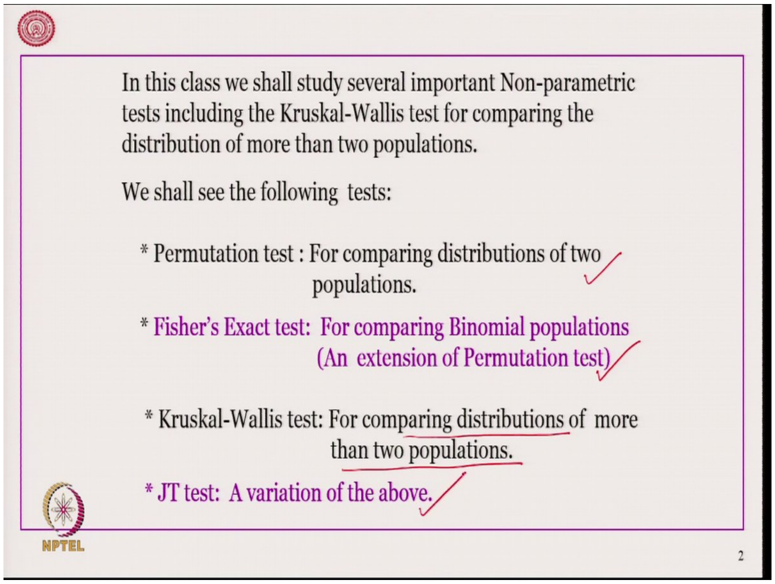
Nonparametric Statistical Inference
Professor Niladri Chatterjee
Department of Mathematics
Indian Institute of Technology, Delhi
Lecture 10
Nonparametric-Statistical Inference

Welcome students to the MOOCs series of lectures on nonparametric statistical inference. This is lecture number 10. And also, this is going to be the last lecture of this series.

(Refer Slide Time: 00:33)



In this class we shall study several important Non-parametric tests including the Kruskal-Wallis test for comparing the distribution of more than two populations.

We shall see the following tests:

* Permutation test : For comparing distributions of two populations.

* Fisher's Exact test: For comparing Binomial populations (An extension of Permutation test)

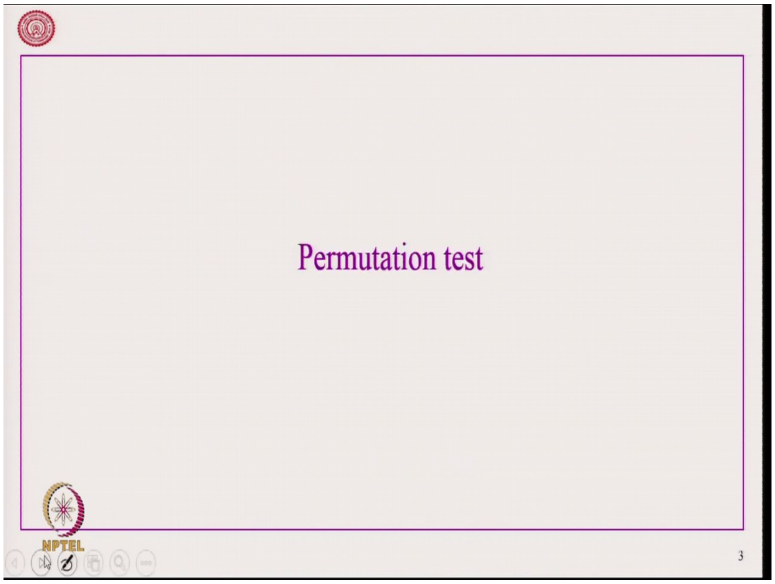
* Kruskal-Wallis test: For comparing distributions of more than two populations.

* JT test: A variation of the above.

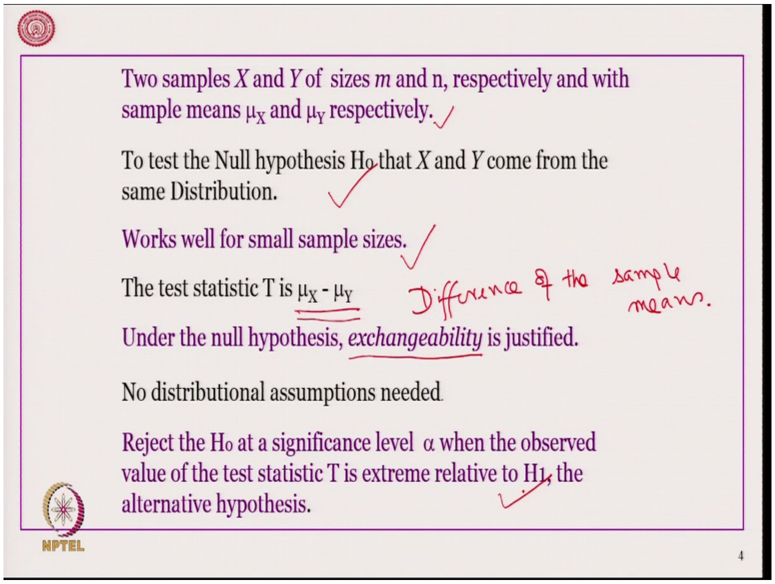
At the end of the last class I said that, I will continue with different algorithms in this class. In particular, we shall study four different tests namely, permutation test which is used for comparing the distributions of two populations, Fisher's Exact test which is used for comparing binomial populations or two binomial populations, and we shall see that it is an extension of the permutation test.

Then, I shall talk about Kruskal-Wallis test, it is different from the earlier ones in the sense that here we will be comparing more than two populations. That is, we shall focus on the distributions of more than two populations. And then we will go to JT test which is a variation of the above. That means here also we shall compare more than two population distributions.

(Refer Slide Time: 01:47)



Permutation test



Two samples *X* and *Y* of sizes *m* and n, respectively and with sample means μ_X and μ_Y respectively.

To test the Null hypothesis Ho that *X* and *Y* come from the same Distribution.

Works well for small sample sizes.

The test statistic T is $\mu_X - \mu_Y$ Difference of the sample means.

Under the null hypothesis, *exchangeability* is justified.

No distributional assumptions needed.

Reject the H_0 at a significance level α when the observed value of the test statistic T is extreme relative to H_1, the alternative hypothesis.

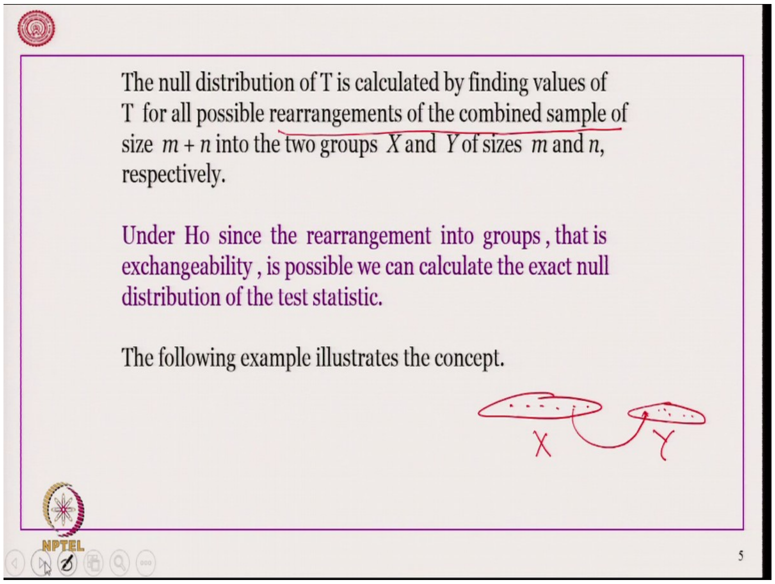
So, let us begin with permutation test. So, the background is very familiar we have two samples X and Y of sizes m and n and with sample means μ_X and μ_Y respectively. To test the null hypothesis Ho, that X and Y have come from the same distribution. So, it is the same problem, we are comparing the distribution of two different populations and we are checking whether they can be considered to be coming from the same distribution or not.

This works well for small sample sizes. The test statistic T is $\mu_X - \mu_Y$ that is the difference of the sample means. Under the null hypothesis exchangeability is justified. So, it is a new term that we are coming across I am explaining what it is, no distributional assumptions are needed.

Therefore, we are not assuming anything about the distributions, but reject Ho at a significance level α when the observed value of the test statistic T is extreme relative to H1 that is the alternative hypothesis.

Extreme means, since you are considering $\mu_X - \mu_Y$, if $\mu_X - \mu_Y$ is significantly greater than 0 that means in the direction of positive and if it is significantly less than 0 that means, in the direction of the negative, then depending upon the alternative hypothesis, we will have to reject the null Ho.

(Refer Slide Time: 03:51)



The null distribution of T is calculated by finding values of T for all possible rearrangements of the combined sample of size $m + n$ into the two groups X and Y of sizes m and n, respectively.

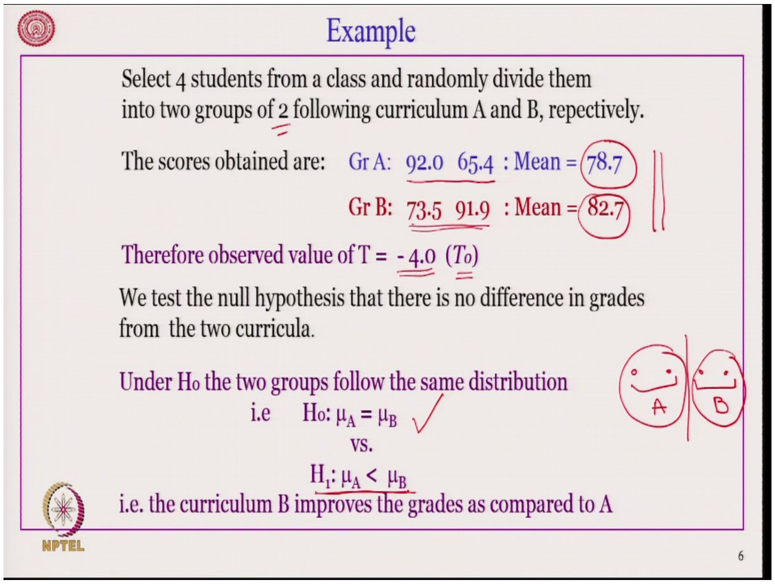
Under Ho since the rearrangement into groups, that is exchangeability, is possible we can calculate the exact null distribution of the test statistic.

The following example illustrates the concept.

Now, the question is how do we calculate the null distribution of T? It is calculated by finding the values of T for all possible rearrangements of the combined sample of size m + n into the two groups X and Y of sizes m and n. What does it mean? Suppose, this is my X and there are certain observations and this is my Y and there are certain observations. Now, if X and Y are coming from the same distribution, then each possible X could have been a member of Y and each possible Y could have been a member of X.

That is, this total m + n many observations, we could find a μ among themselves such that here comes m and here comes n of them and under the null hypothesis they should all be equally likely. This is what exchangeability that we say X observation can we exchange with and Y observation and vice versa.

(Refer Slide Time: 05:13)



Example

Select 4 students from a class and randomly divide them into two groups of 2 following curriculum A and B, repectively.

The scores obtained are: Gr A: 92.0 65.4 : Mean = 78.7

Gr B: 73.5 91.9 : Mean = 82.7

Therefore observed value of T = -4.0 (T_o)

We test the null hypothesis that there is no difference in grades from the two curricula.

Under Ho the two groups follow the same distribution

i.e Ho: $\mu_A = \mu_B$

vs.

H_1: $\mu_A < \mu_B$

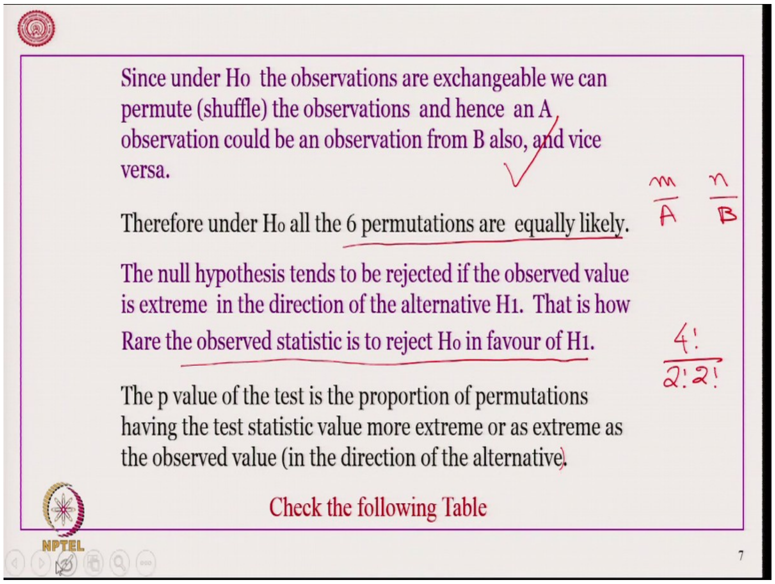
i.e. the curriculum B improves the grades as compared to A

So, the following example illustrate the concept in a better way, select 4 students from a class and randomly divide them into two groups of 2 following the curriculum A and B. That means, there are 4 persons then we are dividing it into two groups of 2 and 2, suppose we could divide it like that, and suppose these 2 are following curriculum A and these 2 are following curriculum B. Now, the average that we got is for group A, since individual scores are 92 and 65.4, the average coming out to be 78.7. On the other ends, the group B students, they have scored 73.5 and 91.9 and their mean has come out to be 82.7.

Therefore, the observed value of T which we are calling say T_o is coming out to be the difference μ_A - μ_B, which is coming out to be minus 4.0. Now, we want to test under the null hypothesis, that there is no difference in grades from the two curricula, that means, which are coming from this population as students who are coming from this population, there is no variation in their grades.

That is under the null hypothesis, since you are checking the equality of central location, you are looking at Ho is $\mu_A = \mu_B$ and what is going to be the alternative? Looking at these we will feel that the average of A is less than the average of B. Therefore, for this particular problem, we can choose that H1 is equal to $\mu_A < \mu_B$. That is, the curriculum B improves the grades as compared to the curriculum A.

(Refer Slide Time: 07:24)



Since under Ho the observations are exchangeable we can permute (shuffle) the observations and hence an A observation could be an observation from B also, and vice versa.

Therefore under H_0 all the 6 permutations are equally likely.

$\frac{m}{A}$ $\frac{n}{B}$

The null hypothesis tends to be rejected if the observed value is extreme in the direction of the alternative H_1. That is how Rare the observed statistic is to reject H_0 in favour of H_1.

$\frac{4!}{2!2!}$

The p value of the test is the proportion of permutations having the test statistic value more extreme or as extreme as the observed value (in the direction of the alternative).

