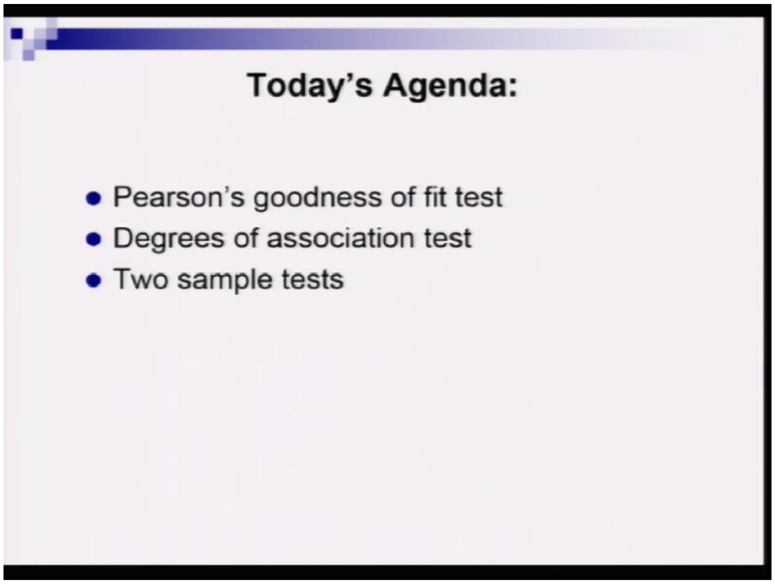
Applied Statistics and Econometrics
Professor Deep Mukherjee
Department of Economic Sciences
Indian Institute of Technology, Kanpur
Lecture-13
Hypothesis Testing (Part III)

Welcome back to the lecture series on Applied Statistics and Econometrics. So, we have been discussing the case of hypothesis testing techniques. And last time we could not finish the Pearson Goodness of Fit Test. So today, we are going to start with that concept and we are also going to talk about a lot of new tests.

(Refer Slide Time: 00:36)

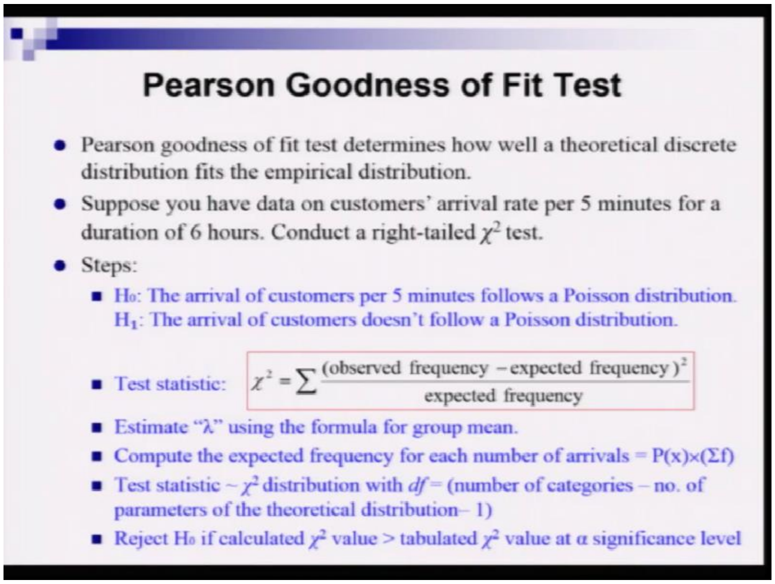


So, here is the agenda items for today's lecture. So, after the Pearson's Goodness of Fit Test, we are going to talk about another chi-square test and that is called degrees of association or test of independence. And then, we are going to finally talk about 2 sample tests.

Many times, an applied researcher may be interested to see, if he or she tries to fit a model based on the frequency observations from the real-life data, then how good or bad that model approximation is doing. So, let me give you an example, suppose you have some discrete variable and you think that from your textbook knowledge, you feel like either a binomial or poison distribution probably would feed the data well. But that is a theoretical distribution and real life data is noisy data.

So, when you try to fit a particular theoretical distribution, that is basically you are trying to model the real life observations, it may be possible that the distributional assumption that you are making in this case either binomial or poisson or it can be anything normal uniform or gamma, whatever, you name it.

(Refer Slide Time: 01:55)



So, it may be the case that the data that you have at hand may not come from that particular theoretical distribution that you have assumed. And hence, there has to be check for goodness of fit. Now, there are 2 types of goodness of fit tests, 1 is for the discrete distributions and the other 1 is for continuous distributions. So, when you are trying to fit a discrete distribution like binomial or poisson to real life frequency observations, then you make use of the Pearson's chi-square goodness of fit test.

