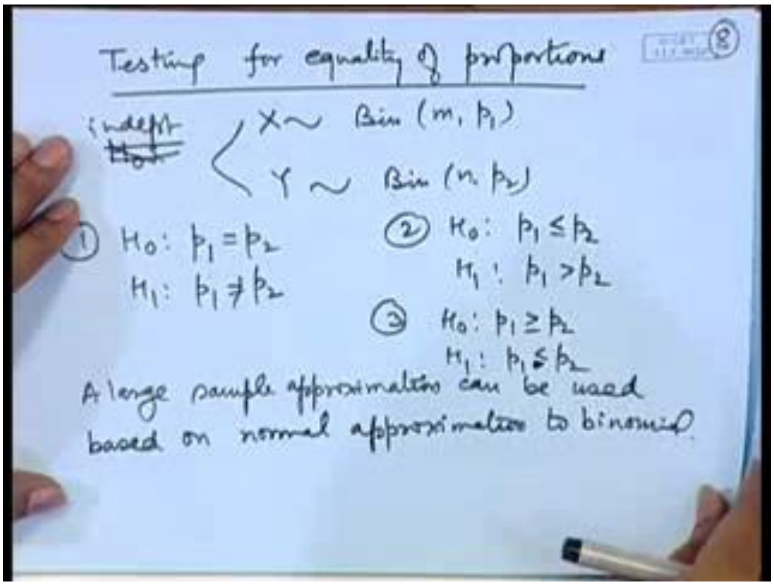
Probability and Statistics
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology, Kharagpur

Lecture – 75
Testing Equality of Proportions

We also have the normal test when we are comparing the proportions of the 2 binomial populations. So, for example, the data is recorded not in the numerical measure, but in the characteristic form. For example, we want to see certain opinions; we want to see the effect of certain say learning procedure. So, that may be the result may be in the form of that a test is conducted, so for example, a set of a student is start a certain instructional material, another set of a student start another instructional material, a common test is conducted we want to see how many passed in the first set and how many passed in the second. Is there a significant difference in the proportions? That means, we want to see whether the instructional material one is better or the instructional material two is better.

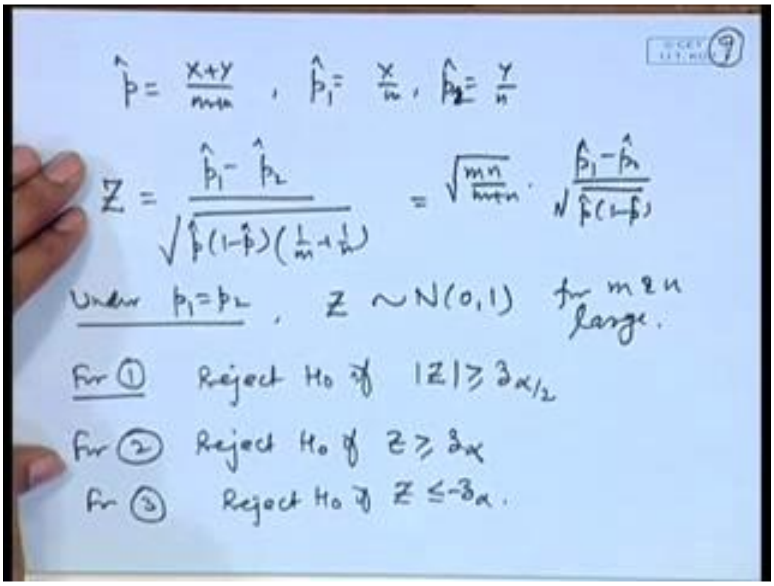
(Refer Slide Time: 01:23)



So, basically the situation is of the following form testing for equality of proportions. So, the general model is of the say X follows say binomial m, p 1 and Y follows binomial n, p 2.

We assume these 2 samples to be independently taken here. So, we are interesting in testing hypothesis of the form p 1 is equal to p 2, against say p 1 is not equal to p 2 say this is 1. We may test p 1 is less than or equal to p 2, against p 1 is greater than p 2 or say H naught; p 1 is greater than or equal to p 2, against H 1; p 1 is less than p 2 etcetera. A large sample approximation can be used based on normal approximation to binomial.