Check the following Table

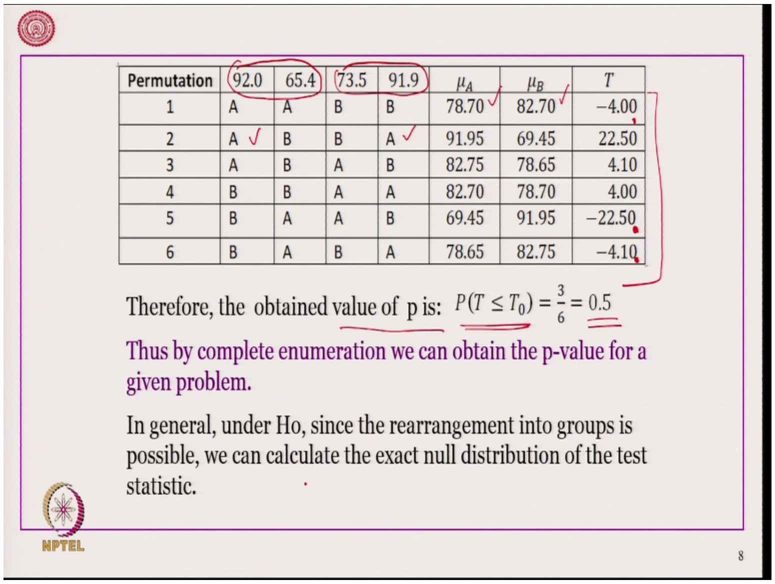
Since under Ho, the observations are exchangeable, we can permute a shuffled the observation and hence, and A observation could be an observation from B also and vice versa. So, this is what I was talking about that this is what is called the exchangeability and element of A could have come from B as well as an element of B could have come from A. However, we have to remember that the m and n these are the numbers corresponding to the population A and corresponding to population B respectively.

So, only thing is that that has to be maintained. Therefore, under Ho all the 6 permutations are equally likely. Why 6 permutations? Because, 4! upon 2! x 2! that is going to be the total number of arrangements and that gives us 6 many different permutations. The null hypothesis tends to be rejected, if the observed value is extreme in the direction of the alternative H1 that means, we shall check if the difference μ_A - μ_B is too small or that is absolute value is very large in the negative direction.

And if that happens, then we will check how rare the observed statistic is to reject Ho in favor of H1 that is the normal thing we do; we look at this, how rare it is what is the probability of getting this value if that is too small, then we are going to reject the null hypothesis, we will assume that actually A is giving lower performance in terms of grade improvement in comparison with the curriculum B.

The p value of the test is the proportion of permutations having the test statistic value more extreme or as extreme as the observed value in the direction of the alternative. So, with that, let us check the following table.

(Refer Slide Time: 09:39)



Permutation	92.0	65.4	73.5	91.9	μ_A	μ_B	T
1	A	A	B	B	78.70 ✓	82.70 ✓	−4.00
2	A ✓	B	B	A ✓	91.95	69.45	22.50
3	A	B	A	B	82.75	78.65	4.10
4	B	B	A	A	82.70	78.70	4.00
5	B	A	A	B	69.45	91.95	−22.50
6	B	A	B	A	78.65	82.75	−4.10

Therefore, the obtained value of p is: $\underline{P(T \leq T_0)} = \frac{3}{6} = \underline{\underline{0.5}}$

Thus by complete enumeration we can obtain the p-value for a given problem.

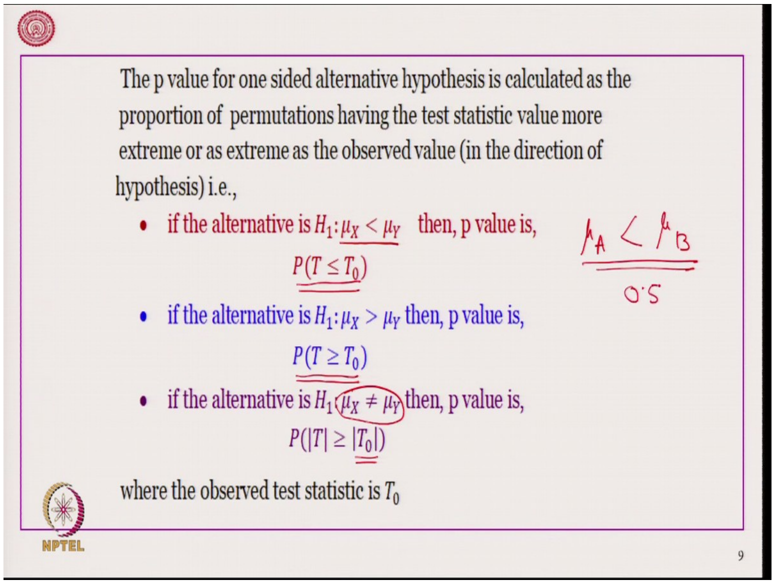
In general, under Ho, since the rearrangement into groups is possible, we can calculate the exact null distribution of the test statistic.

We have computed for these observations, all the 6 permutations. What does it mean? The first row says that these two could have been observations from A and these two could have been observations from B. Whereas, permutation two is saying that this and this are coming from observation A and these two are coming from observation B, the middle two, 65.4 and 73.5. So, like that for all the 6 permutations, we have computed, what is going to be the μ_A and what is going to be the μ_B and we have computed the value of the statistic T for all the 6 permutations. What is the advantage?

This gives us the probability of getting 1 particular value of T. Therefore, the obtained value of p is probability $T \leq T_o$. Our T_o has come out to be - 4. Therefore, how many of them are less than or equal to - 4? 1, 2 and 3, therefore, that probability is coming out to be 3 by 6 is equal to 0.5. That is my complete enumeration, we can obtain the p value for a given problem.

Quite naturally, we can do it only when the number m and n are quite small. Of course, if they are very large, then we have to go to computer programming. In general under Ho since the rearrangement in two groups is possible, we can calculate the exact null distribution of the test statistic as we have computed here.

(Refer Slide Time: 11:42)



The p value for one sided alternative hypothesis is calculated as the proportion of permutations having the test statistic value more extreme or as extreme as the observed value (in the direction of hypothesis) i.e.,

- if the alternative is $H_1: \mu_X < \mu_Y$ then, p value is, $P(T \leq T_0)$
- if the alternative is $H_1: \mu_X > \mu_Y$ then, p value is, $P(T \geq T_0)$
- if the alternative is $H_1: \mu_X \neq \mu_Y$ then, p value is, $P(|T| \geq |T_0|)$

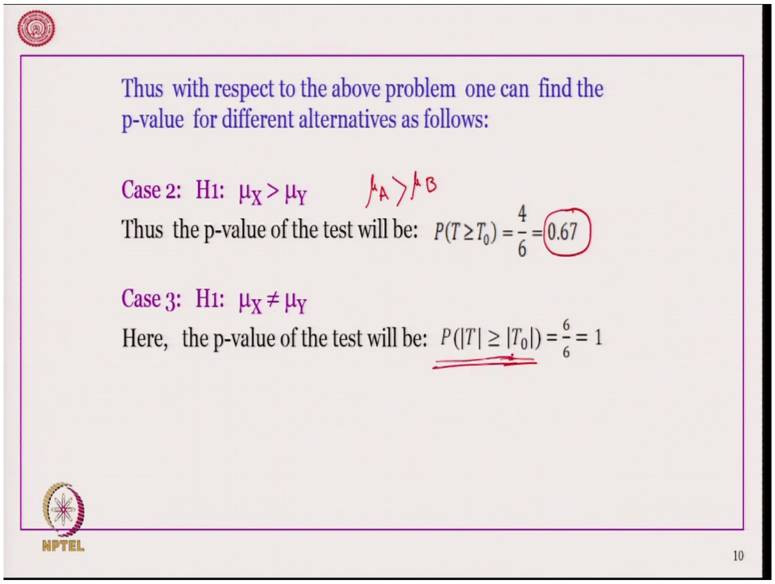
where the observed test statistic is T_0

$\mu_A < \mu_B$

0·5

The p value for one sided alternative hypothesis is calculated as the proportion of permutations having the test statistic value more extreme or as extreme as the observed value in the direction of the hypothesis, which we have said before. Therefore, if the alternative is $\mu_X < \mu_Y$, then p value is probability $T \leq T_0$. If the alternative is $\mu_X > \mu_Y$, then p value is probability $T \quad T_0$. However, if the alternative is two sided, that is $\mu_X \neq \mu_Y$, then we shall look at the absolute value of T the absolute value of T_0 that we have got and what is the probability for that one. For this case, our alternative was that mu A is less than mu B, therefore, we will stick to this one. And we have calculated that this probability is coming out to be 3 by 6 is equal to 0.5.

(Refer Slide Time: 12:55)



Thus with respect to the above problem one can find the p-value for different alternatives as follows:

Case 2: H1: $\mu_X > \mu_Y$ $\mu_A > \mu_B$

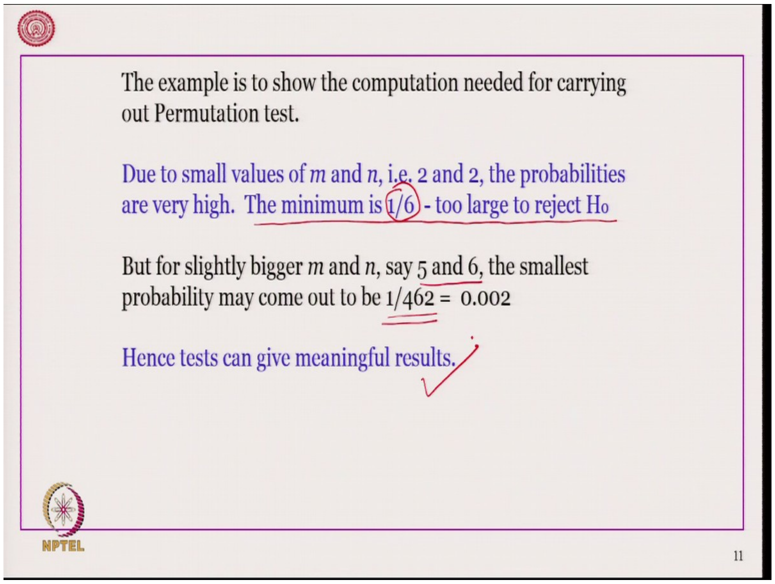
Thus the p-value of the test will be: $P(T \geq T_0) = \frac{4}{6} = \boxed{0.67}$

Case 3: H1: $\mu_X \neq \mu_Y$

Here, the p-value of the test will be: $\underline{P(|T| \geq |T_0|)} = \frac{6}{6} = 1$

Similarly, if the alternative would have been $\mu_X > \mu_Y$, that is $\mu_A > \mu_B$, then that would have given us the value 0.67 for p. And if would have looked at probability of $|T|$ $|T_0|$ that would have given us the value 1.

(Refer Slide Time: 13:17)



The example is to show the computation needed for carrying out Permutation test.

Due to small values of *m* and *n*, i.e. 2 and 2, the probabilities are very high. The minimum is $\boxed{1/6}$ - too large to reject H₀

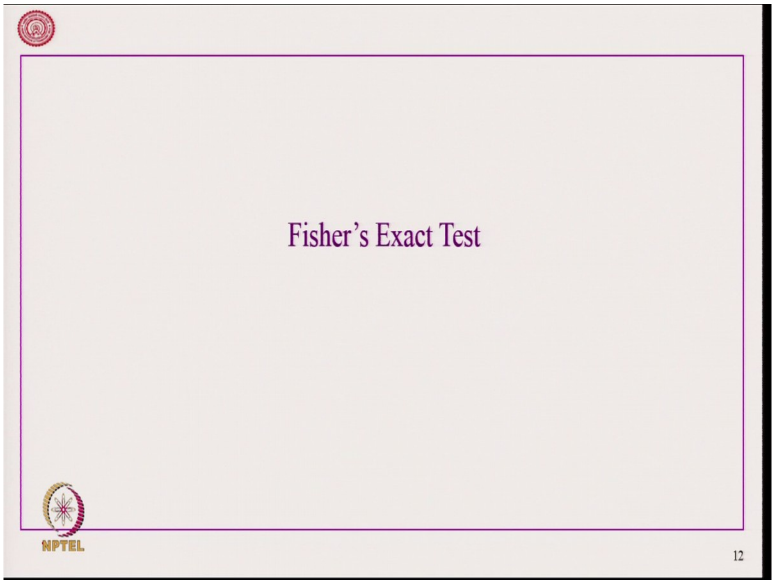
But for slightly bigger *m* and *n*, say 5 and 6, the smallest probability may come out to be 1/462 = 0.002

Hence tests can give meaningful results. ✓

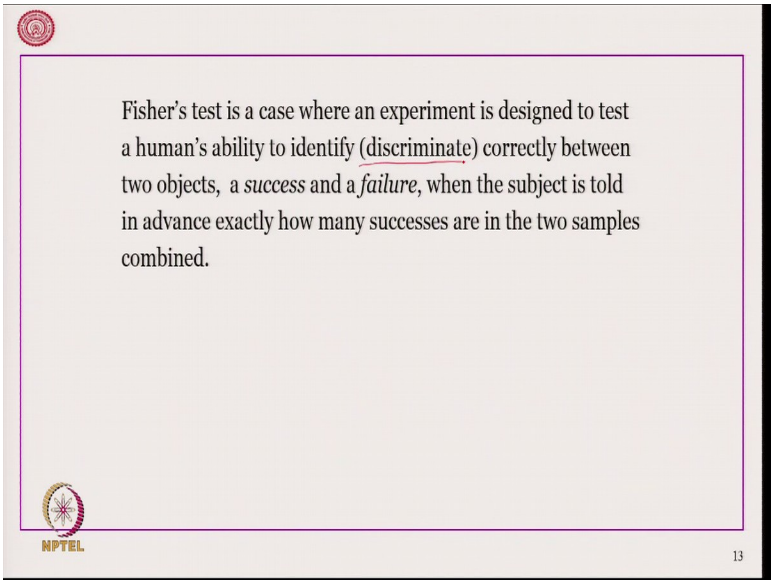
The example is there to show the competition needed for carrying out a permutation test. Due to the small values of m and n, that is, we have taken only 2 and 2, the probabilities coming out to be very high. This is because the total number of permutation is 6. Therefore, the minimum probability that we will get for any event is 1 by 6, it cannot be less than that, that itself is pretty high to reject the null hypothesis at say 5 % level of significance.

But for slightly higher value of m and n, say suppose, we have taken 5 and 6 to be the sample sizes from the 2 populations, then the smallest probability may come out to be 1 upon 462 which is 0.002. And therefore, when we apply permutation test here, we could get some meaningful result even at 5 % level or 1 % level of significance.

