of hypothesis testing.

(Refer Slide Time: 09:27)



So, in this slide, I am going to show you the illustration, where I am going to explain you how hypothesis testing can be seen from the treatment effects angle as well. So, let us change the story slightly and as we are in the statistics course, then let me take example of students of course, it is a hypothetical story. So, let us assume that we have a population of students and there is one behavioral economics or psychology study going on, where the researcher actually wants to see the impact of providing some help during the exam on the behavior of the student, the day before exam.

So, this is basically an experiment, a social experiment. And in that experiment, the intervention is that, that a particular student will be given 1 A4 size cheat sheet or pamphlet, where the formula that are required to do well in the examination will be provided. And suppose the day before exam, you choose 2 random samples from the same population and then one sample, you say it is the treatment group and you provide that cheat sheet or a one page formula sheet to the students in that group. And in the control group, you do not provide such help to the students.

He will say to these students belonging to the treatment group that well, tomorrow's exam, I will provide you one cheat sheet or formula sheet that you can use for your exam. And for the control group, you do not say anything and then that is like a regular exam that these students are going to write. So now, you claim that the students in the treatment group who have learned that tomorrow, they will be given some A4 size cheat sheet or formula sheet, then whether their behavior is going to be impacted by this information.

So what do I mean by that? So I am talking about the time allocation between the study and entertainment. And the claim is that that there is no difference in behavior between the 2 groups. And that is of course, subject to falsification. Now, note that here, the way I am making a claim, it is basically a status quo claim. And status quo means that there is no change. Now, how do I measure? So, of course, I cannot measure how many hours they are studying and how many hours they are spending on entertainment, that is not in our control.

So, basically, what is measurable? Measurable is basically the exam scores. And then, let me tell you about the diagram that you see at the bottom part of the slide. Suppose, previously I have the test scores from the previous semester in another statistic scores and then, I see that well, the mean test or was mu naught and then, the treatment is provided to the treatment group and you see that the mean score is now mu1 and that is coming from another PDF drawn in red color.

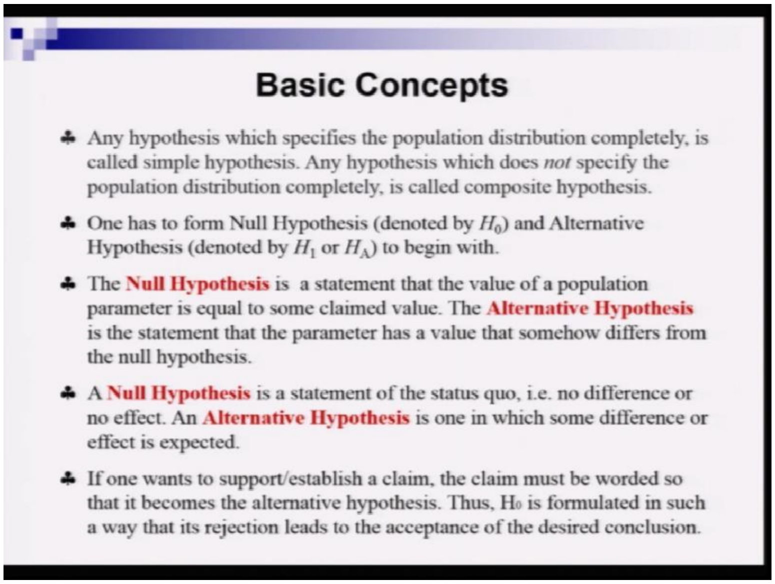
Now, you see that the variation is less. So, basically the standard deviation has gone down in this case and mu1 is also I am anticipating to be higher than mu naught, but this is all my conjecture. This is all hypothesis. So, here we are looking for a before and after comparison. So, the orange colored PDF will tell me about the before treatment story and as the sample is taken from the population, there is one control group, I assume that there is no difference in the control group from this distribution that we are observing from previous semesters statistics scores.

And but I conjectured that there will be some impact on the mean and the variance for those students who are in the treatment group who have been provided with this cheat sheet. So, I expect the mean score is going to be higher and the standard deviation is going to be less compared to the orange PDF that was there for the previous semesters performances.