And if you are trying to fit a distribution on continuous data, if you are trying to fit a continuous distribution on relative frequency, then basically you can make use of Kolmogorov–Smirnov Anderson–Darling kind of tests. But we are not going to see the Kolmogorov–Smirnov or Anderson–Darling test, because I have not taught the related concepts for those 2 tests. So, I will skip those 2, but I am going to talk at length about the Pearson's Goodness of Fit Test.

So, yes, if you are guessing that is this test somehow connected to the father figure of mathematical statistics, Karl Pearson, you are absolutely right. So, these tests was proposed by Karl Pearson, back in 1900. And actually, in connection to this test only, he invented the idea of chi-square distribution.

So, let us have a look at Pearson's Goodness of Fit Test. So, suppose, I give you a data set on customer's arrival at some store or some bank or it can be arrival of ships in a port. And you have this data set at every fifth minute. So, you have data, arrival data for objects or customers per 5 minutes. And then, you have these data sets available for say some 5, 6 hours.

Now, you know that from statistical textbooks, you can guess that, well, in these type of data sets, maybe a poisson distribution could give me very good fit. But you never be sure, you have to try to model this data by assuming a poisson distribution and then, match the fitted values of your model to the observed values in the data set. And then, if they are close enough, then only you can say that, well your assumption of poisson distribution that you want to apply as a model to this dataset is the correct one. So how to conduct a test for that? And here, you have to apply a right-tailed chi-square test.

Now here in this slide, I am showing you the different steps. Of course, in the very first step, you have to set the null hypothesis and the alternative hypothesis. So, let me show you the steps assuming the story that I started with. So, it is basically arrival of customers in a shop or in a bank. So, the arrival of customers per 5 minutes follows a poisson distribution that is the hypothesis you make.

And then, the alternative is that no, the arrival of customers does not follow a poisson distribution. So here, in this case, the test statistic is given by this formula in the red box here. So that is chi-square and that is basically sum of the square of deviation of the expected frequency from the observed frequency divided by expected frequency, it looks a little bit complicated, but calculation wise, it is not that complicated. I am going to show you how you can get these things done.

So, note that, your observed frequency actually is the data that you have and the expected frequency actually is coming from the model. So, here, in this case, the task at hand is how to come up with the expected frequency. And this expected frequency will come from some assumption regarding the probability distribution of the random variable.

So, if you are assuming poisson here in this case, you adopt a poisson PMF and then, you try to calculate the probabilities from the poisson PMF and then, by multiplying that probability, with the total frequency or the total number of observations in the data, you can get the expected frequency.

So, let me talk about the steps in detail. So, if you want to calculate the poisson probability, what comes to your mind? You have to know the parameter value for the poisson distribution, is not it? So, for a poisson distribution, we have to know the value of lambda. So, lambda is the mean of the poisson random variable.

So how can we get a proxy for lambda? So, you need to estimate lambda by using the formula of group mean, right. So, you have the data, suppose x is basically the variable and then f is basically the corresponding frequencies. And then, by group mean, I mean that you have to calculate x bar equal to summation xi fi divided by summation fi.

So now you got your lambda and then you apply this lambda parameter proxy value in the poisson PMF formula and then you can calculate the poisson probabilities for different values of x. Once they are computed, you can compute or calculate the expected frequency for each number of arrivals.

So, for that, actually, you have to multiply the total number of cases, which is sum of all frequencies with the corresponding probability value. So, if your x is taking values like 0, 1, 2, 3, 4, so you have to multiply probability of x equal to 1 with that sum of f, then you have to get the probability of x equal to 1 and you need to multiply that with some of f. And that way, you can generate the expected frequencies for values x equal to 0 and x equal to 1, respectively.

So now, the next step is to get the test statistic value and this test statistic value, I have shown you, the formula in the red box. So now you have got your expected frequencies.

And then, you need to basically subtract this expected frequency from the observed frequency square them and then divide this difference by the expected frequency and then, you have to sum. And by this way, you can calculate the chi-square test statistic.

Now, if you remember, that chi-square distribution always works with degrees of freedom. And here, let us see what degrees of freedom we have to assume. So, here, the relevant degrees of freedom will be the number of categories. So, if you have intervals of access, then you can as you how many intervals do you have or if you just have the discrete numbers like 0, 1, 2, 3, then you have to basically have to count how many numbers do you have and then, you have to subtract the number of parameters of the theoretical distribution that you are fitting minus 1.