(Refer Slide Time: 14:24)



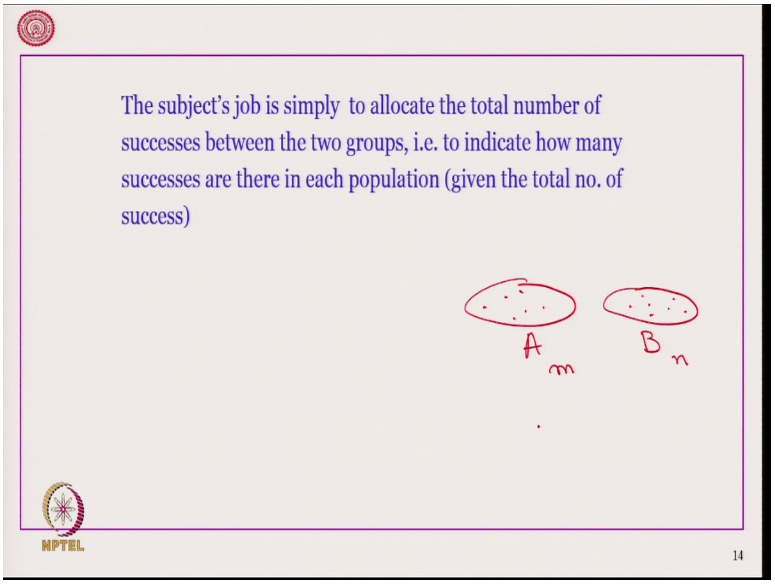
Fisher's Exact Test



Fisher's test is a case where an experiment is designed to test a human's ability to identify (discriminate) correctly between two objects, a *success* and a *failure*, when the subject is told in advance exactly how many successes are in the two samples combined.

Let us now go to the next test, which is called Fisher's exact test. As I said earlier, here also we shall use the permutation test, but in a slightly more tricky way. As we see, as Fisher's exact test is for comparing two binomial populations. So, what is Fisher's test? It is a case where an experiment is designed to test a human's ability to identify or to discriminate correctly between 2 objects, a success and failure when the subject is told advanced, exactly how many successes are there in that two samples combined? Complicated, but we will understand what I mean.

(Refer Slide Time: 15:15)

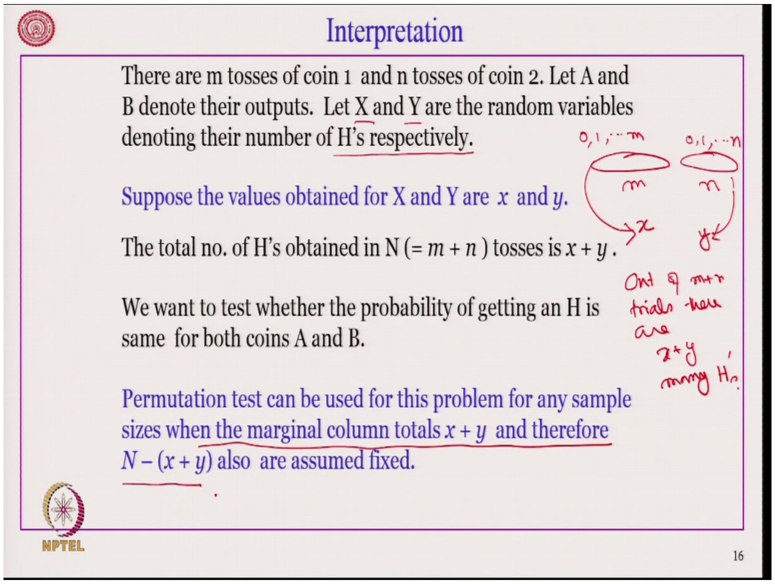


The subject's job is simply to allocate the total number of successes between the two groups, i.e. to indicate how many successes are there in each population (given the total no. of success)

The subject's job is to allocate the total number of successes between the two groups, that is to indicate how many successes are there in each population given the total number of success. So, suppose as before there is A population and there is B population of size say m and n. And there are some success and failure here and some success and failure here.

And we want to check if there is any difference between the success rate of A and B. What is given to us that total out of m + n many elements, how many successes are there? Based on that, if we have the ability to actually discriminate, how many of them are real success, and how many of them are failure.

(Refer Slide Time: 16:18)



Interpretation

There are m tosses of coin 1 and n tosses of coin 2. Let A and B denote their outputs. Let X and Y are the random variables denoting their number of H's respectively.

Suppose the values obtained for X and Y are x and y.

The total no. of H's obtained in N (= $m + n$) tosses is $x + y$.

We want to test whether the probability of getting an H is same for both coins A and B.

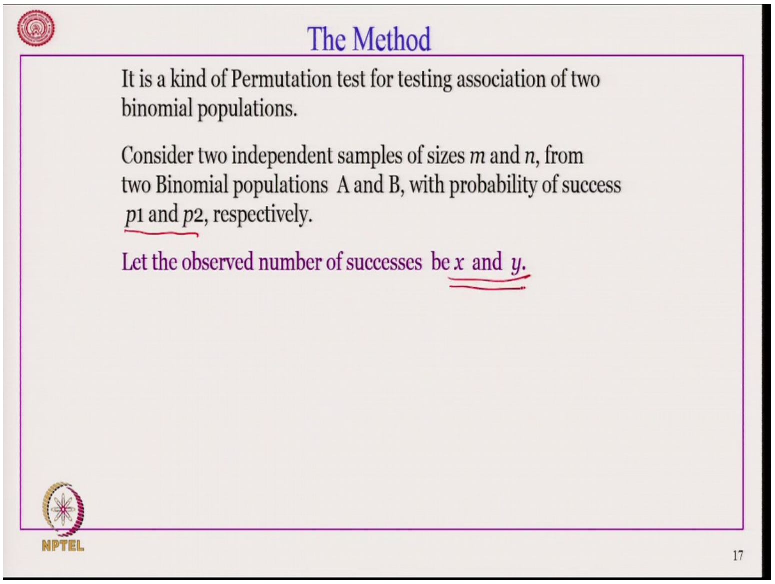
Permutation test can be used for this problem for any sample sizes when the marginal column totals $x + y$ and therefore $N - (x + y)$ also are assumed fixed.

Out of m+n trials there are x+y many H's.

So, what does it interpret to? So, suppose there are m tosses of a coin 1 and n tosses of coin 2. Let A and B denote their outputs, let X and Y are the random variables denoting the number of heads respectively. So, suppose there are m tosses here and n tosses here. Therefore, the number of heads could be 0, 1, 2 up to m and number of heads here could be 0, 1, 2 up to n. It is given that there are x many heads here, and they are y many heads here in second population, what does it mean?

We are giving that out of m + n trials, there are our x plus y many success that is heads. We want to test whether the probability of getting an H is same for both coins A and B. So, permutation test can be used for this problem for any sample sizes, when the marginal column totals x + y for success and N – (x + y) for failure or assumed to be fixed.

(Refer Slide Time: 18:09)



The Method

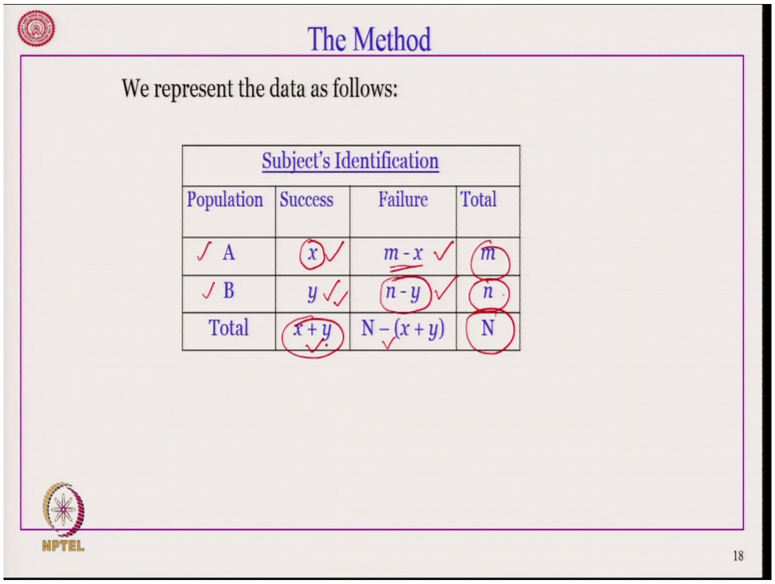
It is a kind of Permutation test for testing association of two binomial populations.

Consider two independent samples of sizes m and n, from two Binomial populations A and B, with probability of success $p1$ and $p2$, respectively.

Let the observed number of successes be x and y.

So, in some sense, we are also using the permutation test for testing the association of the two binomial populations. So, consider two independent samples of sizes m and n from 2 binomial populations A and B with probability of success being p1 and p2 respectively, they need not be equal, whether we can guess it or not, if we guess then we will give equal to both of them, but whether we could guess the actual probabilities or not, that is the question. Let the observed number of successes be x and y.

(Refer Slide Time: 18:48)



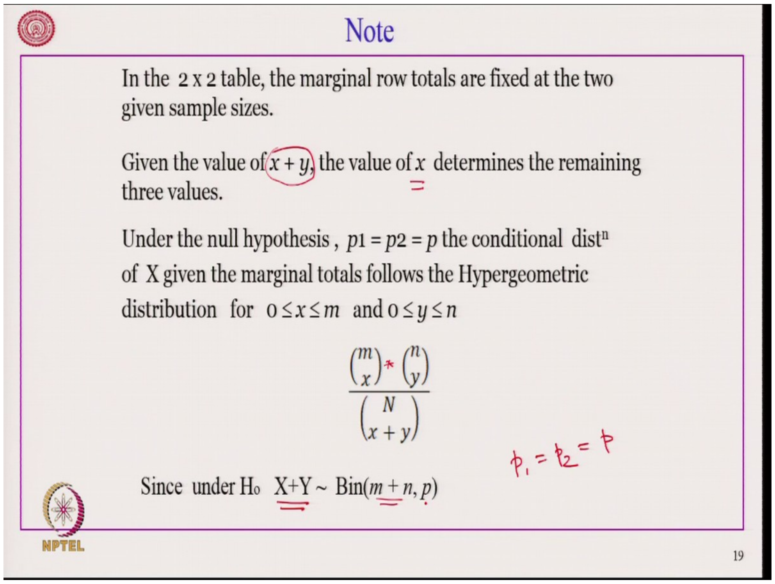
The Method

We represent the data as follows:

Subject's Identification			
Population	Success	Failure	Total
✓ A	x ✓	$m - x$ ✓	m
✓ B	y ✓✓	$n - y$ ✓	n
Total	$x + y$ ✓	$N - (x + y)$ ✓	N

So, what we are saying that this data can be represented in a tabular form there are out of m trials x success and n - x failure for the first population out of n trials, there are y many success and n - y many failure for the second population. Therefore, in total we have x + y many success n - (x + y) many failure where capital N is small m + small n is the total number of samples.

(Refer Slide Time: 19:28)



Note

In the 2 x 2 table, the marginal row totals are fixed at the two given sample sizes.

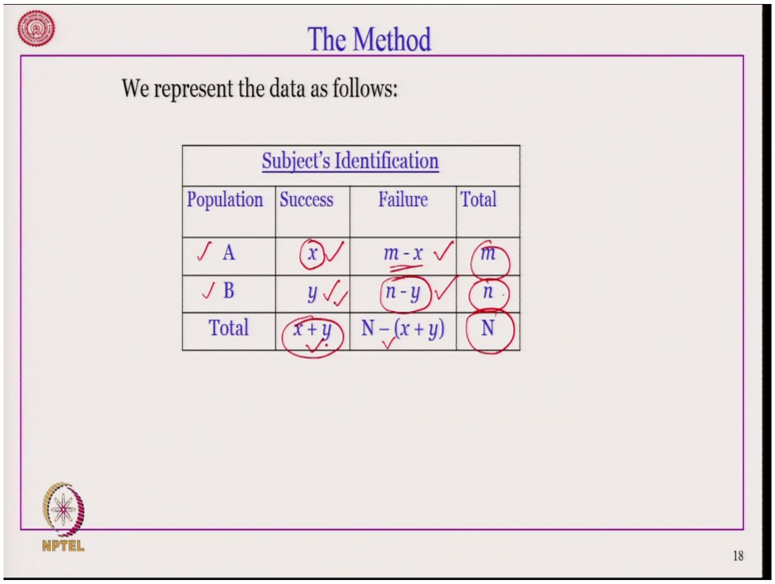
Given the value of $x + y$, the value of $\underline{x}$ determines the remaining three values.

Under the null hypothesis, $p1 = p2 = p$ the conditional dist^n of X given the marginal totals follows the Hypergeometric distribution for $0 \le x \le m$ and $0 \le y \le n$

$$\frac{\binom{m}{x} * \binom{n}{y}}{\binom{N}{x+y}}$$

$p_1 = p_2 = p$

Since under H₀ $\underline{X+Y} \sim \text{Bin}(\underline{m+n}, p)$



The Method

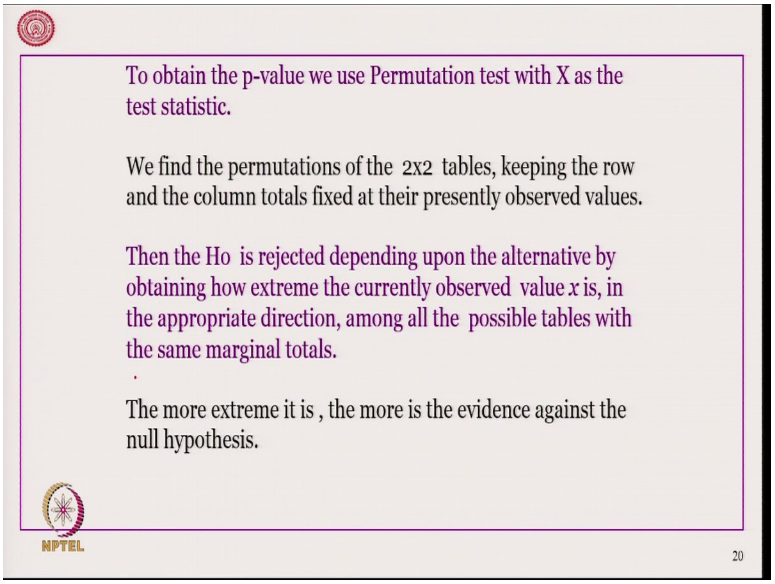
We represent the data as follows:

Subject's Identification			
Population	Success	Failure	Total
✓ A	x ✓	$\underline{m - x}$ ✓	m
✓ B	y ✓✓	$n - y$ ✓	n
Total	$x + y$ ✓	$\text{N} - (x + y)$ ✓	N

In the 2 x 2 table the marginal row totals are fixed at the two given sample sizes and also x + y is given. Therefore, given X we can easily determine the remaining three values, because we are given m and n these are fixed and also you are given x + y. Therefore, there is only one independent value, let us call it x because the moment we know it x, we know what is m - x, we know what is y. And of course, therefore, we know what is n - y.

That is what we are saying that given the value of x, it determines the remaining three values once x + y is fixed. Under the null hypothesis p1 = p2 = p, then the conditional distribution of x given the marginal totals follows a hypergeometric distribution, that means, what is going to be the probability of x out of m many trials that is going to be ^{m}Cx, ^{n}Cy the product divided by capital $^{N}Cx+y$. This happens, because under the null hypothesis that p1 = p2 = p. This sum of x plus y itself is also a binomial random variable with parameters m+n and p.

(Refer Slide Time: 21:12)



To obtain the p-value we use Permutation test with X as the test statistic.

We find the permutations of the 2x2 tables, keeping the row and the column totals fixed at their presently observed values.