(Refer Slide Time: 03:03)

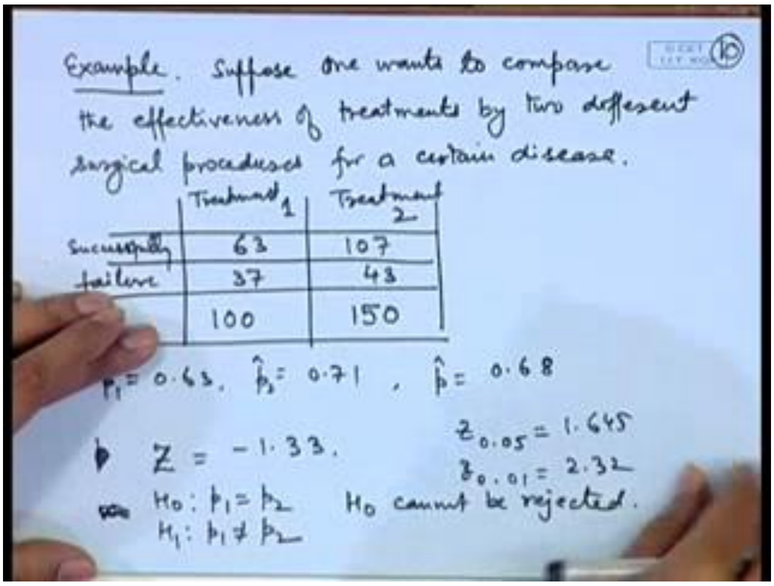


So, we may calculate here p as hat X plus Y by m plus n, p 1 hat is equal to X by m, p 2 hat is equal to Y by n. So, this is the first proportion, this is the second proportion sample proportion, this is the pooled proportion. So, we construct a statistic a Z as p 1 hat minus p 2 hat divided by square root p hat into 1 minus p hat, 1 by m plus 1 by n, which is actually equal to root of m n by m plus n, p 1 hat minus p 2 hat divided by root p hat into 1 minus p.

Under the assumption that p 1 is equal to p 2 Z is approximately normal 0, 1 for m and n large. So, we may make a test based on this for 1 reject H naught if modulus Z is greater than or equal to z alpha by 2. For 2 it will be reject H naught, if Z is greater than or equal to z alpha. For 3 reject H naught, if Z is less than or equal to minus z alpha. So, this is an approximate test, if the assumption that m and n are large is not true in that case we may have to go for a exact procedure, but that procedure will make use of the distribution which is calculated from the binomial.

So, under p 1 equal to p 2, the distribution of X plus Y is again binomial and one can make use of the distribution of X given X plus Y, which is hyper geometric and there is a test procedure for that, but we are not going to discuss that in this particular course here. Let me give an application of this here.

(Refer Slide Time: 05:27)



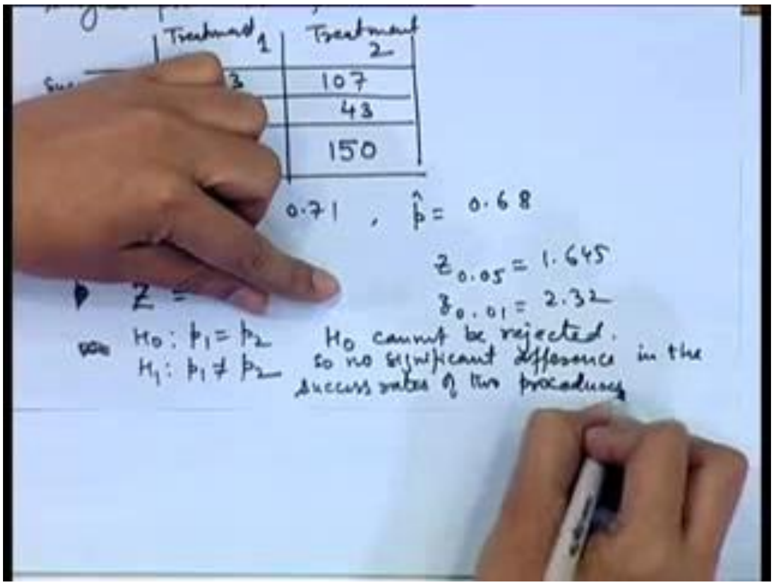
Suppose one wants to compare the effectiveness of treatments by two different surgical procedures for a certain disease.