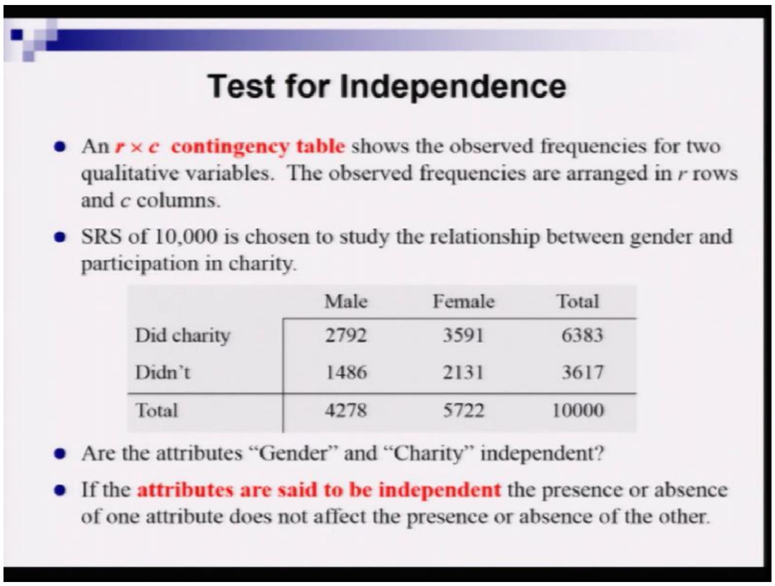
So here in this case, as you were trying to fit a poisson, poisson has only 1 parameter lambda. So basically, here in this case, it will be number of categories minus 2, that is going to be the degrees of freedom. So once the degrees of freedom is determined, now we have to set the decision rule. And note that, the decision rule that we are going to present here is not very different from what we have seen earlier.

So here, the decision rule can be laid out as you reject null hypothesis if the calculated chi-square value, which is basically the test statistic value is greater than the tabulated chi-square value at alpha significance level. So, of course, needless to say that you have to fix the alpha value at some numbers 0.05 or 0.01 and then, you go to the chi-square table and then basically, you match these statistic value and the tabulated value. So, that was the case of Pearson Goodness of Fit Test.

And now, we are going to talk about another interesting test. Remember, at the very beginning, I talked about organization of categorical data or qualitative data and then, I briefly introduced the concept of a contingency table to you in that matrix form, if you remember, I have 2 attributes and the attributes can have different levels; and then, in the sales you have the frequencies. So, that was basically the structure of the contingency table.

Now, one may ask a very interesting question. So, if I am dealing with 2 characteristics, which are qualitative in nature, so, these are attributes in other words, are they somewhere related? So, is there any association between these 2 qualitative features of the population? And that question can be addressed by another chi-square test. And this is called the degrees of association test. That is what we are going to study next.

(Refer Slide Time: 11:23)



So, let me assume that in the general setup, the most general case, we have an r cross c contingency table and basically, that shows the observed frequencies for 2 qualitative variables or attributes. And you can name them a1 and a2 and the observed frequencies are arranged in the sales.

Now, the best way to understand this test for independence or association is through a hypothetical example. Let me talk about in our hypothesis. Many times, social scientists say that gender actually is associated with person's kindness and kindness actually means the behavior regarding charity, making donations or the sensitivity towards environmental quality. So, these are basically all related with the gender attribute. Now, if someone wants to test that out from real life data then how to go about that. So, this example that I am going to present here is going to talk about 1 such problem.

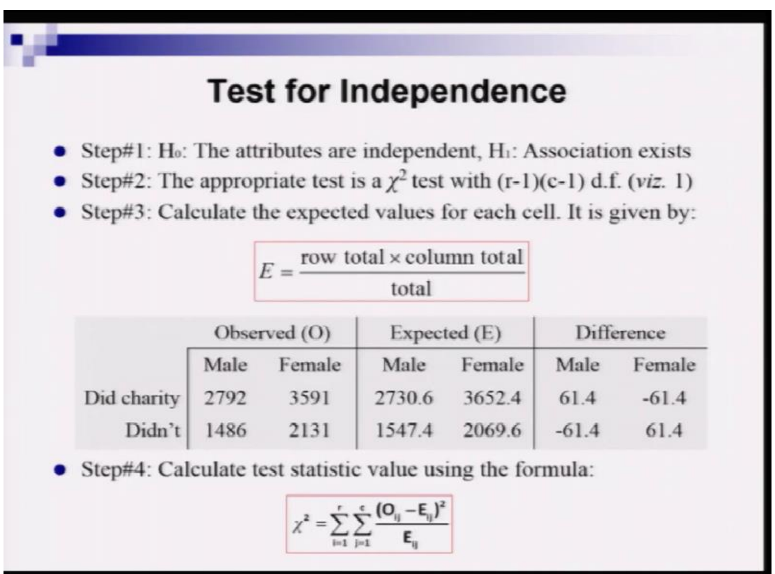
So, let us assume that there is a social scientist who wants to study, whether gender has any association with the participation in charity activities. So, to test that hypothesis or to study this association between these 2 attributes gender and charity behavior, a simple random sample of 10,000 is chosen and the frequency is arranged in a contingency table.