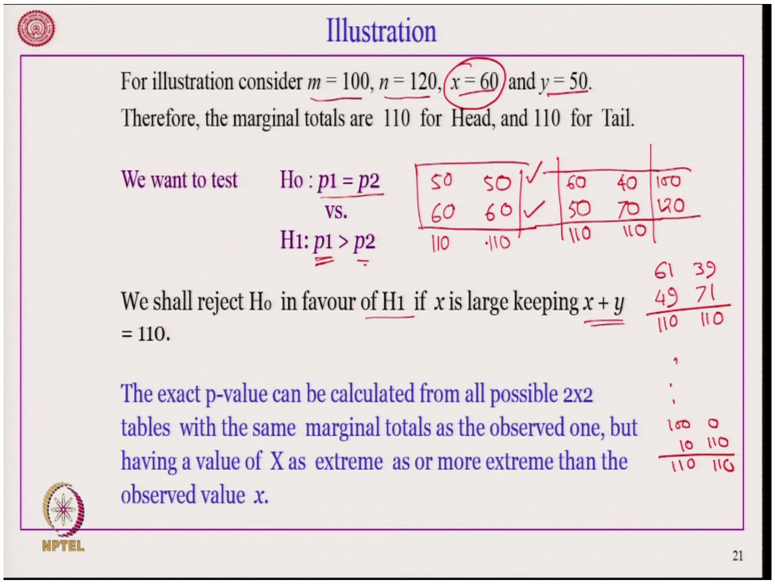
Then the Ho is rejected depending upon the alternative by obtaining how extreme the currently observed value *x* is, in the appropriate direction, among all the possible tables with the same marginal totals.

The more extreme it is , the more is the evidence against the null hypothesis.

To obtain the p value, we use permutation test, with x as the test statistic. To find the permutations of the 2 x 2 tables, keeping the row and column totals fixed at their presently observed values. That means, whatever value is given to us for x + y that will keep fixed and therefore, we should be able to interchange x and y observations among themselves keeping m, n and x+ y fixed.

Then Ho is rejected depending upon the alternative by obtaining how extreme the currently observed value X is in the appropriate direction among all the possible tables with the same marginal totals, the more extreme it is the more is the evidence against the null hypothesis. I will give you an example you will understand it right now.

(Refer Slide Time: 22:15)



Illustration

For illustration consider $m = 100$, $n = 120$, $x = 60$ and $y = 50$.
Therefore, the marginal totals are 110 for Head, and 110 for Tail.

We want to test Ho : $p1 = p2$
vs.
H1: $p1 > p2$

50	50	✓
60	60	✓
110	110	

60	40	100
50	70	120
110	110	

We shall reject Ho in favour of H1 if x is large keeping $x + y$ = 110.

61	39
49	71
110	110

The exact p-value can be calculated from all possible 2x2 tables with the same marginal totals as the observed one, but having a value of X as extreme as or more extreme than the observed value x.

100	0
10	110
110	110

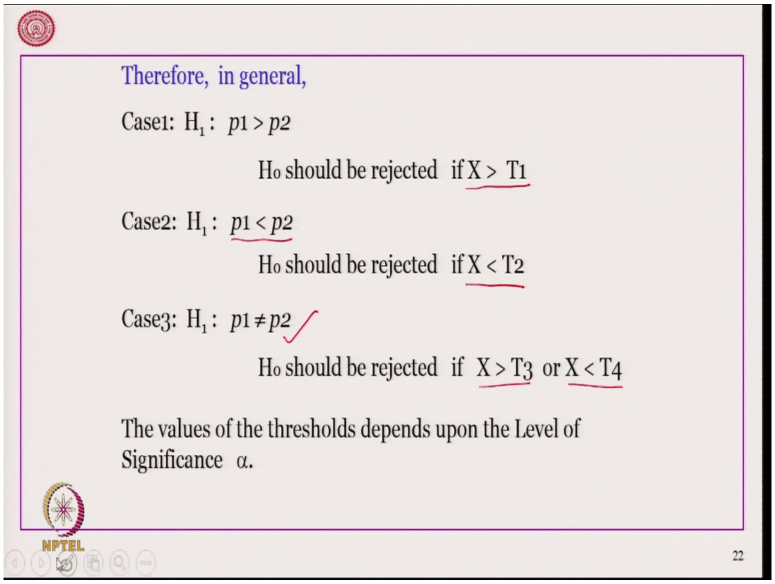
So, let us consider that m is equal to 100 and n is equal to 120, X is equal to 60 and y is equal to 50. That means, if we look at the earlier table, we will get something like this, there are 100 many trials for the first one and 120 many trials for the second one, we have got 60 heads here and therefore, we have got 40 failures here, we have got total 110 many success, therefore, 50 many heads there and therefore 70 many tails here and this is also going to be 110. So, this is the table that we could construct.

We shall reject Ho in favor of H1. Since, we are testing p1 > p2, as we can see that more number or more proportion of trials are ending in head for the first population. Therefore, what we are suspecting that the probability of success may be more for the first population, than the second one. therefore, we are testing the equality of the proportions against the alternative hypothesis that p1 > p2 such that x + y is equal to 110.

Therefore, the exact p value can be calculated from all possible 2 cross 2 tables with the same marginal totals as the observed one, but having a value of x as extreme as or more than that observed value of x. For example, one possible table could have been 61, 39, 49 and 71 then that would have give us again, 110 and 110. Like that, we could go perhaps maximum that we could do is 100, 0, 10 and 110 and then you also get 110, 110.

So, this is extreme. This could have been one possibility. Of course, some other possibility could have been that there are only 50 heads and 50 tails for the first one, and 60 heads and 60 tails for the second one. As you can understand that there can be many possible ways we could exchange the x and y observations under the constraint that this row total is 100, this row total is 120, this column total is 110. And this column total is also 110. Therefore, we shall actually be doing a permutations of all the possible values, and then we will try to see what is the probability of the obtained configuration of the table.

(Refer Slide Time: 25:47)



Therefore, in general,

Case1: H_1: $p1 > p2$

Ho should be rejected if $\underline{X > T1}$

Case2: H_1: $\underline{p1 < p2}$

Ho should be rejected if $\underline{X < T2}$

Case3: H_1: $p1 \neq p2$ ✓

Ho should be rejected if $\underline{X > T3}$ or $\underline{X < T4}$

The values of the thresholds depends upon the Level of Significance α.

Therefore, in general for H1 where p1 > p2, then Ho should be rejected if the value of X is greater than some threshold. Similarly, if p1 < p2, we shall reject the null hypothesis if the value of X is less than some threshold T2. And for the two sided case, we will look at whether X is greater than some threshold or X is less than some threshold T4. And the values of the thresholds will depend upon the level of significance α that, of course, all of us understand.

(Refer Slide Time: 26:31)

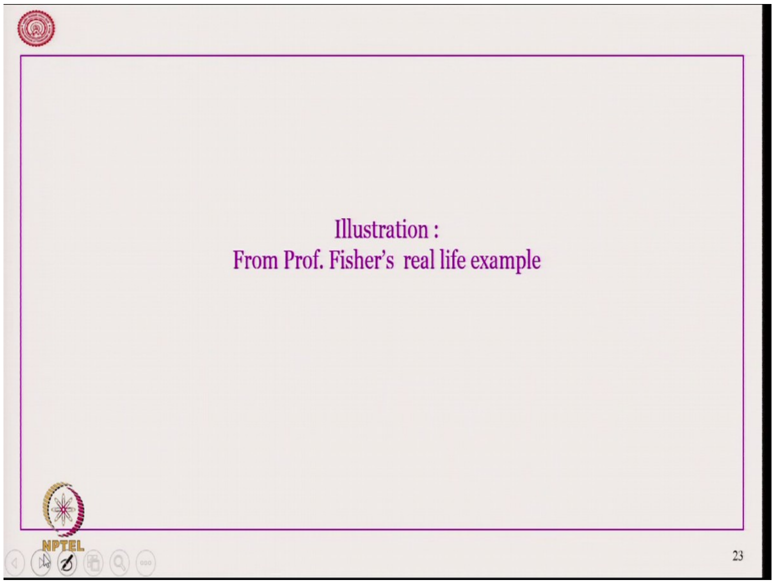
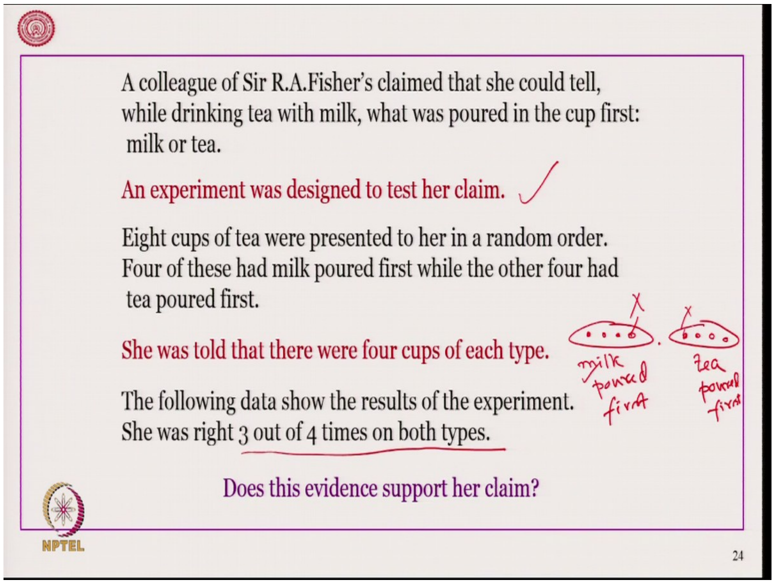


Illustration :
From Prof. Fisher's real life example



A colleague of Sir R.A.Fisher's claimed that she could tell, while drinking tea with milk, what was poured in the cup first: milk or tea.

An experiment was designed to test her claim. ✓

Eight cups of tea were presented to her in a random order. Four of these had milk poured first while the other four had tea poured first.

She was told that there were four cups of each type.

The following data show the results of the experiment. She was right 3 out of 4 times on both types.

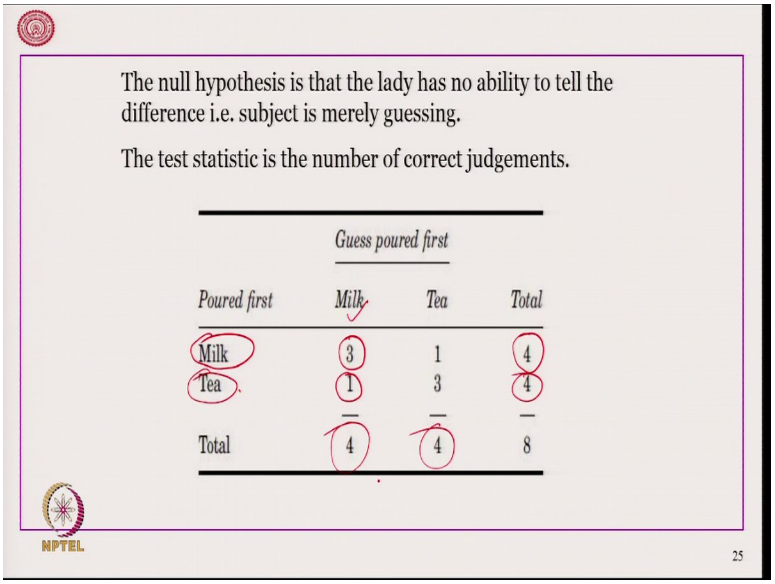
Does this evidence support her claim?

Let me illustrate the same from Professor Fisher's real life example, a colleague of Sir Fisher claimed that she could tell while drinking tea with milk, what is poured in the cup first, it is milk tea. So, a person is drinking a cup of tea with milk. So, while making the tea, one can put milk first in the cup, or one can put tea first in the cup. So, the lady claimed that she can sip and tell which one is poured fast in the cup. So, an experiment was designed to test her claim, 8 cups of tea were presented to her in a random order. 4 of these had milk poured first, while the other 4 had tea poured first.

So, these are the 2 populations, there are 4 cups of tea. And these are 4 populations where tea is first poured. So, this is milk poured first. And here tea poured first. So, she was told that there are 4 cups of each type. And she has to sip and tell whether milk was poured first in that cup or tea, she does not know which cup is coming from which population. So, the following data show that she was right, 3 out of 4 times on both types.

So, does this evidence support our claim? That means out of these 4, she could rightly guess for three, that milk was put first. But suppose for one of them she was wrong. And similarly, out of these 4, she could tell that for these three tea was put first, but for the fourth one or one of them, again, she failed. Does this evidence support her claim? That means we are checking whether she is randomly guessing, or she actually could decide whether milk was put first or tea.

(Refer Slide Time: 29:02)



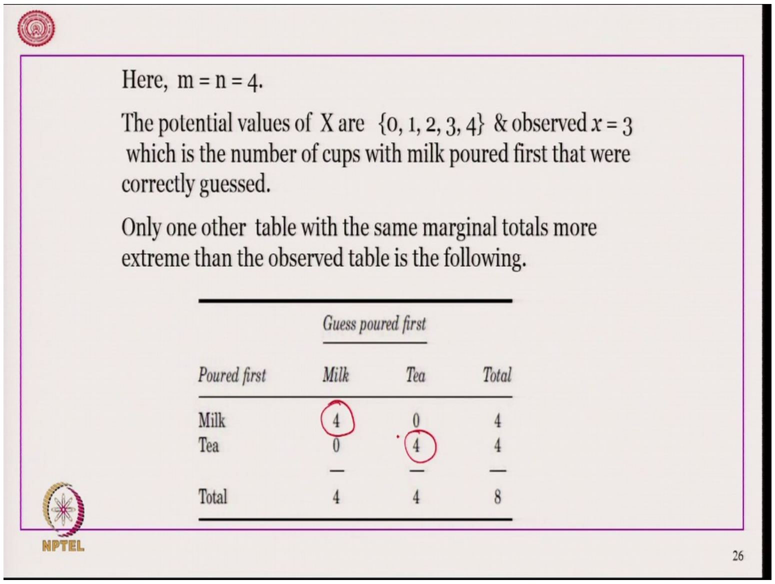
The null hypothesis is that the lady has no ability to tell the difference i.e. subject is merely guessing.

The test statistic is the number of correct judgements.

	Guess poured first		
Poured first	*Milk*	*Tea*	*Total*
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

So, the tables that we get is actually milk is poured first in these 4 cups. Tea was put first in these 4 cups. And she guessed that milk was poured first for 3 out of them and 1 out of these. And naturally the other 2 values are fixed and the column totals are 4 and 4. So very, very simple table, we have to analyze the data to see whether she actually has any such talent.

(Refer Slide Time: 26:44)



Here, m = n = 4.