So, now how to do this testing? There are two things changing as I am saying here the mean and variance. Well, there could be only one change, it could be only told in terms of the mean also, but here there is a measure which is very popular. So, this measure is called Cohen’s D. And this Cohen’s D actually measures the effect size. What do I mean by effect size? Effect size means that how large is an impact or how small is the impact of the treatment. So, that is basically in a nutshell Cohen’s D measures.

Now, what is the formula for Cohen’s D? So, here we are talking about two different groups and group 1 could be treatment group and group 2 could be the control group. And we need to compute the mean for these two groups separately and we need to take difference and then, this difference should be divided by the pooled standard deviation for the two groups and that is the way you measure Cohen’s D. And this Cohen’s D now can be used to do some statistical testing.

(Refer Slide Time: 15:48)



Now, we are going to start our discussion on hypothesis testing theory by looking at some basic concepts and these concepts are fundamental in nature. So here in this slide, I am going to define two very important concepts called null hypothesis and alternative hypothesis and you will see that these are going to be some pivotal terms or jargons, when you are dealing with hypothesis testing.

So, I will start with a very small definition on simple hypothesis and the composite hypothesis, because sometimes you will come across these terms. So, any hypothesis which specifies the population distribution completely is called a simple hypothesis. And a hypothesis which does not specify the population distribution completely is called the composite hypothesis.

So, now, I move on to the pivotal terms; null hypothesis and alternative hypothesis. One has to form this null hypothesis and that is noted by H naught and alternative hypothesis is denoted by H1 or HA. And what is a null hypothesis? So it is a statement that the value of population parameter is equal to some claimed value. And the alternative hypothesis is the statement that the parameter has a value that somehow differs from the value that have been assumed under the null hypothesis.

So, you see, that is basically the definition from the perspective 1, when I defined the first definition of hypothesis testing. So, now, I am going to revisit the definition of null hypothesis and alternative hypothesis. But this time, I am going to look at these two concepts from the perspective of a treatment effect type of point of view and that is linked to the definition number 2, what I had laid out in slide 1.

So, if you remember that we had talked about the impact of a particular treatment on a particular population. And we want to now test that. So, how to tell the same story in terms of null hypothesis and alternative hypothesis? So, if I take this treatment effect or impact evaluation kind of perspective, then a null hypothesis is a statement of the status quo, it implies that no difference or no effect actually is found after a treatment has been conducted or given to a group of subjects in a population.

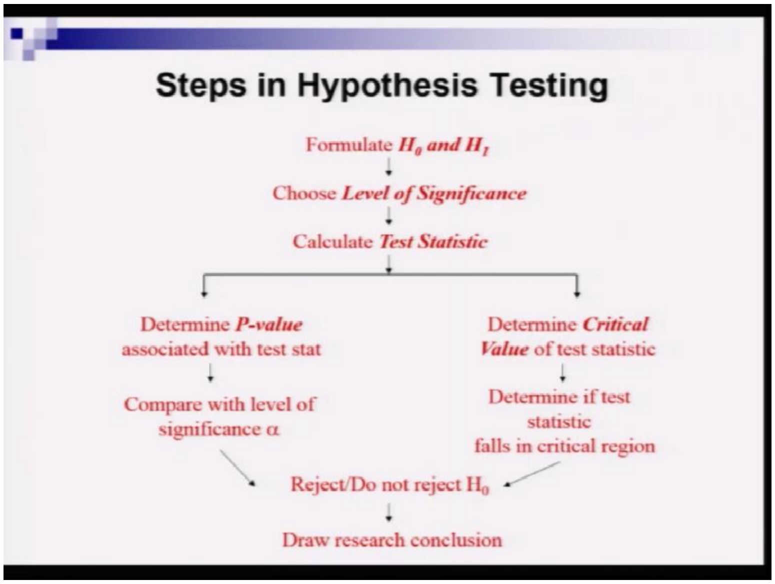
Whereas, an alternative hypothesis is one in which some difference or effect is expected. So, how to set a null hypothesis or alternative hypothesis in applied research. So, I will end this particular slide by describing or summarizing what wise men say, so many textbooks or renowned statisticians make this comment that if one wants to support or establish a claim, the claim must be worded, so that it becomes the alternative hypothesis. Thus, h naught is formulated in such a way that its rejection leads to the acceptance of the desired conclusion.