So, for this set of patients is taken on one set of patients, one surgical procedure is adopted and we observe the proportion of success. So, let us make the data in this particular fashion, suppose 100 patients are there on which treatment procedure 1 is adopted, we see how many are successfully treated and how many of them are failures. Suppose in out of 100 year, it turns out that 63 are successfully treated 37 are unsuccessful whereas, using the treatment procedure 2; suppose 100 patients 150 patients were given this procedure out of that 107 where successfully treated and 43 are not successfully treated; that means, on them the surgical procedure did not yield any positive result.

Let us look at the proportions here p 1 hat is 0.63, p 2 hat is equal to 0.71 and p hat is equal to 0.68. So, if we calculate the Z statistic here, that is p 1 hat minus p 2 hat divided by root of p hat into 1 minus p hat into m n by m plus n, this value turns out to be minus 1.33. If we are considering any reasonable level of significance say Z 0.05 that is 1.645, suppose I take z 0.01, that is 2.32 etcetera; then you can see that we if we consider say hypothesis H naught: p 1 is equal to p 2 against say H 1, p 1 is not equal to p 2.

Then at level of significance say 10 percent, 2 percent etcetera. H naught cannot be rejected, because the absolute value of z that is 1.33 is a smaller than this values.

(Refer Slide Time: 09:01)

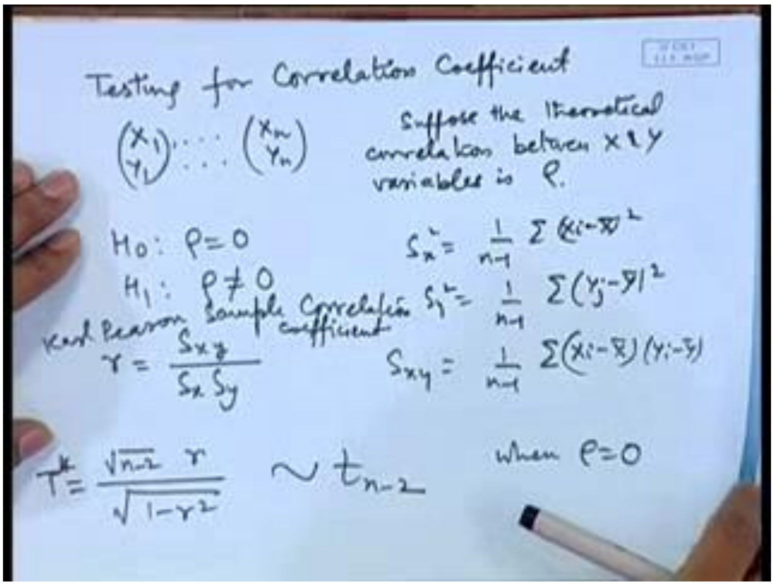


So, that means, there is no significant difference in the 2; in the success rate of the 2 surgical procedures. So, no significant difference in the success rates of 2 procedures. Although from here it looks that the second procedure is more effective, but statistically speaking there is no significance difference here.

We have seen here that sometimes the assumption of the correlated observations is required as in the case of exercise program etcetera, but we have seen the use of t whether t test should be done under certain checks; that means, we should use it with k f for example, we are adopting a paired t test procedure, but the value of the correlation is say 0, in that case sigma 1 square plus sigma 2 square will be the variance of X i minus Y i; that means, we have unnecessary used reduced over degree of freedom here, we are using only n minus 1 degrees of freedom, consequently our power of the test will reduce.

So, it is not advisable to go for a paired t test here; that means, in such a case it will be a reasonable option. Firstly, to check whether the correlation is 0 or not in the given data set. So, fortunately we can find actually a test for the correlation coefficient being significantly different from 0 or not.