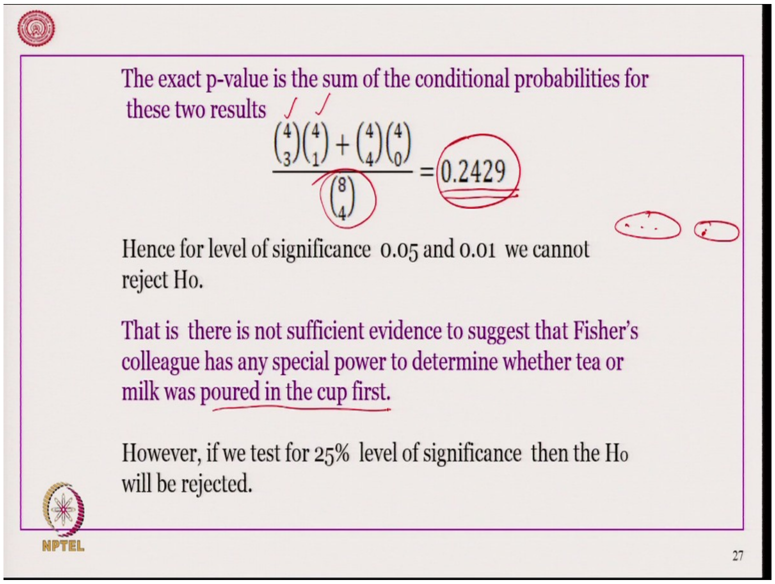
The potential values of X are {0, 1, 2, 3, 4} & observed $x = 3$ which is the number of cups with milk poured first that were correctly guessed.

Only one other table with the same marginal totals more extreme than the observed table is the following.

	Guess poured first		
Poured first	*Milk*	*Tea*	*Total*
Milk	4	0	4
Tea	0	4	4
Total	4	4	8

Now, the potential values of x at 0, 1, 2, 3, 4 and observed is 3 which is the number of cups with milk poured first that were correctly guessed. The only other table with the same marginal totals, more extreme then the observable is the following. That she could guess, all 4 correctly. And she could guess, all 4 correctly. That is 4, 4.

(Refer Slide Time: 30:16)



The exact p-value is the sum of the conditional probabilities for these two results

$$\frac{\binom{4}{3}\binom{4}{1}+\binom{4}{4}\binom{4}{0}}{\binom{8}{4}} = 0.2429$$

Hence for level of significance 0.05 and 0.01 we cannot reject Ho.

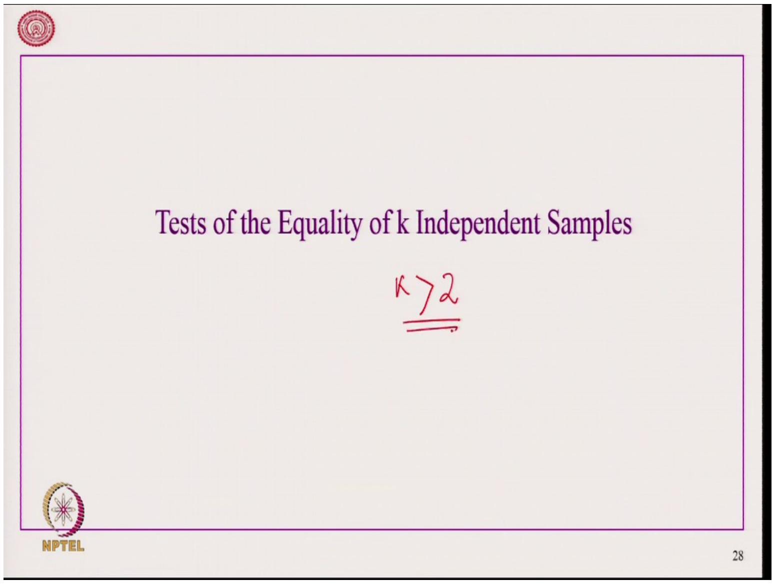
That is there is not sufficient evidence to suggest that Fisher's colleague has any special power to determine whether tea or milk was poured in the cup first.

However, if we test for 25% level of significance then the Ho will be rejected.

And the exact p value is the sum of the conditional probabilities for these two results, which is 0.2429. Because for the first one, it is ${}^4C_3$, multiplied by ${}^4C_1$ out of these 4, she could correctly guess 3, that is $^4C_3$ out of these 4 she could correctly guess for one, she could guess only for 1 that is $^4C_1$ and for the extreme 1, it is going to be $^4C_4$ multiplied by $^4C_0$.

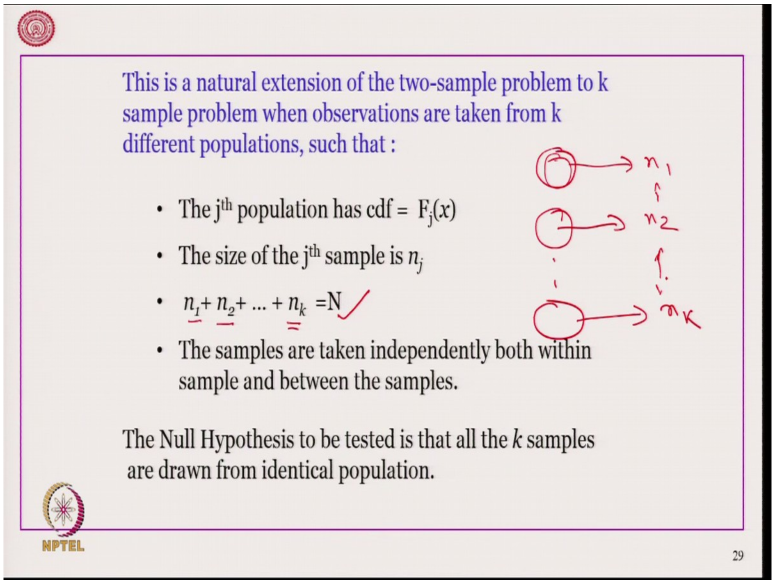
And of course, the total number of possibilities is $^8C_4$ the result is coming out to be 0.2429. Hence, for level of significance 0.05 and 0.01 we cannot reject the null hypothesis, because this value is pretty high that is, there is not sufficient evidence to suggest that Fisher's colleague has any special power to determine whether tea or milk was put first in the cup. However, if we test for 25 percent level of significance, then we could reject the Ho and we could say that actually she is not guessing she really has the talent to identify which one is poured first, ok students.

(Refer Slide Time: 31:52)



Tests of the Equality of k Independent Samples

$\underline{\underline{k > 2}}$



This is a natural extension of the two-sample problem to k sample problem when observations are taken from k different populations, such that :

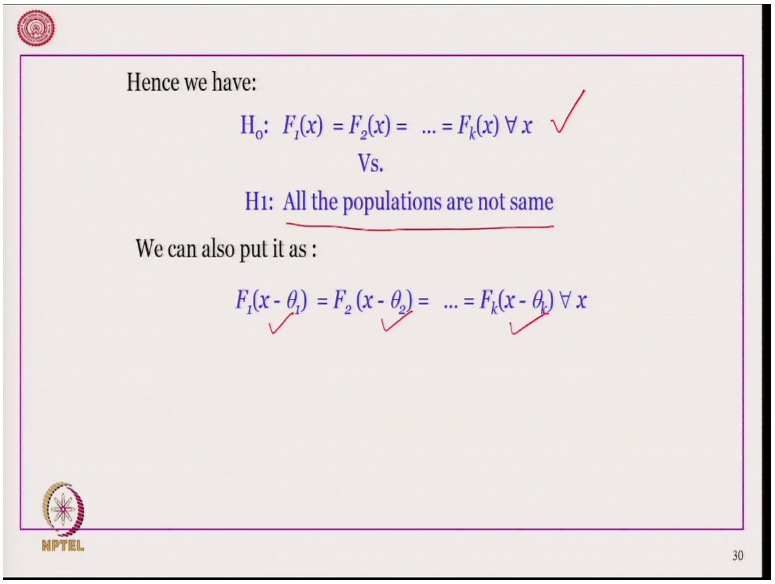
- The jth population has cdf = $F_j(x)$
- The size of the jth sample is n_j
- $\underline{n_1} + \underline{n_2} + \ldots + \underline{\underline{n_k}} = N$ ✓
- The samples are taken independently both within sample and between the samples.

The Null Hypothesis to be tested is that all the k samples are drawn from identical population.

Let us now move to one of the most prominent or important nonparametric test, which is test of equality for k independent samples when value of k is strictly greater than 2. So, it is a natural extension of the 2 sample problem to k sample problem when observations are taken from k different populations such that the jth population has cdf Fj. The size of the jth sample is nj and the total number of sample is n which is n1 + n2 + … + nk. These samples are taken independently both within sample and between the samples, that means, suppose these are the populations we have taken n1 samples from here n2 samples from here and nk samples from here.

So, this sampling is taken independently while taking from the population as well as they are all independent of each other the null hypothesis to be tested is that all the k samples are drawn from identical population. That means, we want to check if all of them have the same distribution.

(Refer Slide Time: 33:20)



Hence we have:

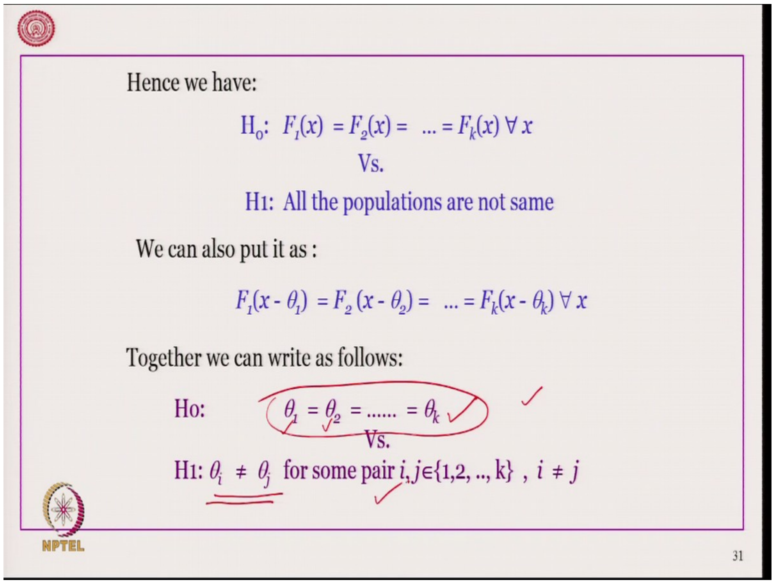
$$H_0:\ F_1(x) = F_2(x) = \ldots = F_k(x)\ \forall x \quad ✓$$

Vs.

H1: $\underline{\text{All the populations are not same}}$

We can also put it as :

$$F_1(x-\theta_1) = F_2(x-\theta_2) = \ldots = F_k(x-\theta_k)\ \forall x$$



Hence we have:

$$H_0:\ F_1(x) = F_2(x) = \ldots = F_k(x)\ \forall x$$

Vs.

H1: All the populations are not same

We can also put it as :

$$F_1(x-\theta_1) = F_2(x-\theta_2) = \ldots = F_k(x-\theta_k)\ \forall x$$

Together we can write as follows:

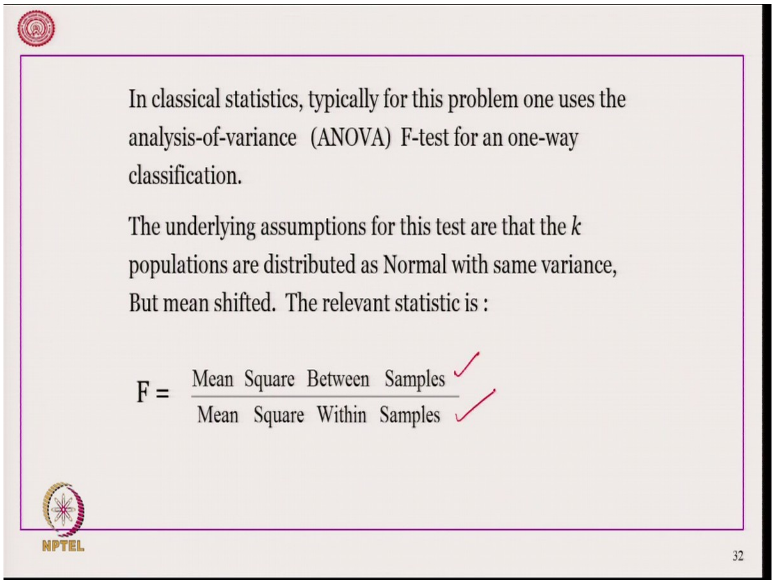
Ho: $\theta_1 = \theta_2 = \ldots\ldots = \theta_k$ ✓

Vs.

H1: $\underline{\theta_i \neq \theta_j}$ for some pair $i, j \in \{1,2,..,k\}$, $i \neq j$

Hence, we have the Ho is that F1(x) = F2(x) = .. = Fk(x) for all x versus all populations are not same, that means, they do not have the same distribution. So, we can rewrite it in the form F1(x) - θ1 = F2(θ1 = x) - θ2 = .. = Fk(x) - θk such that under the null hypothesis θ1 = θ2 = … θk, and under the alternative there has to be at least 1 pair θi θj which are not the same. Because, if they are same for all of them, essentially we are getting Ho. So, this is the null hypothesis, this is the alternative. We are testing if we can accept or reject Ho.

(Refer Slide Time: 34:30)



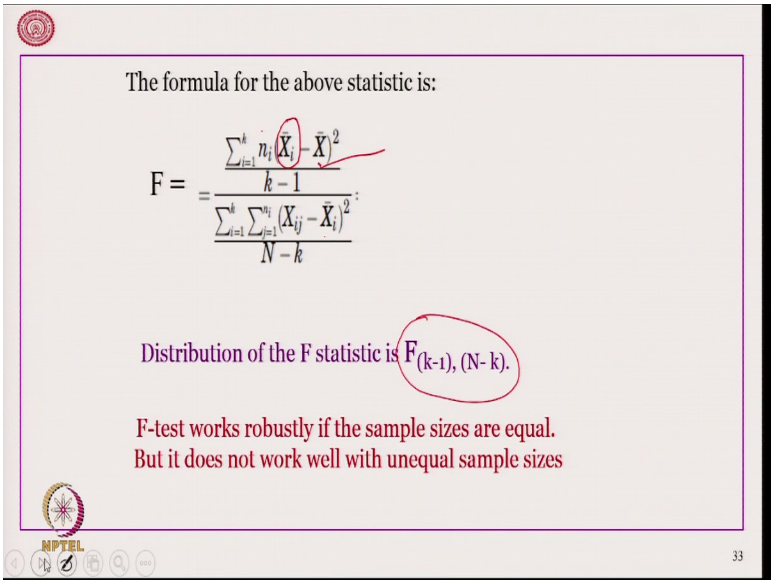
In classical statistics, typically for this problem one uses the analysis-of-variance (ANOVA) F-test for an one-way classification.

The underlying assumptions for this test are that the *k* populations are distributed as Normal with same variance, But mean shifted. The relevant statistic is :

$$F = \frac{\text{Mean Square Between Samples} \checkmark}{\text{Mean Square Within Samples} \checkmark}$$

In classical statistics typically, for this problem 1 uses analysis of variance if test for one-way classification. The underlying assumptions for this test are that the k populations are distributed as normal as you all know that under parametric setup, we assume normality with same variance, but mean shifted and the relevant statistic for them is F statistic, which is the mean square between the samples divided by mean square within the samples.