So, now, I will describe or explain the last point that we had seen in the last slide. And that is basically how to frame null and alternative hypothesis. So, let me take a very simple example from economics, microeconomics, and we have discussed demand function in some previous lecture. And then, we said that demand function is downward sloping. So if for a commodity, price of the commodity increases then quantity demanded falls and vice versa. So that is basically law of demand. And we want to actually test that out using real life data. So in that case how do I frame my null and alternative hypothesis?

So, here, note that, I want to establish the fact that demand curve indeed is downward sloping. So, if I fit a straight line demand function, the simplest possible functional form that one can assume for demand function obeying law of demand that can be written as quantity demanded q equals to alpha plus beta times p, the price of the commodity. And if, indeed law of demand exists, then I expect negatively sloped, downward sloping demand function or straight line.

So, then the coefficient of beta should be negative. So, then in that case, I can frame the alternative hypothesis first as beta is negative and then, I have to frame the null hypothesis accordingly whose rejection will lead me to the conclusion that well, indeed my demand function is downward sloped. So, then I can say, null hypothesis H naught colon beta equal to 0. So, I hope this is clear.

(Refer Slide Time: 21:10)

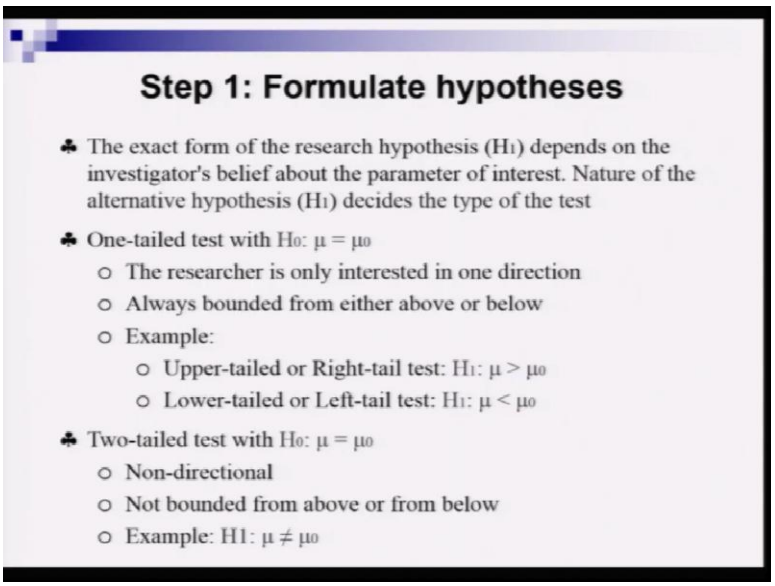


Now, in this slide, I am going to show you the 5 steps of conducting a hypothesis testing. So, you have to begin by formulating your null and alternative hypothesis in very clever way, so that, you can draw meaningful conclusions. Then, you have to choose level of significance, then you have to calculate a test statistic. And then, after that the step 4 you can do in two ways and that is one called the p value approach the other one is called traditional approach.

So, there are two different approaches and we will first say that there is this p value approach and that is associated with the test statistic and then the second approach is the critical region approach associated with the test statistic. So, if I follow the p value method, then actually you need to compare the p value with the level of significance that you have already chosen before in step 2 and that is alpha.

And in the other way, you determine whether the test statistic falls in the critical region or not. And then either approaches should help you to decide whether to reject or not to reject null hypothesis. And then, you finish your project by drawing a conclusion from the sample. Now, note that, here I am not saying that we can accept a null hypothesis, we can only reject or do not reject null hypothesis. So, please never write accept null hypothesis.

(Refer Slide Time: 22:52)



So, now, we will begin our discussion on step 1 that is how to formulate hypothesis. So, as I explained before, the exact form of the research hypothesis, H1 depends on the investigator's belief about the parameter of interest and nature of the alternative hypothesis decides the type of the test 2.