(Refer Slide Time: 10:53)

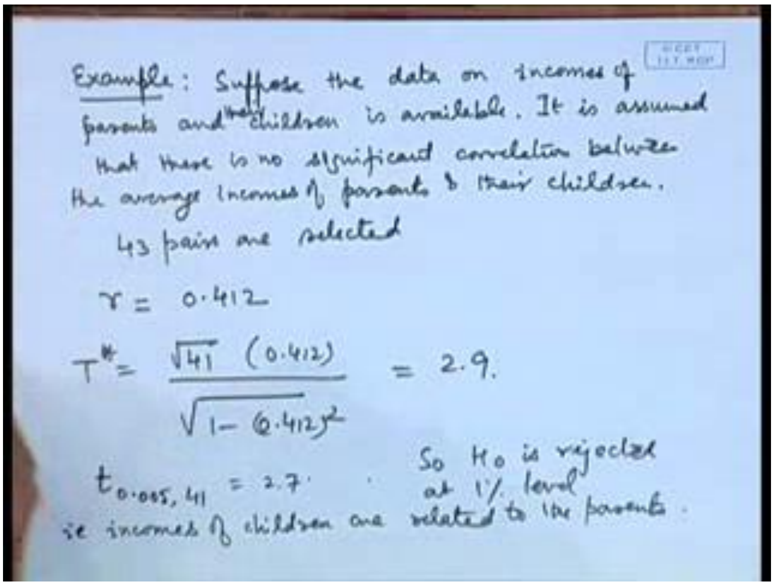


So, testing for correlation coefficient; so we have the data X 1, X 2, X n; Y 1, Y 2, Y n. So, in the 2 data sets we consider the correlate sample correlation coefficient and the population correlation coefficient suppose the theoretical correlation between X and Y variables is rho.

So, we want to test whether rho is equal to 0, against say rho is not equal to 0. For this we calculate the sample correlation coefficient that is r, that is equal to S x y divided by S x S y; where this S x square is 1 by n minus 1, sigma X i minus X bar whole square, S Y square is 1 by n minus 1, sigma Y j minus Y bar whole square and S x y is equal to 1 by n minus 1 sigma X i minus X bar into Y i minus Y bar, that is the sample standard sample variances for the 2 samples and this is the sample covariance.

So, based on this we define the Karl Pearson sample correlation coefficient, then it has been observed that square root n minus 2 into r divided by the root 1 minus r square this is having a t distribution on n minus 2 degrees of freedom, when rho is equal to 0, let me call it t star. So, one can make use of this for testing the significance of correlation; that means, whether there is a significant correlation between the 2 variables or not.

(Refer Slide Time: 13:26)

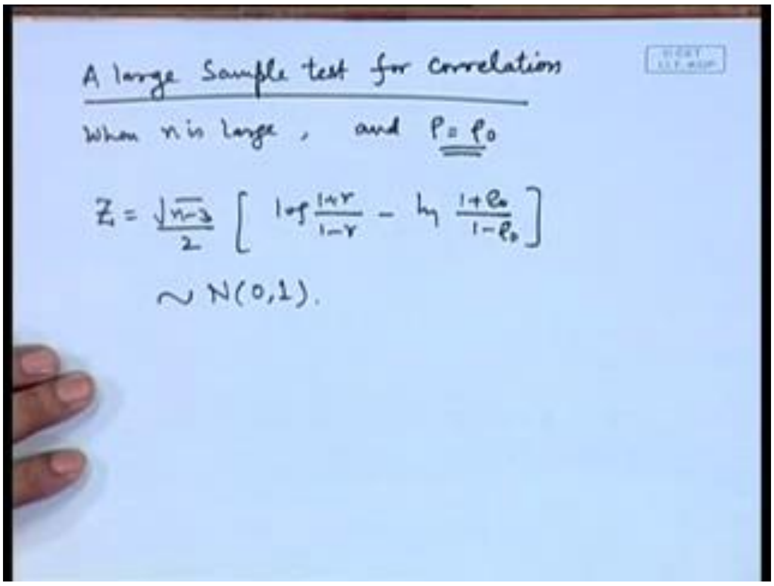


Let us consider one example here; suppose the data on incomes of parents and children their children is available, it is assumed that there is no significant correlation between the average incomes of parents and their children. So, 43 pairs are selected and their sample correlation was calculated which turned out to be 0.412. So, from here this T star value that is square root 41 into 0.412 divided by square root 1 minus 0.412 square is calculated, this value turns out to be 2.9.