(Refer Slide Time: 35:06)



The formula for the above statistic is:

$$F = \frac{\dfrac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{k-1}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{N-k}}$$

Distribution of the F statistic is $F_{(k-1),(N-k)}$.

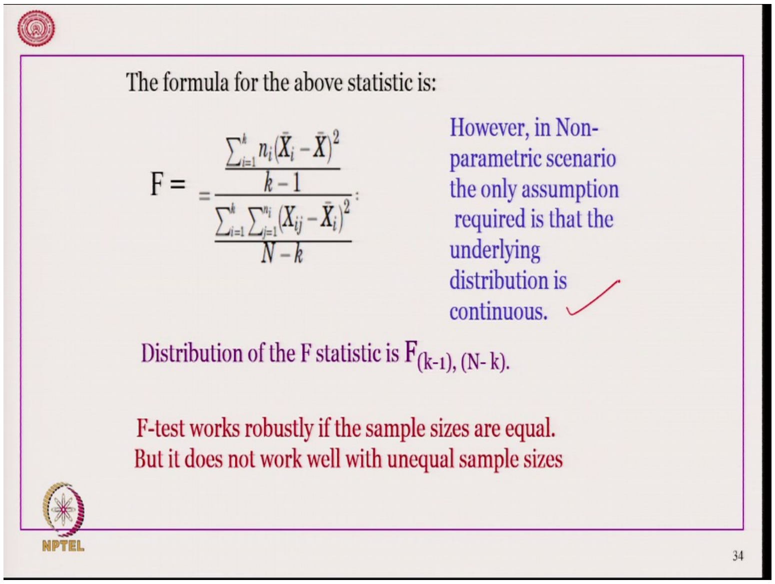
F-test works robustly if the sample sizes are equal.
But it does not work well with unequal sample sizes

What does it mean? It means the following that for the ith population, we are looking at the mean of the ith sample which is $\overline{X_i}$ and we are looking how much different it is from the overall sample mean considering all the samples. So, $(\overline{X_i} - \overline{X})^2$ and that we are multiplying with a ni because there are ni many observations from the i th population and that we are summing it over i = 1 to k.

And we are dividing by k - 1 because all k of them are not independent because the total population is constant. So, therefore, if we know n1 n2 up to n k-1, then nk is fixed and in the denominator we have the overall sample variance $\sum_i \sum \left(X_{i,j} - \overline{X_i}\right)^2$ that is the variance for the i th population and that we are summing for j is equal to 1 to a ni for the i th population and i is equal to 1 to k.

We are not going into the details, but the corresponding statistic will follow F distribution with k-1 and n-k as its parameters. Now, F test works robustly, if the sample sizes are equal. But it does not work well with unequal sample sizes. That is 1 of the problem of parametric setup on the top of the normality.

(Refer Slide Time: 36:50)



The formula for the above statistic is:

$$F = \frac{\dfrac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2}{k-1}}{\dfrac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2}{N-k}}$$

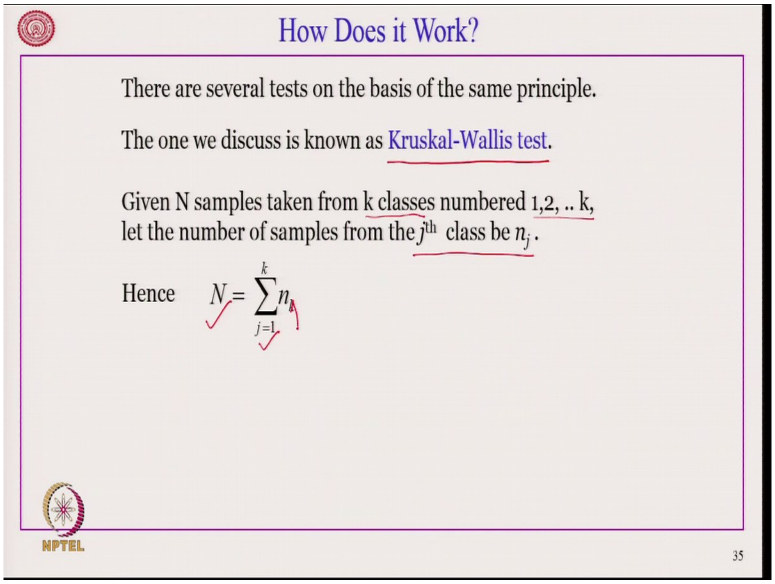
However, in Non-parametric scenario the only assumption required is that the underlying distribution is continuous.

Distribution of the F statistic is $F_{(k-1),(N-k)}$.

F-test works robustly if the sample sizes are equal. But it does not work well with unequal sample sizes

However, in nonparametric scenario, the only assumption required is that the underlying distribution is continuous. That is one assumption that we have to make for nonparametric treatment of the above problem.

(Refer Slide Time: 37:04)



How Does it Work?

There are several tests on the basis of the same principle.

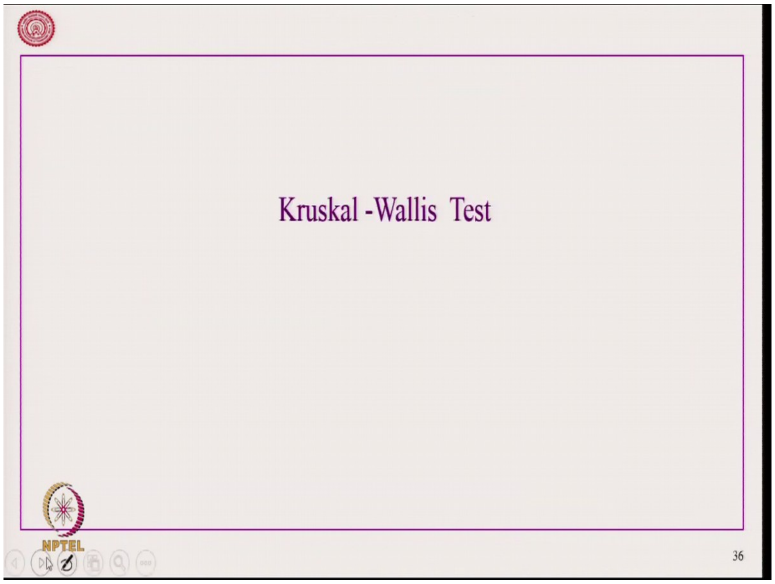
The one we discuss is known as Kruskal-Wallis test.

Given N samples taken from k classes numbered 1,2, .. k, let the number of samples from the j^{th} class be n_j.

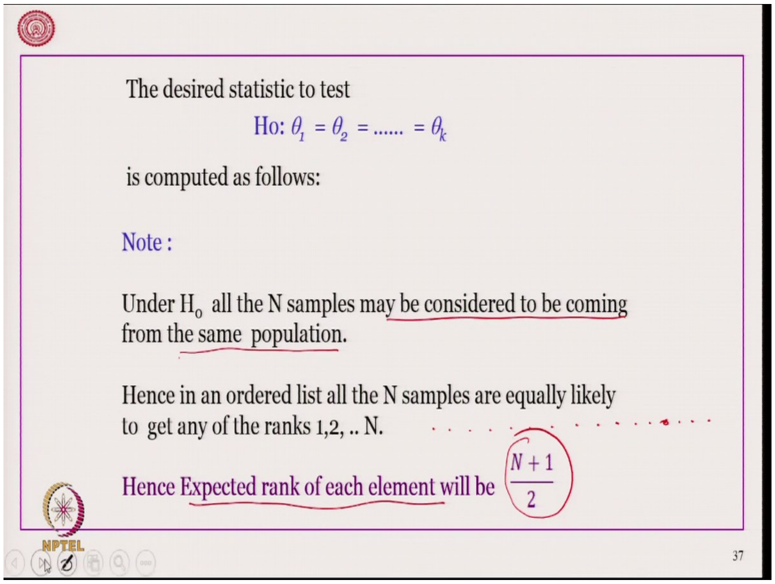
Hence $N = \sum_{j=1}^{k} n_j$

So, there are several tests on the basis of the same principle. The first we are going to discuss is called the Kruskal-Wallis test, as I said, it is a very important nonparametric test and it works as follows. Given n samples taken from k classes, numbered 1, 2, 3 up to k. Let the number of samples from the jth class be nj as we have already defined and therefore, capital N the total number of samples is equal to summation over j = 1 to k, nj.

(Refer Slide Time: 37:46)



Kruskal -Wallis Test



The desired statistic to test

Ho: $\theta_1 = \theta_2 = \ldots\ldots = \theta_k$

is computed as follows:

Note :

Under H_0 all the N samples may be considered to be coming from the same population.

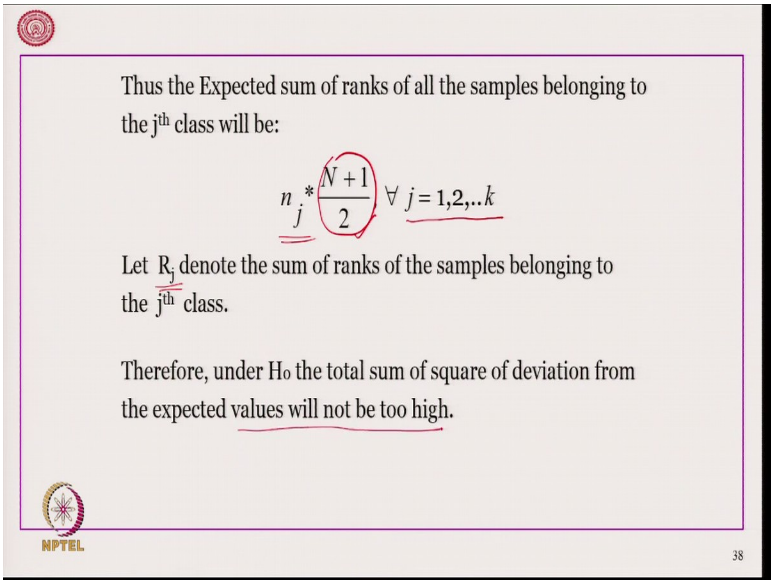
Hence in an ordered list all the N samples are equally likely to get any of the ranks 1,2, .. N.

Hence Expected rank of each element will be $\frac{N+1}{2}$

What we do in the Kruskal-Wallis test, we do the following we are testing if θ1 = θ2 = … θk and it is computed as follows. Under Ho all the n samples may be considered to be coming from the same population as you all know. Hence, in an ordered list, all the n samples are equally likely to get any of the ranks 1, 2, 3 up to n that means, what we are doing?

We are arranging all the capital N many observations in ordered way. Now, if all the distributions are equal, then any 1 of them can come from any 1 of the population. Hence, the expected rank of each element is going to be (N +1) / 2 that we all know there are capital N many observations and therefore, the mean is going to be (N +1) / 2.

(Refer Slide Time: 38:52)



Thus the Expected sum of ranks of all the samples belonging to the j^{th} class will be:

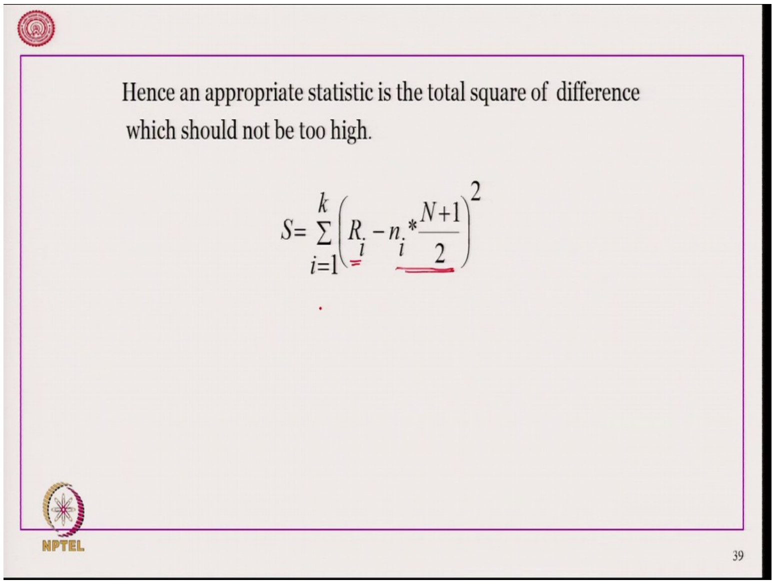
$$n_j * \frac{N+1}{2} \quad \forall\ j = 1,2,..k$$

Let R_j denote the sum of ranks of the samples belonging to the j^{th} class.

Therefore, under Ho the total sum of square of deviation from the expected values will not be too high.

Therefore, the expected sum of ranks of all the samples belonging to the j th class is going to be nj multiplied by (N +1) / 2 for all j = to 1, 2, 3 up to k. Because each element has the average rank this under the null hypothesis, therefore, the total sum of rank is expected to be nj * (N +1) / 2. Now, let us calculate actual rank that we have got from the sample. Let Rj denote the sum of ranks of the samples belonging to the jth class. Therefore, under Ho the total sum of square of deviation from the expected values will not be too high.

(Refer Slide Time: 39:48)

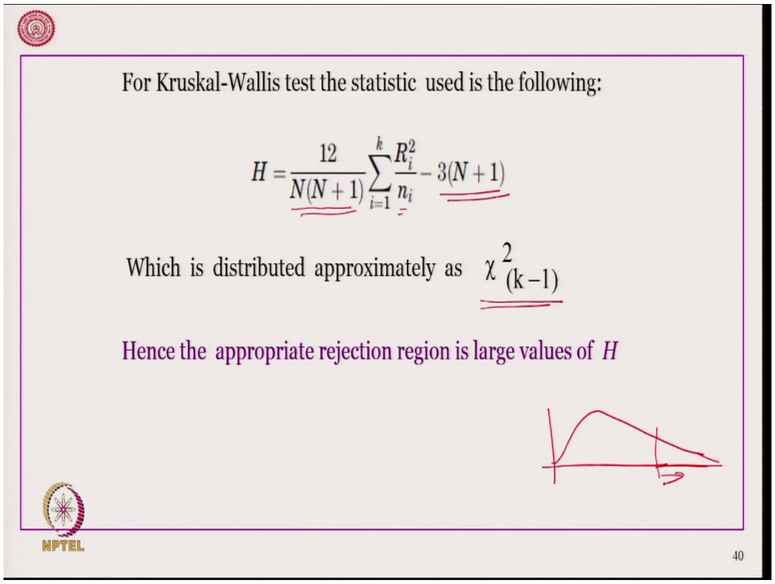


Hence an appropriate statistic is the total square of difference which should not be too high.

$$S = \sum_{i=1}^{k} \left(R_i - n_i * \frac{N+1}{2} \right)^2$$

Therefore, an appropriate statistic is going to be how much the expected sum of rank is different from the obtained sum of rank. So, we are making the sum of the square of the differences for i = 1 to k. So, that is in some sense giving us the idea of how much it is varying from equality of distribution.