So, here, you see that the first 2 columns are talking about 2 levels of the attribute gender, and they are male and female. And then along the rows, I present another attribute, which is charity behavior. And again, 2 levels are there, the person did not charity or the person did some charity. Okay?

So here, the sales are showing the frequencies and you see that total is 10,000. Of course, it is a very simple example of a contingency table. But we do not have to make it complicated to understand even a simple example of the sort will be enough.

So here, the question at hand is that are the attributes gender and charity independent? So, we are talking about independence and association. So, let us briefly define what do I mean by independence. So, if the attributes are said to be independent then the presence or absence of 1 attribute does not affect the presence or absence of the other attribute, that is basically in a nutshell independence means in the context of qualitative variables.

(Refer Slide Time: 14:27)



So, now, we are going to look at the steps of conducting a chi-square degree of association test. So, needless to say that you have to frame the null and alternative hypothesis at the very first step and here we set our null hypothesis h naught as the attributes are independent. So basically, they are status quo and then, this null hypothesis should be tested against the alternative hypothesis h1, which says that there is an association between these 2 attributes.

So, in step 2, what we are going to do we have to, we know theoretical result and this is going to be a chi-square test. So, in this case, the chi-square test statistic is going to follow the chi-square distribution with r minus 1 times c minus 1 degrees of freedom. So, in this case, we have 2 rows and 2 columns. So then basically you are talking about degrees of freedom 1. So, it is a simple case.

Now in step 3, we have to calculate the expected values for each cell, as we did for the previous test also. But here, the complication is that, well, it is not a complication. Here, we do not have to assume a theoretical distribution to compute the expected value. So that is good news.

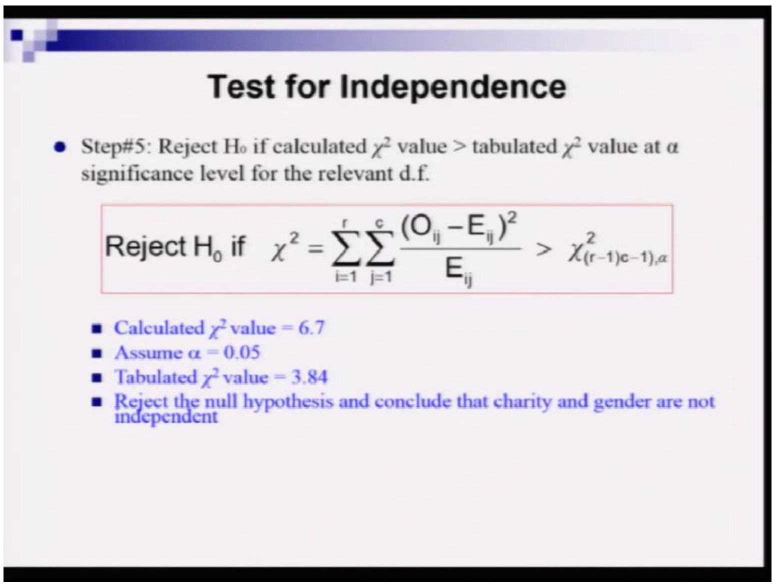
So, how to actually get the expected values? So here, we have to use this formula that I am showing you here in this red box. And that is basically you have to multiply the row total times column total and you need to divide it by the total number of observations and that is the way you can get the expected frequencies.

So here, the below table, I am showing you how the expected frequencies are matched against the observed frequency. So, here, you see the observed frequencies are denoted by big Okay and these are the numbers with which we started and then expected frequency is denoted by capital E. And we are using the formula in the box and we are getting this expected frequency numbers. Note that, as these are calculated numbers, they can be fractions. So, do not you worry about that.

Now, you need to take the difference, because if you remember, last time also, we spoke about a difference between the observed frequency and the calculated frequency. So, here also you need to take the difference and the difference is shown here in the third part of the table. But you see here, you will observe 61.4 as the common value in all 4 cells with plus minus signs reading. Now, this is just by chance, there is no statistical reason why they should be equal, they are just by fluke.

So, once the differences are calculated, you calculate the test statistic by using this following formula, it is slightly complicated compared to the previous case, because here you have both rows and columns, so you need to actually use the double sum formula. So, once you calculate the test statistic value, then basically, the next step should be to set your decision rule.