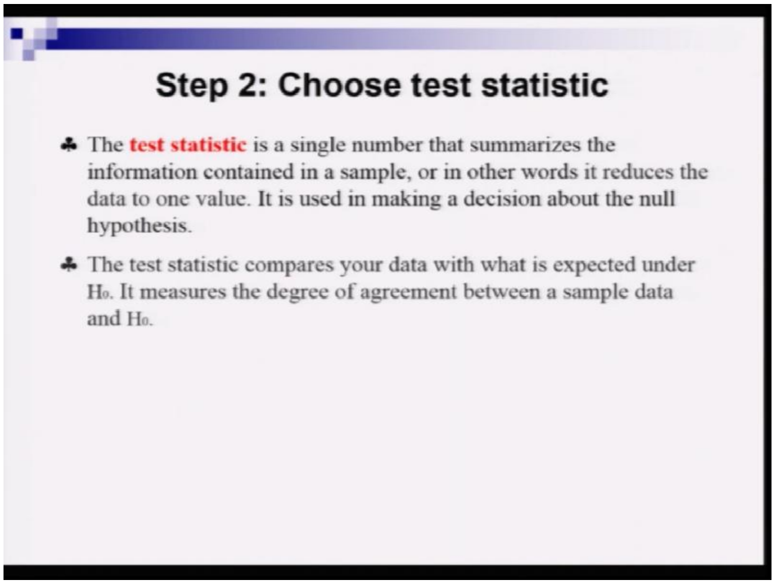
So, there are two types of tests and as I said that depends on the alternative hypothesis that one makes. So, one is called one-tailed test and the other one is called two-tailed test. So, let me first talk about one-tailed test, it begins with an all hypothesis mu equal to mu naught. So, mu is basically say the population parameter, it could be anything, it could be mean, it could be median, it could be variants. But we start with a specific value of that unknown population parameter in mind and we said mu equal to mu naught.

So then, the researcher is only interested in one direction, that is an assumption. Suppose, you are interested only in one direction and then, this is always bounded from either above or below. So again, this one-tailed test now can be of two types, one is called the upper-tailed or right-tailed test. And here, you see I set an alternative hypothesis H1 as mu greater than the particular value of parameter and that is mu naught. And then, there is lower-tailed or left-tailed test where I set the alternative hypothesis H1 as mu less than a particular value of the population parameter and that is mu naught.

So remember, in the demand function case that we have been discussing, so you can say that beta was the slope coefficient for the demand function. So, you can write a one-tailed test by framing the null hypothesis as h naught colon beta equal to 0. And then you can actually write alternative hypothesis and conduct a lower-tailed or a left-tailed test, where you can write H1 as beta less than 0. Now, we move on to the case of two-tailed test.

It starts with a null hypothesis, mu equal to mu naught as usual and this is different from the previous one because now it is non-directional and the earlier one was unidirectional and as it is non directional there is no bound from either above or from below as it was the case in the one-tailed test. So here, if we write this alternative hypothesis then we can write as mu naught equal to mu naught. So again, if we go back to the demand function example, then we can write H naught as beta naught equal to 0. So, then beta value could be either positive or could be negative.

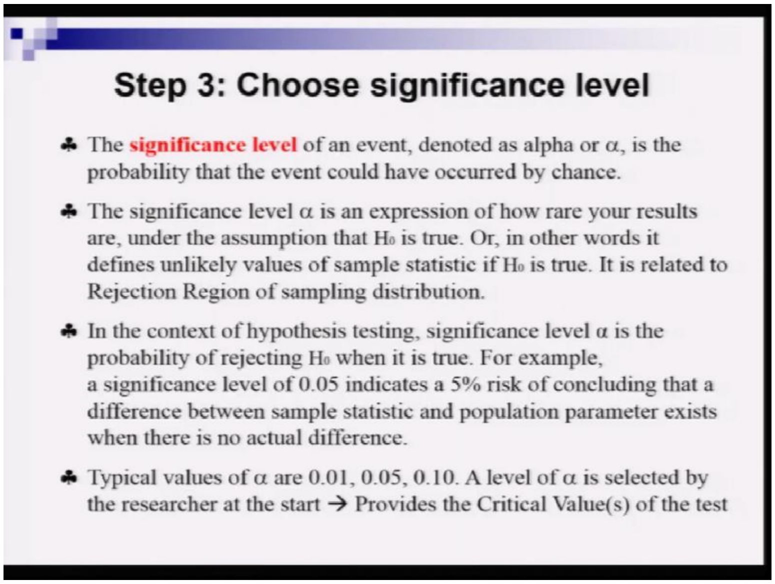
(Refer Slide Time: 26:00)



Now, quickly, we look at the step 2, which is choosing the test statistic now. This is kind of half baked in the sense that here I am not giving you example, because example actually varies from one test to the other. So, I just leave you with the definition and then when in the next class we will discuss about specific particular test then I can give you examples of test statistic.