(Refer Slide Time: 40:18)



For Kruskal-Wallis test the statistic used is the following:

$$H = \frac{12}{\underline{N(N+1)}}\sum_{i=1}^{k}\frac{R_i^2}{\underline{n_i}} - \underline{3(N+1)}$$

Which is distributed approximately as $\underline{\chi^2_{(k-1)}}$

Hence the appropriate rejection region is large values of H

NPTEL

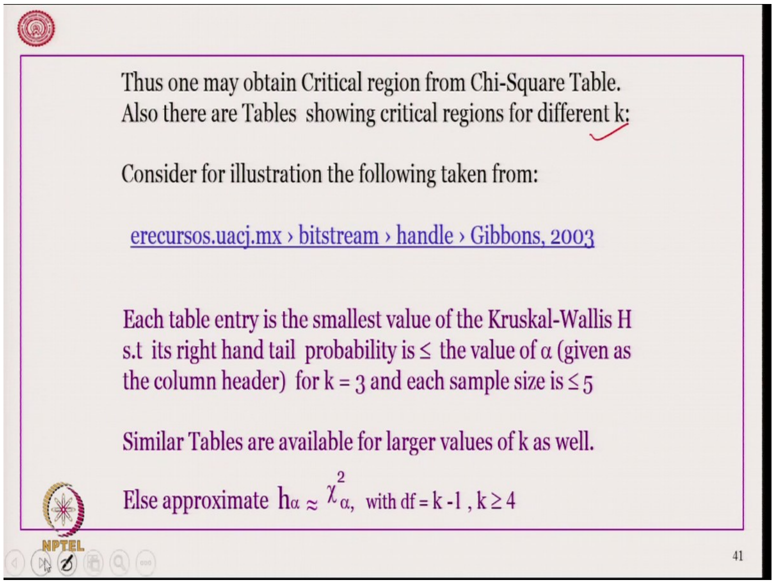
40

For Kruskal-Wallis test, the statistic used is

$$\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1)$$

Let us remember that, we are not going to derive it and these can be shown that is distributed approximately as chi square with k - 1 degrees of freedom. As you all know, that chi square is having this type of shape. Therefore, we shall reject the null hypothesis if the value obtained is too high, that is also intuitively clear, because somewhere we are looking at Ri minus the expected sum of rank whole square. So, if that value is too high, then we are going to reject the null hypothesis.

(Refer Slide Time: 41:16)



Thus one may obtain Critical region from Chi-Square Table.
Also there are Tables showing critical regions for different k:

Consider for illustration the following taken from:

erecursos.uacj.mx › bitstream › handle › Gibbons, 2003

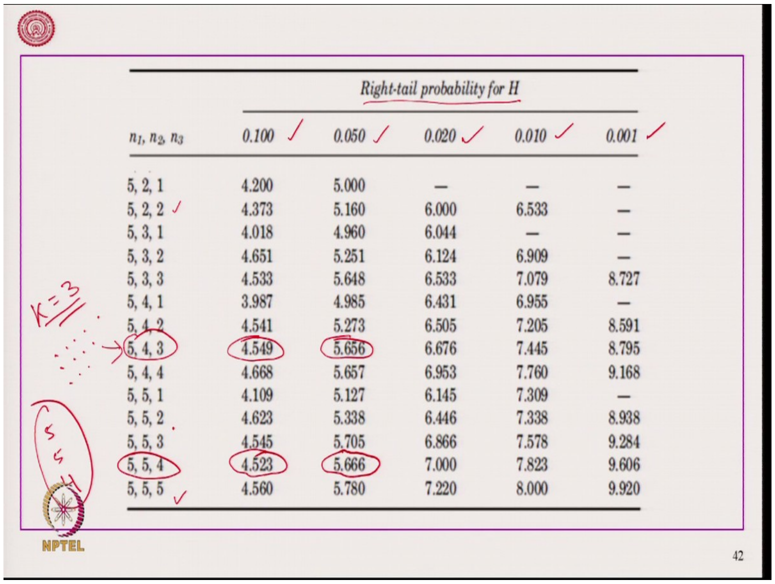
Each table entry is the smallest value of the Kruskal-Wallis H s.t its right hand tail probability is ≤ the value of α (given as the column header) for k = 3 and each sample size is ≤ 5

Similar Tables are available for larger values of k as well.

Else approximate $h_\alpha \approx \chi^2_\alpha$, with df = k - 1 , k ≥ 4

Thus, one may obtain critical region from the chi square table. Also, there are tables showing the critical regions for different k, Kruskal-Wallis tables are available. Consider for illustration the following taken from this source, let us go to the table first so, it looks like this.

(Refer Slide Time: 41:39)



	Right-tail probability for H				
n_1, n_2, n_3	0.100	0.050	0.020	0.010	0.001
5, 2, 1	4.200	5.000	—	—	—
5, 2, 2	4.373	5.160	6.000	6.533	—
5, 3, 1	4.018	4.960	6.044	—	—
5, 3, 2	4.651	5.251	6.124	6.909	—
5, 3, 3	4.533	5.648	6.533	7.079	8.727
5, 4, 1	3.987	4.985	6.431	6.955	—
5, 4, 2	4.541	5.273	6.505	7.205	8.591
5, 4, 3	4.549	5.656	6.676	7.445	8.795
5, 4, 4	4.668	5.657	6.953	7.760	9.168
5, 5, 1	4.109	5.127	6.145	7.309	—
5, 5, 2	4.623	5.338	6.446	7.338	8.938
5, 5, 3	4.545	5.705	6.866	7.578	9.284
5, 5, 4	4.523	5.666	7.000	7.823	9.606
5, 5, 5	4.560	5.780	7.220	8.000	9.920

k = 3

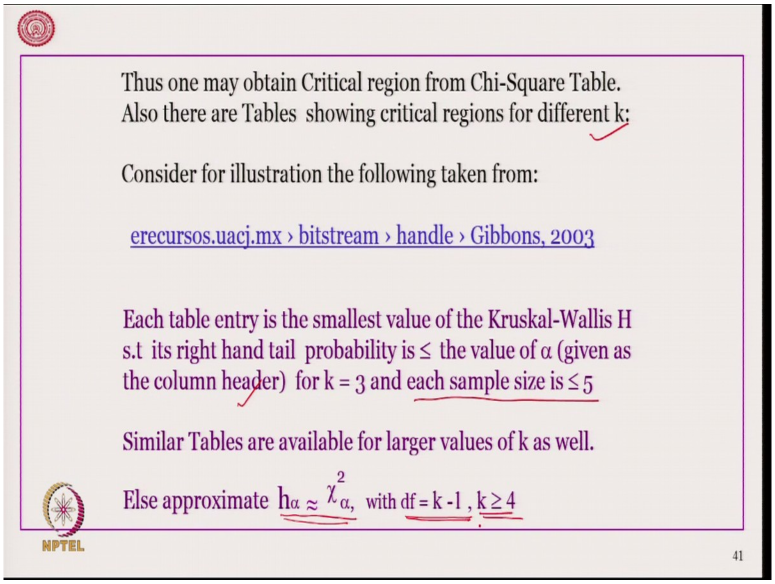
5 5 4

These are the alpha for several values of alpha the values have been obtained the right tail probability for H and the segment that I am showing it is for k = 3 and the highest number of observation for at least 1 class is 5 and therefore, these are different combinations 5 2 1, 5 2 2 up to 5 5 5.

In particular, say consider 5 5 4. What does it mean? That means that there are 5 observations from 1 class 5 observations from another class and 4 observations from the third class. If that is the type of observation, then the critical value for 5% level of significance is 5.666 and for 1% level of significance is 4.523. Similarly, for any combination of this type for k is equal to 3 and the highest observation from any class is 5.

For example, 5 4 3 that means, there are 5 observation in 1 class 4 observation in another class and 3 observation is another class. Then the critical values are going to be for 5% level of significance 5.656 at 5 percent level and 4.549 at 1% level. So, that is how we need to read the table.

(Refer Slide Time: 43:32)



Thus one may obtain Critical region from Chi-Square Table.
Also there are Tables showing critical regions for different k:

Consider for illustration the following taken from:

erecursos.uacj.mx › bitstream › handle › Gibbons, 2003

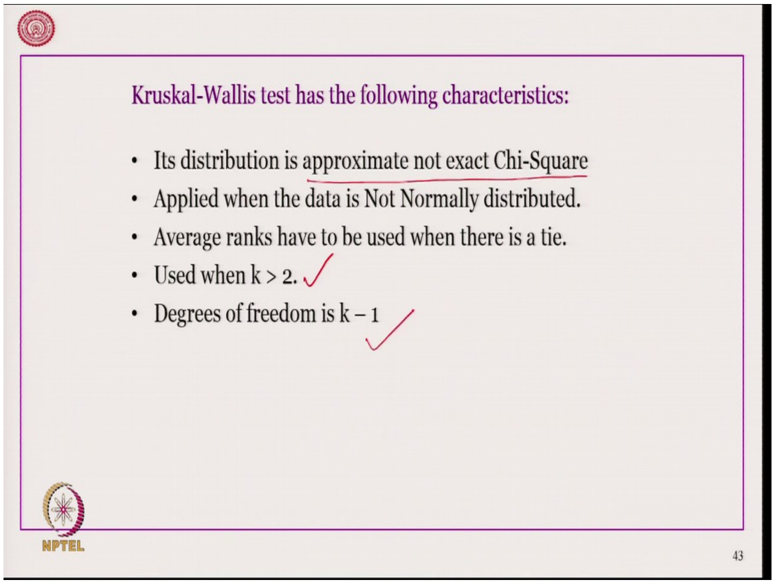
Each table entry is the smallest value of the Kruskal-Wallis H s.t its right hand tail probability is ≤ the value of α (given as the column header) for k = 3 and each sample size is ≤ 5

Similar Tables are available for larger values of k as well.

Else approximate $h_\alpha \approx \chi^2_\alpha$, with df = k - 1 , k ≥ 4

Now, let us go back here, each table entry is the smallest value of the Kruskal-Wallis each such that its right hand tail probability is ≤ the value of α given as the column header for k = 3 and each sample size is ≤ 5. Similar tables are available for larger values of k also, but, in practice, 1 can approximate that with a chi square distribution. Therefore, the critical values obtained for a chi square table can be used here, when a for a particular case the degrees of freedom is going to be k - 1 when case is 4.

(Refer Slide Time: 44:22)

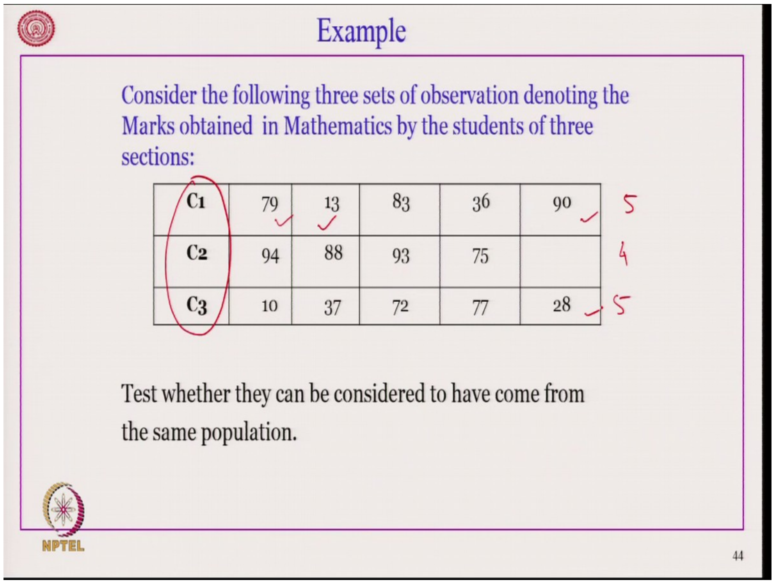


Kruskal-Wallis test has the following characteristics:

- Its distribution is approximate not exact Chi-Square
- Applied when the data is Not Normally distributed.
- Average ranks have to be used when there is a tie.
- Used when k > 2.
- Degrees of freedom is k – 1

Kruskal-Wallis test has the following characteristic. Its distribution is approximately, but not exactly chi square, you have to remember that it is only an approximation to chi square, it is applied when the data is not normally distributed as we have mentioned many times. Average ranks have to be used, when there is a tie, if two observations are coming to be the same, then as usual, we have to give them the average length. It is used when k is greater than 2 and degrees of freedom is going to be k - 1.

(Refer Slide Time: 45:01)



Example

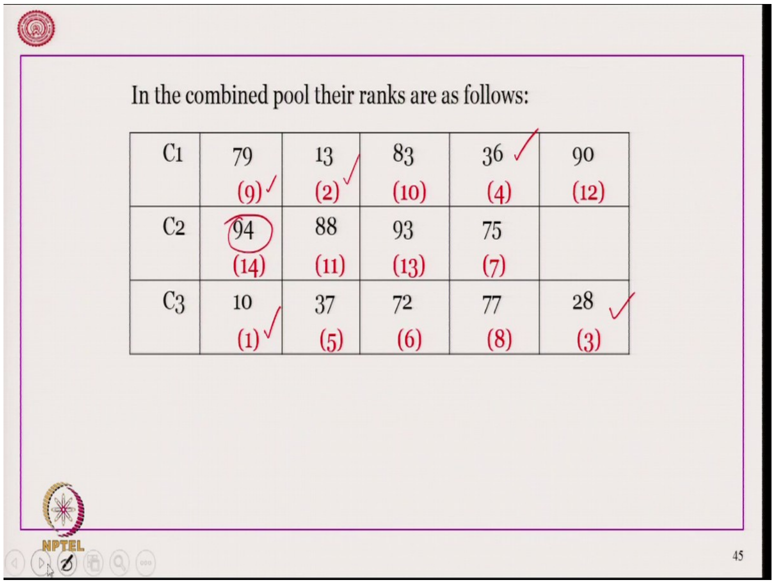
Consider the following three sets of observation denoting the Marks obtained in Mathematics by the students of three sections:

C1	79	13	83	36	90	5
C2	94	88	93	75		4
C3	10	37	72	77	28	5

Test whether they can be considered to have come from the same population.

So, let us give an example. Consider 3 class problems C1, C2, C3, there are 5 observations from 1 class, 5 observation from the second class and 4 observations from the third one and the values given are like this 79, 13 up to 90 and you can see that these are the 14 values given 5 + 4 + 5 as coming out from that 3 classes. We need to test whether these can be considered to have come from the same population.