(Refer Slide Time: 17:35)



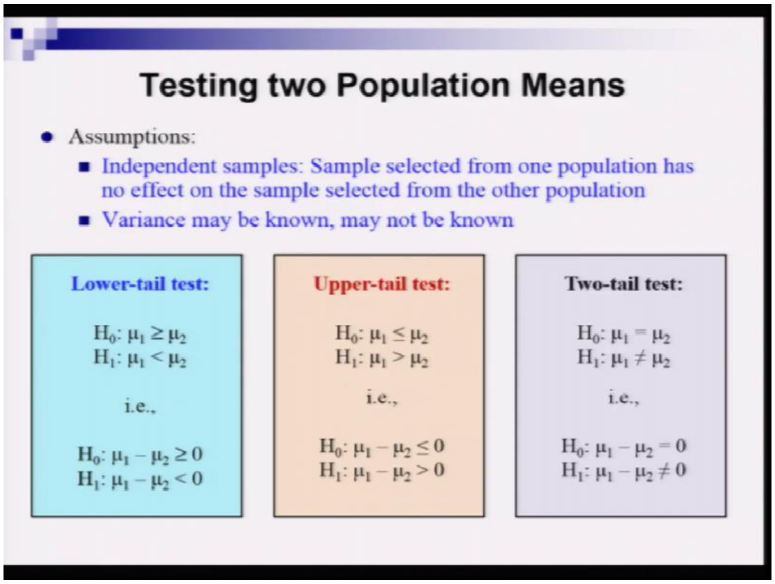
So, let us set the decision rule; and here, we are going to follow the traditional method or approach of hypothesis testing. So, you reject the null hypothesis, if calculated chi-square statistic value is greater than the tabulated chi-square value at alpha significance level for the relevant degrees of freedom.

So here, I am again showing you this decision rule in the symbolic form in this box and let us actually go back to the example that we have. So, given the contingency table that we have here, we have a calculated chi-square value of 6.7. Now, let us assume we are talking about level of significance alpha set at 0.05 following the standard norm. And then, for this alpha 0.05 and degrees of freedom v equal to 1 we have to go to the chi-square table and we actually find out the tablet chi-square value being 3.84.

So, as the calculated chi-square value is higher than the tabulated chi -square value, we reject the null hypothesis and conclude that charity and gender are not independent. So, there is some association between these 2 attributes. So, interestingly, note that, this particular test will not tell us what is the direction of the association. So, here we cannot draw certain inferences like whether the female gender has a positive impact on the charitable behavior or sensitivity towards the environment or poor people or whatever. We just know that they are related.

So far, we have talked about 1 sample tests. So, we had 1 sample at hand and we want to draw some inferences based on that sample about the population characteristics and then, whether we succeed or we cannot succeed that was the subject matter of this hypothesis testing exercises.

(Refer Slide Time: 19:56)



But now, what if, if you have 2 samples? So now we are going to study some interesting tests based on the tests that we have done previously? What will happen if we want to draw inferences by comparing 2 different random samples? So first, we are going to talk about testing 2 population means, because mean is probably the most important characteristic about any data set. And that is the first thing that comes to anybody’s mind. So, we will start our discussion by talking about means.

So here, we have to make 2 assumptions and these are the assumptions. We have to assume that the samples selected from 1 population has no effect on the sample selected from the other population. So, that means that the samples are completely independent.

Now, the second assumption says that the variance of the population may be known or may not be known. So, the random variable that we are generating for that the population variance is no not known to us. So, there are different testing procedures to solve this particular problem, variance is known then we will do typical z test because that is what we have done previously also. But if we do not know the value of variances then we will land up doing the t test and that is also not new to you, we have spoken about t tests.

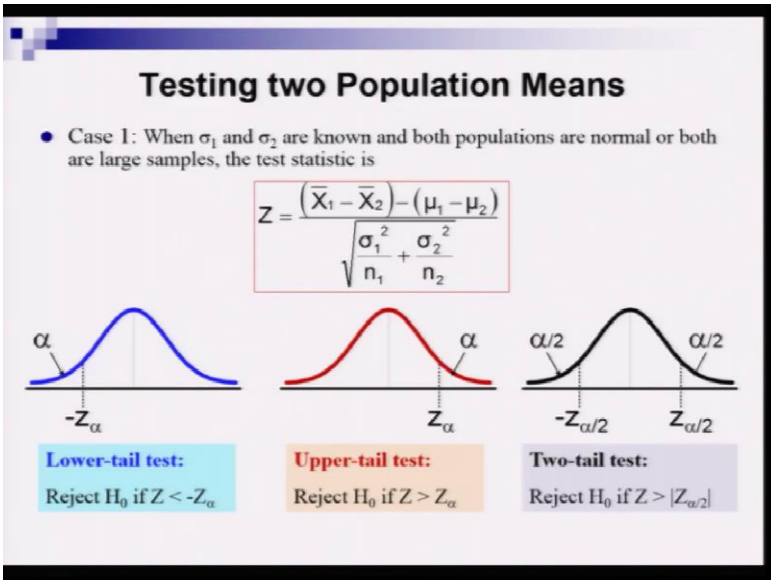
So as usual, the null hypothesis and alternative hypothesis are to be framed first. And based on the nature of the alternative hypothesis, we can have 3 different types of tests, one is the left-tail or lower-tail test. And here, you can see that I assumed that my mu1, the population mean from population 1 is greater than equal to the population mean from population 2. And my alternative hypothesis says that no, mu2 is actually higher than mu1.