So, what is a test statistic if someone asks me know how to define? So, then in a very simple language, a test statistic is a single number that summarizes the information content in the samples, it is more like likely that, it reduces data into one particular value and this is it on test statistic as of now.

(Refer Slide Time: 27:08)



So, next we move on to step 3 and that is basically going to talk about the selection of level of significance. Now, what is level of significance, we will start with a definition first and then, we will discuss. So, the significance level of an event is denoted by alpha or Greek letter alpha here, and that is the probability that the event could have occurred just by chance.

Now, this significance level alpha is an expression of how rare your results are under the assumption that H naught is true. So, if you remember, I started today's lecture by saying that sometimes you can conclude based on a sample just because of sampling error, because of no hypothesis or no causal explanation, it is just by fluke that you have got a sample and you have decided something and that is actually not the true situation that is actually prevailing in the population.

So, it actually defines unlikely values of the sample statistic, if H naught is true. Now, why I am now bringing the sample statistic here, because remember, in the last slide, I have defined sample statistic in the following manner. So, sample statistic actually is just representation of the entire data set through one single number. So, actually you take a call whether to reject or do not reject your null hypothesis based on the value of the sample statistic.

So, it is very important. So, actually there are likely values of sample statistic if H naught is true and there are unlikely values of sample statistic if H naught is true. So, this significance level actually is related to the rejection region of the sampling distribution. So, this thing will be much more clear in the next class when I am going to talk about particular tests with the aid of diagrams.

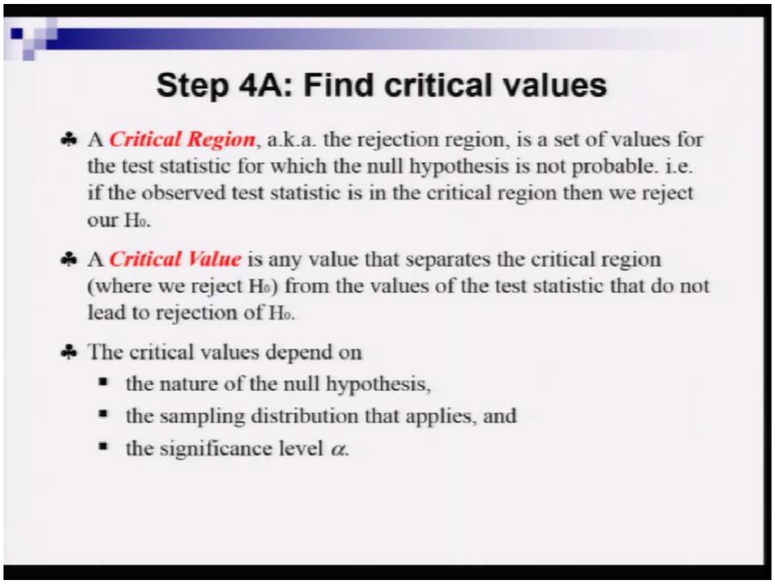
Now, in the context of hypothesis testing significance level alpha is the probability of rejecting H naught when it is indeed true. So, this is another way of describing the significance level. So, you can state it with an example. So, a significance level of 0.05 actually indicates a 5 percent risk of concluding that difference between sample statistic and the population parameter exists when there is no actual difference.

So, I repeat the previous stuff again just to make things clear for you, why we are talking about difference you know, I am emphasizing this point again and again, please remember that we are actually trying to infer about unknown population parameter value from the sample statistic. But sample statistic has sampling error or variation why because, if you compute a particular sample statistic or test statistic value from one particular sample, the next sample you draw from the population, it will give you a different sample statistic and test statistic value.

So, there is risk associated with the inference. So, from one sample you get a nice test statistic value, which says that the difference between sample statistic and population parameter is actually minimal. But you may get another sample and from that sample you can calculate another sample statistic or test statistic value such that now this tells you that there is huge gap between the sample statistic and the population parameter.

Now, this is all theory. When it comes to practice, the typical values of alpha are 0.01. 0.05, and 0.1. So, these are basically the most popular values of alpha in applied statistics and econometrics work. And you also note that you have to choose alpha before you start your empirical project. So, you cannot actually look at the data n, then choose your value of alpha so that you get nice results. So, you declare your alpha value and then, you get the data and you decide.