Suppose we are considering say t on say 0.005 at 41, this value is 2.7. Now you can see here that this is extremely small level of significance we are taking. So, H naught is rejected at any reasonable level of significance, this is actually 1 percent level. So, at 1 percent level itself this is rejected; that means, the incomes of children are related to the parents; that means, higher income parents their children will also tend to earn higher incomes, and lower income parents their children will have lower incomes.

Now, this test is again based on the normality assumption; sometimes the normality assumption may not be valid, in that case there is a large sample test for correlation coefficient.

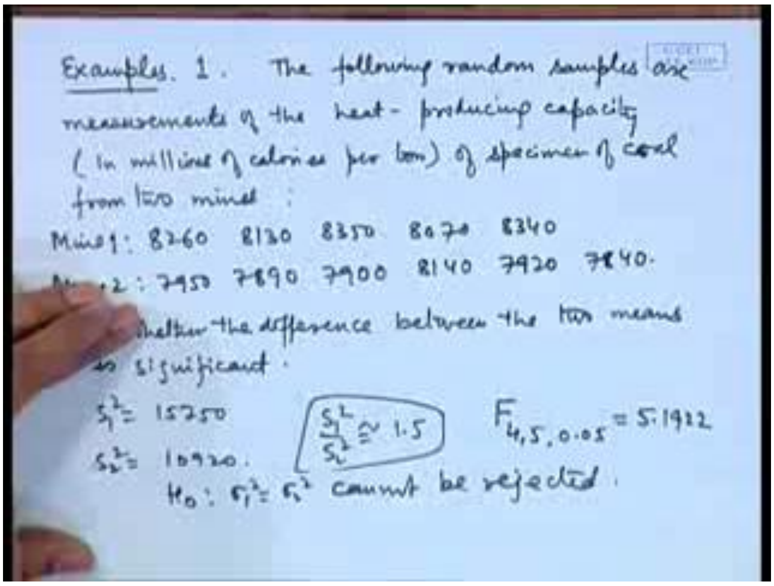
(Refer Slide Time: 16:17)



When n is large and say rho is equal to rho naught. So, we may not test that rho is equal to 0 but some arbitrary value rho naught, then we construct Z that is equal to root n minus 3 by 2, log of 1 plus r divided by 1 minus r minus log, 1 plus rho naught divided by 1 minus rho naught then this has approximately normal 0, 1 (Refer Time: 17:17).

So, consequently if n is large, then we may test about the correlation being equal to any arbitrary value. Of course, if rho is 0 here then this term will vanish and we will have only this particular term. So, this is an approximate normal test here and this is not based on the assumption of normality for the initial samples that is X i's and Y i's here. Let us take a few more examples here of the test that we have discuss today. So, let us take one example based on say normal distributions here.

(Refer Slide Time: 18:23)