And then, these complicated null and alternative hypotheses can be simplified and here you see that I have written it in such a way that I get 0 as the value for reference in both my null hypothesis and the alternative hypothesis. And as the alternative hypothesis has the less than sign, then we know we are talking about the left-tail or the lower-tail of the distribution.

So here, let us move to the second box. And that says that we know we have to conduct upper-tail test. Now, what is an upper-tail test in this context? We have to follow the same discussion that we did for the first box, which is the lower-tail test. Only note that here the direction of the inequality for the null and alternative hypothesis are different here. So, but, you can write, rewrite these null and alternative hypothesis following the similar manner by putting 0 in the right hand side. And as the alternative hypothesis has the inequality greater than inequality sign, then we know that we have to conduct an upper-tail test here.

And let us move to the third and last box that talks about two-tail tests. So, if I now say that mu1 is not equal to mu2, but we do not know the direction in that case, you have to conduct a two-tail test. And again, I am showing you, how you can rewrite the null and alternative hypothesis, so that you can get 0 value in the right-hand side for both your null and alternative hypothesis.

(Refer Slide Time: 23:52)



So, here, we are going to talk about 2 different cases for 2 sample population mean testing; and these 2 cases are when the population variances are known and the case number 2 would be when the population variances are unknown. So, now, let us start with case 1, where sigma 1 and sigma 2 the population standard deviations are known and both populations are normal.

So, in other words, you can say that you are dealing with large samples, so, that is the case of n greater than 30. So, in that case, you first write the null hypothesis and alternative hypothesis and after that, the next task is to come up with the test statistic. So, you can choose any of the 3 boxes in the previous slide and write your null and alternative by following one particular box.

But here, note that, the test statistic is going to be very complicated. And here, you see that the z statistic we are going to use, so, this is basically the difference between the difference of sample means, and the population means, and that grand difference should be divided by the square root of the sigma square 1 divided by n1 plus sigma square 2 divided by n2. It is a complicated formula, but, this is very useful in applied social science research. So, that is why I am showing you this formula.

And now, let us talk about 3 different cases that you can encounter under this case. So, of course, it depends on your particular way of thinking. So, if you have thought about lower-tail test, then I am showing you the diagram here, which is basically at the left hand side.

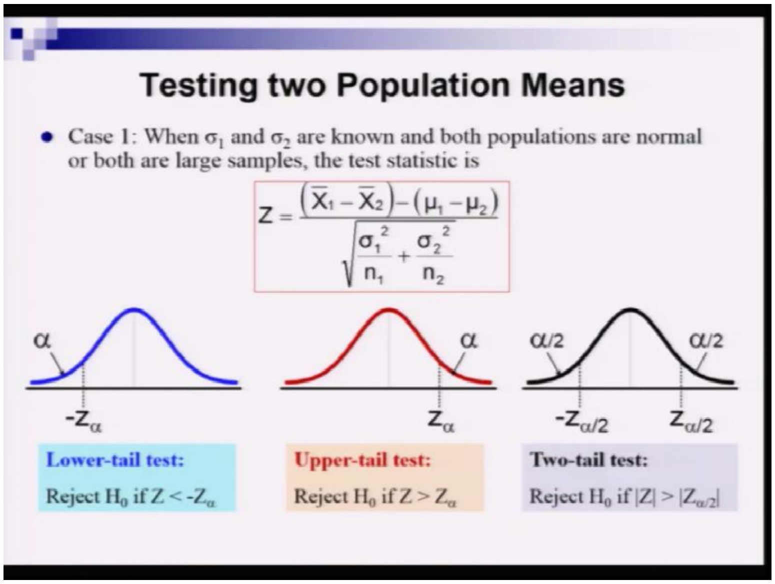
And then, basically here do you see the z distribution you have to conduct a test here, of course, so, the z distribution or standard normal curve is given by this blue bell-shaped curve. And as it is a lower-tail test, you have to basically figure out the critical value from the table which is minus z alpha for a chosen alpha.