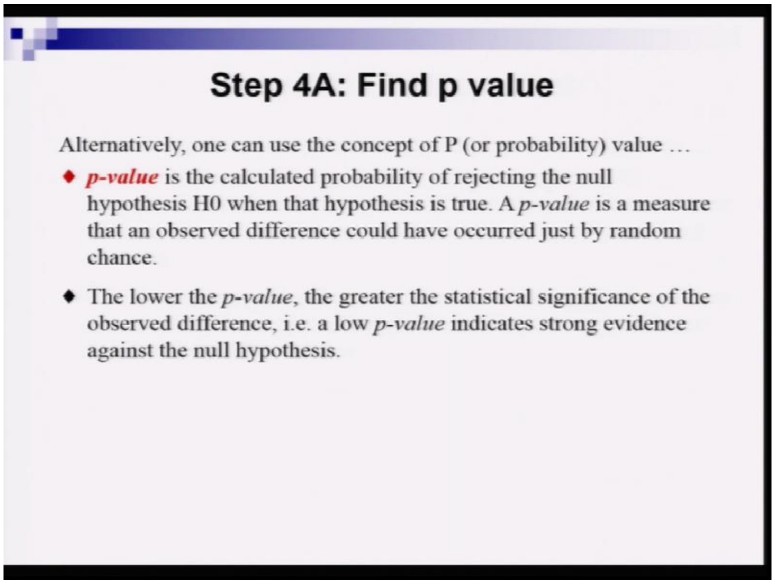
(Refer Slide Time: 31:24)



So, now, I am going to talk about the traditional method that one can adopt in step 4 and I will again, start with two definition, because it is a mostly theory based lecture. So, we start with critical region, what is it? So, the critical region or the rejection region actually is a set of values for the test statistic for which the null hypothesis is not probable. It implies, if the observed test statistic is in the critical region, then we can reject our null hypothesis.

And what is critical value? A critical value is any value that separates the critical region where, if the test statistic value falls in that region then we reject our null hypothesis from the values of the test statistic that do not lead to rejection of H naught. So here, the critical values actually depend on 3 different things; 1, the nature of the null hypothesis; then the sampling distribution that applies to a specific case and then the significance level of alpha. So, I think, again, you have to wait for some time and in the next lecture, when I am going to discuss about particular tests with diagrams and all, this thing is going to be clear.

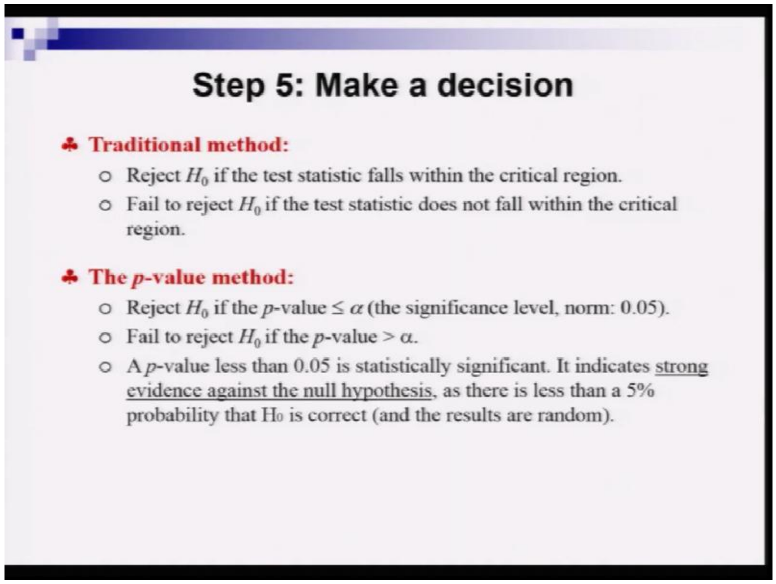
(Refer Slide Time: 32:53)



Now, if I follow the p value approach, which is basically in step 4 or the alternative way to conduct step 4, then we must first know what is p value and then we can adopt this. So, what is p value or probability value? So, it is the calculated probability of rejecting the null hypothesis h naught, when that null hypothesis is actually true. So, a p value is a measure that an observed difference could have occurred just by random chance.

So, what difference we are talking about here? So, we are talking about the difference between the sample statistic and the true population parameter value. Now, note that, lower the p value, the greater the statistical significance of that observed difference between sample statistic and population parameter. So, a low p value indicates strong evidence against the null hypothesis.