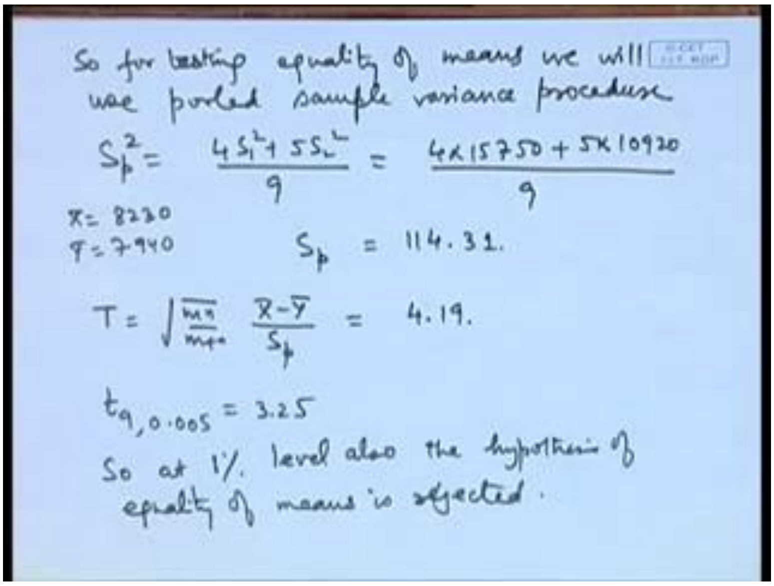


The following random samples are measurements of the heat producing capacity in millions of calories per ton of a specimen of coal from two mines. So, in mine 1: there will be 5 measurements are taken, the values are 8260, 8130, 8350, 8070, 8040. In the mind 2: 6 observations were taken, 7950, 7890, 7900 8140, 7920 and 7840. Test whether the difference between the 2 means is significant.

So, here 2 samples are available to us, and we are having 5 and 6 observations respectively from the 2 samples. So, what we do here, we have to check whether mu 1 is equal to mu 2 or mu 1 is not equal to mu 2, but once again here to test this hypothesis we will have to test about the variances also. So, whether the variances are same? So, here we see that if we consider S 1 square here, S 1 square is equal to 15750 etcetera S 2 square is equal to 10920.

So, if we consider the ratio here, S 1 square by S 2 square that is approximately 1.5. So, if I am looking at the F value on say 1, 2, 3, 4, 5. So, 4 and 5 degrees of freedom, let us see one example here from the tables of the F distribution and say at 0.05 level, if we are seeing F 4 5. So, at 0.05 the value is 5.1922. So, H naught: sigma 1 square is equal to sigma 2 square cannot be rejected. Now if this cannot be rejected then for the equality of means, we will go for the pooled sample variance procedure.

(Refer Slide Time: 22:14)



So, for testing equality of means we will use pooled sample variance procedure. So, if we see S p square that is equal to 4 S 1 square plus 5 S 2 square by 9 that turns out to be. So, that is equal to well that is a huge value here, 4 into 1 15750 plus 5 into 10920 by 9. So, that is equal to some value. So, S p is taken to be square root of that that is 114.31.

So, if we calculate the T variable here that is root m n by m plus n, X bar minus Y bar by S p. Then first thing is we observe here that X bar is equal to 8230 and Y bar is equal to 7940. So, this value turns out to be after calculation 4.19. So, if we see the t values at say 9 degrees of freedom then even at 0.005 this value is 3.25. So, at 1 percent level also, the hypothesis of equality of means is rejected.

So, we conclude that in the 2 mines, the measurements of the heat producing capacity are significantly different. Because it may be due to difference in the type of the coal that is available, it may be due to the type of the mine that you are having may be in, one of the mines you have a very low level roots and various kind of parameters which may be operating in those mines there. So, it may be due to that. So, to sum up if we are comparing the means of 2 normal populations, the first thing is we have to look at is that what type of variances are there. If the variances are known, then we have one type of procedure. If the variances are unknown, then we firstly test whether the variances are same or not; if they are same then we go for a pooled sample variance procedure, if they are not same then we go for a different procedure which is an approximate test.

We also see the correlation if the correlation is present then we may go for a pairing, if we do not have the correlation then we may go for independent samples. So before adopting any procedure, one as to carefully examine the problem and then choose the appropriate test; we have also seen the effect of choosing the null and alternative hypothesis. As I have already mentioned since we are controlling the probability of type one error therefore, it is always reasonable to put a stronger hypothesis or you can say the convection in which we have more that hypothesis as an alternative hypothesis, because rejection of the hypothesis strong conclusion whereas, acceptance of the hypothesis becomes a weaker conclusion. Simply because of the reason that we are actually concluding the probability of type one error.

In the forth coming lecture I will be discussing the chi square test for goodness of fit or testing for the independence.