And then, rejection rule is, simple it is like, previously we have stated that reject your null hypothesis if z is less than minus of z alpha, okay. So, basically what do I mean to say here, if you want to use plain simple English language? So, basically here you have encountered 1 particular value of test statistic which is very far from the center of your sampling distribution.

So here, you can assume that you have encountered 1 particular value of z so off from the center of your z distribution, so that you can make this inference that your sample statistic value on the difference of sample means, is so high that the difference between the population means cannot be 0. So that is you are going to conclude from running this test.

So next we move on to the second case, which is the upper-tail test, the story remains the same here, only the direction of the alpha and the critical value will be changed. And you see this is the exact mirror image of the previous case. So, I do not want to spend a lot of words here. And we have spoken about one-tailed test before, so I leave it up to you.

(Refer Slide Time: 27:32)

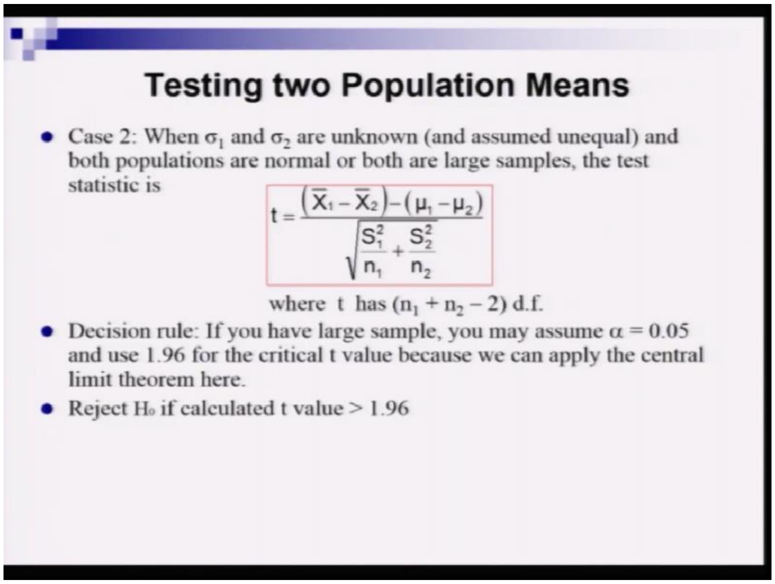


So now, we are going to look at the third case of two-tailed test. So here, look at the diagram placed in the extreme right hand side of the slide. So here, as usual, I am going to show you 2 rejection regions because we are dealing with two-tailed test. And the area that is split into 2 equal parts, so that is why I am showing you these probability areas alpha by 2, alpha by 2 in 2 directions, 1 in lower-tail and 1 in the upper-tail. And what would be the decision rule? How can I reject my null hypothesis or how can I fail to reject my null hypothesis?

So here is the rule. So, as here in this third case, you may end up getting both plus and minus values of or positive or negative values of z. So, you have to better deal with the modulus or at the absolute value of the z and not only the tabulated value, but also, the calculated z value.

So, you take the absolute value of the calculated z and you take the absolute value of the critical value z and then, you compare so if the absolute value of the calculated test statistic is higher than the absolute value of the critical value, then you reject the null hypothesis.

(Refer Slide Time: 29:00)



So now, we are going to look at the case 2 of comparing population means. So, in case 2, we have sigma 1 and sigma 2 unknown. But let us assume that they are unequal and both the populations are normal. Or at least both the samples are large samples.

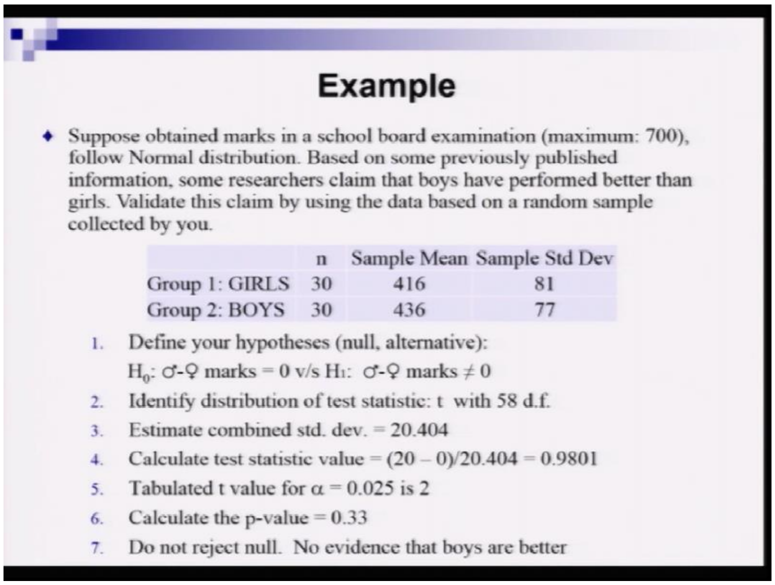
So, in this case, the test statistic can be defined as t and that is basically a very complicated expression. But note that it is not that difficult. So, the first component x bar, 1 minus x bar 2 actually gives you the difference in the sample mean. And then, you have to subtract that difference in mu1 and mu2.