(Refer Slide Time: 34:06)



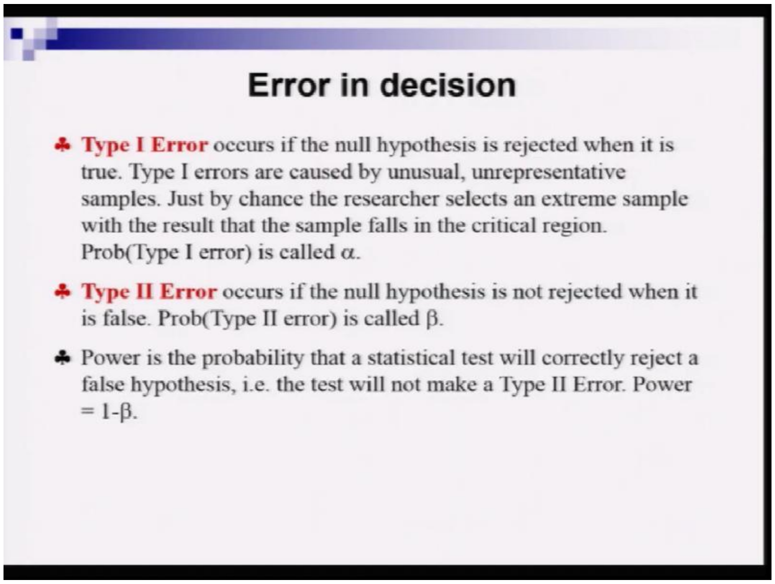
So, once we are done with steps 1 to 4, now it is time to make decision. So, again, based on two different approaches, we can make two types of decision rules and let us have a look at them one by one. So, first I will talk about the traditional method, here you can reject null hypothesis if the test statistic falls within the critical region. And you fail to reject null hypothesis if the test statistic does not fall within that critical region.

And next, we move on to the p value method. So, here we can reject null hypothesis p value is less than the value of alpha, the significance level and the norm in applied statistics and econometrics work is to choose the value 0.05 for alpha. And you fail to reject null hypothesis if p value is greater than your choice of significance level, which is most times 0.05.

So a p value less than 0.05 is statistically significant. We can say by following the statistical jargon and it indicates strong evidence against the null hypothesis. So, there is less than 5 percent probability that H naught is indeed correct and the results actually are random, results mean that what we observe from the sample.

So, after you make a particular decision in a particular hypothesis testing exercise, then there could be only two decisions, either you could be correct or you could be wrong. But again you do not know, because there is sampling error involved, whatever decision you make that has some associated probability of being correct or being wrong. So, in statistics, the statisticians talk about two types of errors type I error and type II error and then there is also a concept called power of a test and next we are going to talk about these concepts.

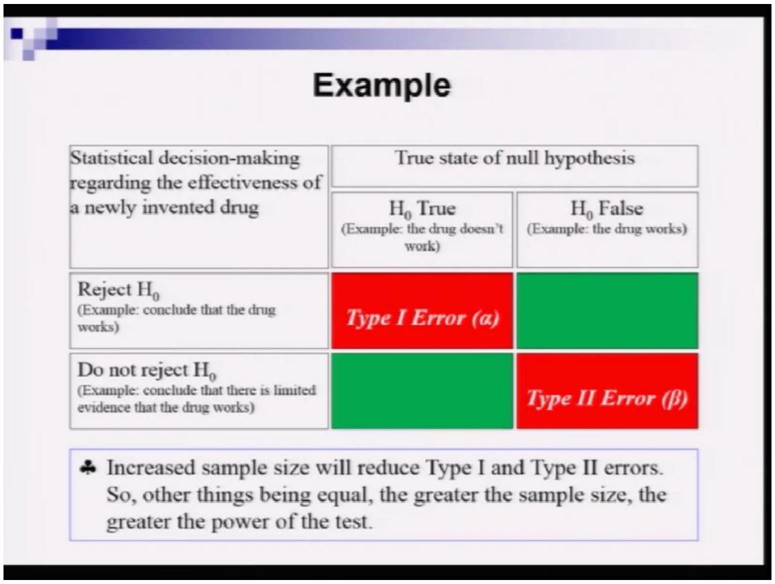
(Refer Slide Time: 36:01)



So, I will begin with formal definition for type I error. Now, this occurs if the null hypothesis is rejected when it is indeed true. So, type I errors are caused by unusual, unrepresentative samples. Just by chance the researcher selects an extreme sample with the result that the sample falls in the critical or the rejection region and the probability of type I error is denoted by that level of significance that is alpha.