So, if you start with an assumption mu1 is equal to mu2, then of course, this will boil down to 0. And then, in the denominator, you see that we have the sample variance of sample 1 and then, that is divided by n1 and then you need to add the ratio of sample variance from the sample 2 divided by the sample size of sample 2. So that is the way you get a combined measure of the variance and standard deviation.

And here, the t statistic will follow the t distribution with n 1 plus n2 minus 2 degrees of freedom. And why, we have to deduct 2 here? Because note that for each sample, you have to compute the sample variance. So for that, actually, you have to first compute the sample means for 2 groups, that is why you have to deduct 2.

Now, what is the decision rule in this case? So, if you have large sample, you may assume alpha equal to 0.05, as that is the standard norm. And then, use 1.96 for the critical t value, because we can here apply the central limit theorem. And then, you reject the null hypothesis, if the calculated t value is greater than the magic figure 1.96.

(Refer Slide Time: 31:18)



Now, we are going to look at an example, how we can actually apply these testing procedures in real life problems. So, we assume that the obtain marks in the school board examination of 700 marks follows normal distribution. And suppose that there is some previously published information.

Here, what could be the information? Suppose the board when it declares results, it also publishes the percentage of the students who have passed in first division or overall pass; and then, they know they give the gender breakup also. So, they separately report the percentage figures for boys and girls. And suppose from these statistics or the available figures, it is shown that the percentage figures a bit higher for the boys compared to the girls.

So, a researcher can claim that boys have performed better than the girls. So, is that actually the true situation? So, you need to validate this claim by using hypothesis testing? And how do you proceed? So, you, of course, do not know about the total population, you do not have entire population data. So, you decided to pick random samples and then, you visited couple of schools and collected some results by talking to the authorities in the school. And here, is the result in front of you.

So, you see, for each group, we have n equal to 30. So, for the group 1, we see that sample mean score is 416. And for boys group or group number 2, the sample mean marks is 436. And of course, you are also reporting the sample standard deviation that we can calculate from the samples that we have collected.

Now here, I am going to show you step by step story how 1 can conduct a hypothesis testing to compare the means. So, first, you have to define your hypothesis null and alternative, both. So, you start with the null hypothesis, where you assume that there is no difference between the mean of the girls population and the boys population, when it comes to the total obtain marks.

So, basically, you say that the difference between the mean marks for boys and girls is 0 and alternative hypothesis you set, so that you can conduct a two-tailed test. So, you assume that the difference in mean marks is not equal to 0. So, that is your alternative.

So, what would be the next step? You have to identify the distribution of the test statistic. So here, as you see that I am dealing with n equal to 30. So, you can assume that okay, this is large sample. So here, also you see the standard deviations that I have calculated for group 1 and group 2, they are slightly different 81 and 77 for 2 groups. So, basically we are in case 2 that we have discussed just in the last slide.

So then, we have to actually write the t statistic and then, this will follow the t distribution with 58 degrees of freedom. Why 58? Because 30 plus 30 minus 2. So, in step 2, you have to estimate the combined standard deviation and if you compute by using the formula I have shown you in the previous slide, you are going to get say 20.404. And then you are all set to calculate the test statistic value and then that will basically give you 0.9801.

And now, it is time to figure out the tabulated t value for alpha value of 0.025. Why it is so? Because, here I am assuming the level of significance to be 5 percent as it is a two-tailed test, you need to split the 0.05 into 2 equal parts. And hence, in each tail you see alpha value equal to 0.025.

And then, you consult the t table and you get the value it is approximately 2. And if someone is interested to calculate the p value that can also be calculated by using software or in online calculators and then, you calculate the p value it is going to be around 0.33. So, based on our observations in step 5 and step 6, we can say that actually we do not have enough evidence so, we do not reject null and there is no evidence that boys are better than the girls. So, in the next lecture, we are going to continue this discussion on hypothesis testing. Thank you.