Now, what is type II error? So, it occurs when the null hypothesis is not rejected when it is false. And probability of type II error is denoted by the Greek symbol beta. Now, power is a concept which is associated with the concept of type II error. So, it is defined as the probability that a statistical test will correctly reject our false hypothesis, it implies the test will not make a type II error. So, power is given by 1 minus beta.

(Refer Slide Time: 38:19)



So, after visiting the formal definitions in sentences on type I and type II errors, let us talk about them through an example. And the example, let me talk about this ongoing event, say there is some pandemic and you can say the COVID and then different medical research organizations they are trying to come up with quality vaccines to prevent us from the pandemic. And when research and development takes place in the lab and scientists are developing new vaccines, it is an experiment.

So, the success of that experiment is a random event. So, there could be even failure. So, now, if a vaccine is proposed, now that vaccine has to pass through different trials, and then only one can conclude whether the vaccine actually works or not. So, based on this context, I am showing you this example, please, look at these tables that I am showing you first with several cells, colored cells we will explain what they are.

So, we concentrate on a statistical decision making problem regarding the effectiveness of a newly invented drug or vaccine, whatever you want to call it. And note that the true state of null hypothesis are of two types. So, either H naught could be true or H naught could be false. So, H naught to true means that the drug actually does not have any effectiveness, so the drug does not work. So, that is basically describing the case of status quo.

And the H naught false case means that actually the drug is effective. So, drug has impact and so, we can say that there is some treatment impact of the drug or the vaccine. Similarly, the decisions at hand are of two types, either you reject the null hypothesis or you do not reject the null hypothesis. So, what would be the case of rejecting null hypothesis that you conclude based on the sample evidence that the drug actually works. So, the drug is effective. So, in that case you reject the null hypothesis.

And what could be an example of do not reject the null hypothesis in this context. So, if you find sample evidence, then you may also conclude that there is very limited or insufficient statistical evidence that the drug actually works. So, now, depending on the true state of the null hypothesis and your decision either reject or do not reject, there could be 4 different cases that are possible.

So, now, we are going to look at them one by one. So, let us consider the first case when H naught is indeed true, but you have rejected null hypothesis and this is and that is given by the red color box and you see I have written type I error there, I have defined type I error previously and the probability of type I error is alpha. So, hopefully now, you can match the definition that I spoke about in the previous slide with this example or case.

Now, the next box just left to it is green and green means that there is no problem that is no standard convention. And of course, it is not a problem because actually H naught is false and you have rejected the null hypothesis. So, what could be better than that? So, actually the reality is that the drug or vaccine is effective, it works and you have rejected the null hypothesis that means that you are concluding that the drug works. So nothing could be better than that.

Now, we come to the case of H naught is true. So, the drug does not work, but you do not reject null hypothesis. So, you say that there is limited evidence that the drug works. So here, also same situation you are happy because, okay, anyway, the drug does not work and you found enough statistical evidence from the sample that that drug does not work. So you do not reject your null hypothesis. So that is also a very good scenario.

But what about the fourth box? And that is again in red, so that is another problem area, where actual situation or the reality of the null hypothesis is that that H naught is false. So, the drug actually works. But you now do not reject the null hypothesis. So, you conclude that from the data, you find very little statistical evidence that the drug works. So, this problematic area because you are drawing wrong conclusion based on the sample is called a type II error and the associated probability is given by beta.

So, what could be a remedy of these type I and type II errors? So, can we do something to put a check on the probabilities alpha and beta? Yes, of course, the remedy is very simple, but it is costly. So, you have to increase the sample size. So, increased sample size will reduce the probability of type I and type II errors. So, other things being equal, the greater the sample size, the greater the power of the test.

So, I know that today's discussion was a bit theoretical and then, I have showed you definitions one after another and probably some concepts were not very clear. But I just wanted to put every theoretical concept in one particular lecture and save the discussion on the applications of these concepts in another lecture. So, please come back for the next lecture, where I am going to actually explain you these concepts that I defined in lecture through particular tests. Like, we are going to show you how to conduct a z test, how to conduct a chi-square test. And till then, bye.