(Refer Slide Time: 45:38)

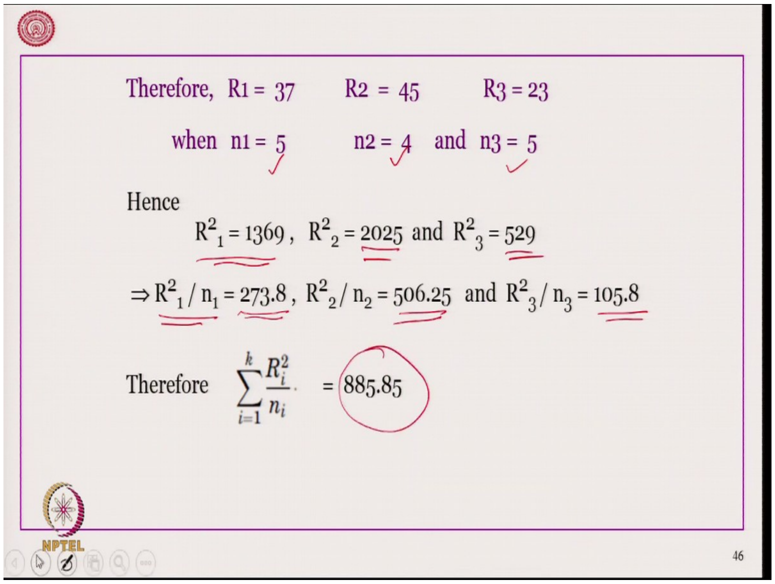


In the combined pool their ranks are as follows:

C1	79 (9) ✓	13 (2) ✓	83 (10)	36 ✓ (4)	90 (12)
C2	94 (14)	88 (11)	93 (13)	75 (7)	
C3	10 (1) ✓	37 (5)	72 (6)	77 (8)	28 ✓ (3)

So, what we did we have ranked the combined population and under this ranking, these are the ranks. This is getting the rank 9, 10 is the smallest therefore, it is getting the rank 1. And similarly, you can see 13 is getting the rank 2, 28 is getting the rank 3, 36 is getting the rank 4 like that highest 1 is 94 which is getting the rank 14.

(Refer Slide Time: 46:12)



Therefore, R1 = 37 R2 = 45 R3 = 23

when n1 = 5 n2 = 4 and n3 = 5

Hence

$R^2_1 = 1369$, $R^2_2 = 2025$ and $R^2_3 = 529$

$\Rightarrow R^2_1 / n_1 = 273.8$, $R^2_2 / n_2 = 506.25$ and $R^2_3 / n_3 = 105.8$

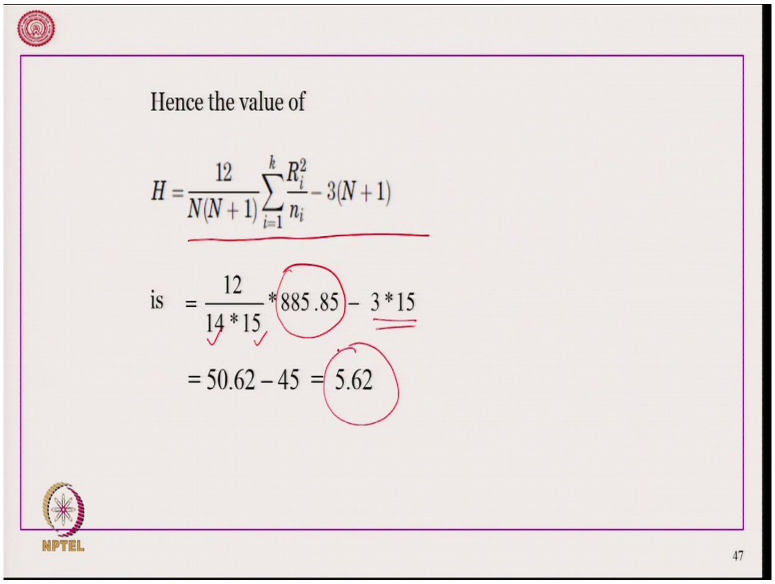
Therefore $\sum_{i=1}^{k} \frac{R_i^2}{n_i} = 885.85$

Therefore, R1 the sum of ranks for the first set is 37, R2 is 45 and R3 is 23 when n1 = 5, n2 = 4 and n3 = 5. therefore, $R1^2$ is coming out to be 1369, $R2^2$ is coming out to be 2025 and $R3^2$ is coming out to be 529. Therefore, we are dividing them by corresponding number of observations. Therefore, $R1^2$ / n1 = 273.8, $R2^2$ / n2 = 506.25 and $R3^2$ / n3 = 105.8 and therefore, the

$$\sum \frac{R_i^2}{n_i}$$

is coming out to be 885.85.

(Refer Slide Time: 47:05)



Hence the value of

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

is $= \frac{12}{14 * 15} * 885.85 - 3 * 15$

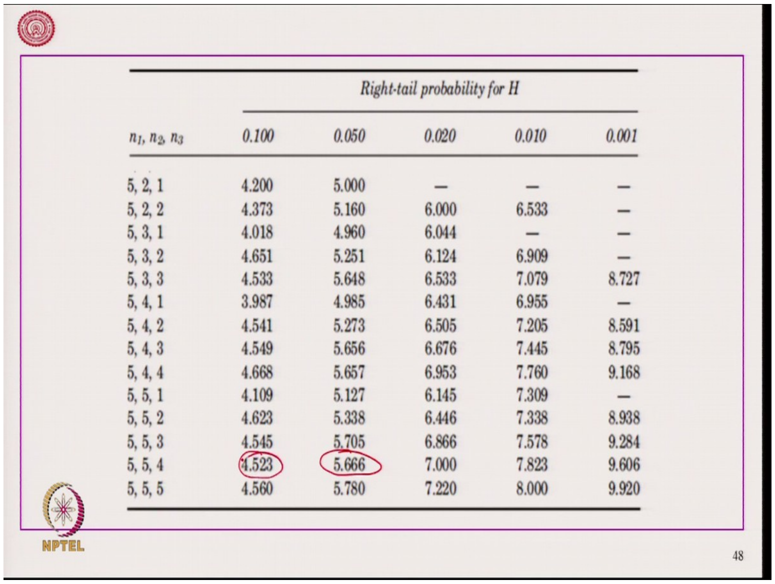
$= 50.62 - 45 = 5.62$

Therefore, the value of the Kruskal-Wallis statistic H for H the formula is given already I have given it to you is coming out to be

$$\frac{12}{14 * 15} * 885.85 - 3 * 15$$

which when we simplify, we are getting 5.62.

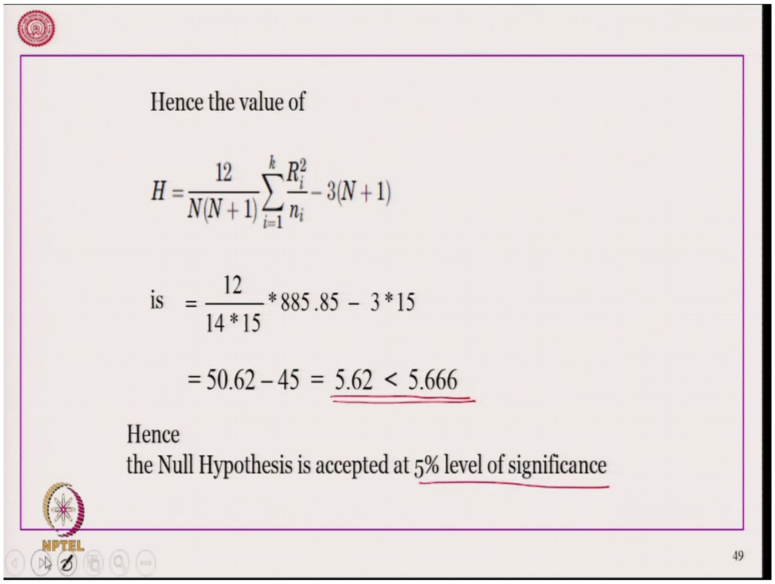
(Refer Slide Time: 47:35)



	Right-tail probability for H				
n_1, n_2, n_3	0.100	0.050	0.020	0.010	0.001
5, 2, 1	4.200	5.000	—	—	—
5, 2, 2	4.373	5.160	6.000	6.533	—
5, 3, 1	4.018	4.960	6.044	—	—
5, 3, 2	4.651	5.251	6.124	6.909	—
5, 3, 3	4.533	5.648	6.533	7.079	8.727
5, 4, 1	3.987	4.985	6.431	6.955	—
5, 4, 2	4.541	5.273	6.505	7.205	8.591
5, 4, 3	4.549	5.656	6.676	7.445	8.795
5, 4, 4	4.668	5.657	6.953	7.760	9.168
5, 5, 1	4.109	5.127	6.145	7.309	—
5, 5, 2	4.623	5.338	6.446	7.338	8.938
5, 5, 3	4.545	5.705	6.866	7.578	9.284
5, 5, 4	4.523	5.666	7.000	7.823	9.606
5, 5, 5	4.560	5.780	7.220	8.000	9.920

As I have already shown to you that this value is coming out to be 5.666 that is the critical value at 5% level and 4.523 the critical value at 1% level.

(Refer Slide Time: 47:56)



Hence the value of

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

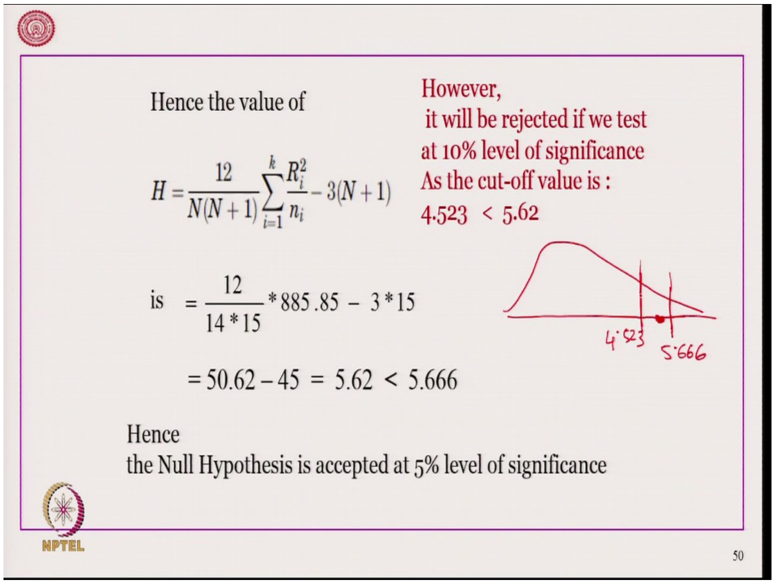
is $= \frac{12}{14 * 15} * 885.85 - 3 * 15$

$= 50.62 - 45 = \underline{\underline{5.62 < 5.666}}$

Hence
the Null Hypothesis is accepted at 5% level of significance

Therefore, what decision we can arrive at is the following that since 5.62 is less than 5.666. Therefore, the null hypothesis is accepted at 5% level of significance.

(Refer Slide Time: 48:07)



Hence the value of

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

is $= \frac{12}{14 * 15} * 885.85 - 3 * 15$

$= 50.62 - 45 = 5.62 < 5.666$

Hence
the Null Hypothesis is accepted at 5% level of significance

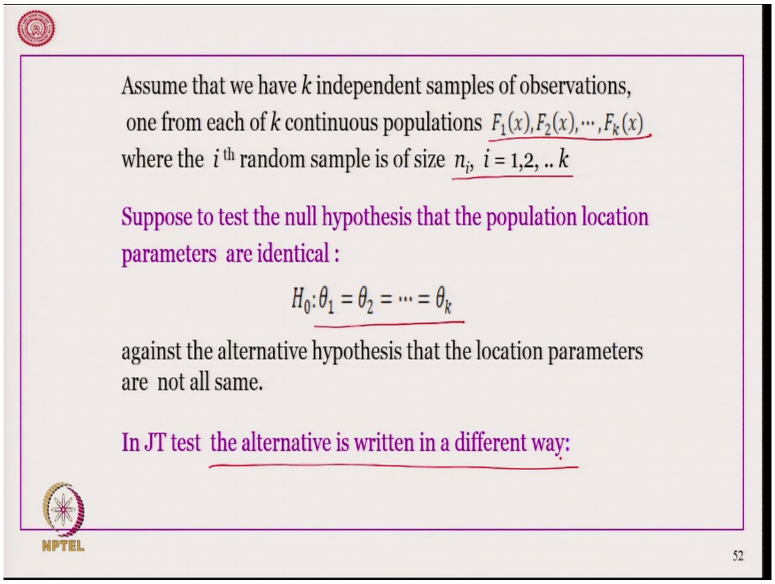
However,
it will be rejected if we test
at 10% level of significance
As the cut-off value is :
4.523 < 5.62

However, it will be rejected, if we test as 10% level of significance as the cutoff value there is 4.523 that is less than 5.62. So, basically what we are saying we get a chi square type of distribution at 5% level, we have 5.666, at 10% level the cutoff is 4.523 and our observed value is somewhere here. Therefore, we cannot reject the null hypothesis at 5% level of significance, but we can reject the null hypothesis at 10% level significance.

(Refer Slide Time: 48:55)

JT Test

Jocnckheere and Terpstra



Assume that we have k independent samples of observations, one from each of k continuous populations $F_1(x), F_2(x), \cdots, F_k(x)$ where the i^{th} random sample is of size n_i, $i = 1, 2, .. k$

Suppose to test the null hypothesis that the population location parameters are identical :

$$H_0 : \theta_1 = \theta_2 = \cdots = \theta_k$$

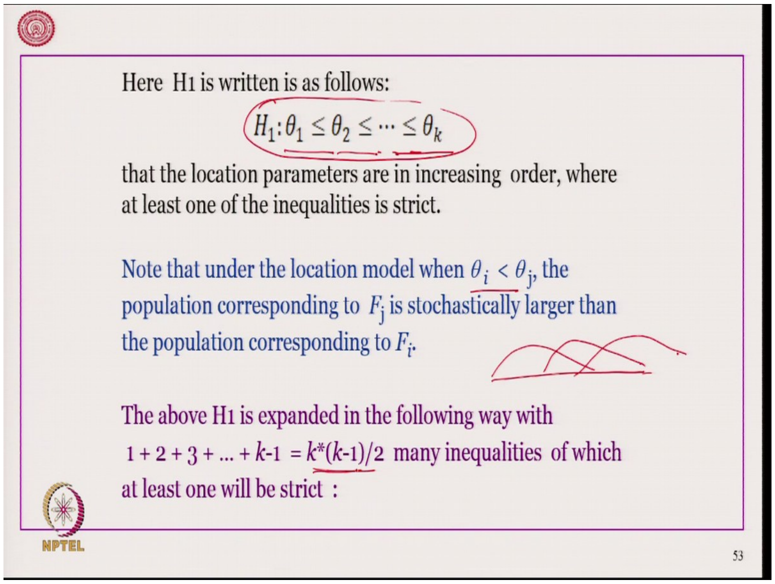
against the alternative hypothesis that the location parameters are not all same.

In JT test the alternative is written in a different way:

Let me now talk about another test which is called JT test named after these two scientists, it is similar to Kruskal but there is a difference in the treatment of that alternative. So, the set of is same, we have ni many observations from the ith population i = 1 to k and we have k many continuous populations with the corresponding distribution functions as F1(x), F2(x) up to Fk(x).

Suppose, we want to test whether $\theta 1 = \theta 2 = \ldots \theta k$, it is the alternative hypothesis that the location parameters are not all same. Therefore, in JT test the alternative is written in a different way as I have already said.

(Refer Slide Time: 49:48)



Here H1 is written is as follows:

$$H_1: \theta_1 \le \theta_2 \le \cdots \le \theta_k$$

that the location parameters are in increasing order, where at least one of the inequalities is strict.

Note that under the location model when $\theta_i < \theta_j$, the population corresponding to F_j is stochastically larger than the population corresponding to F_i.

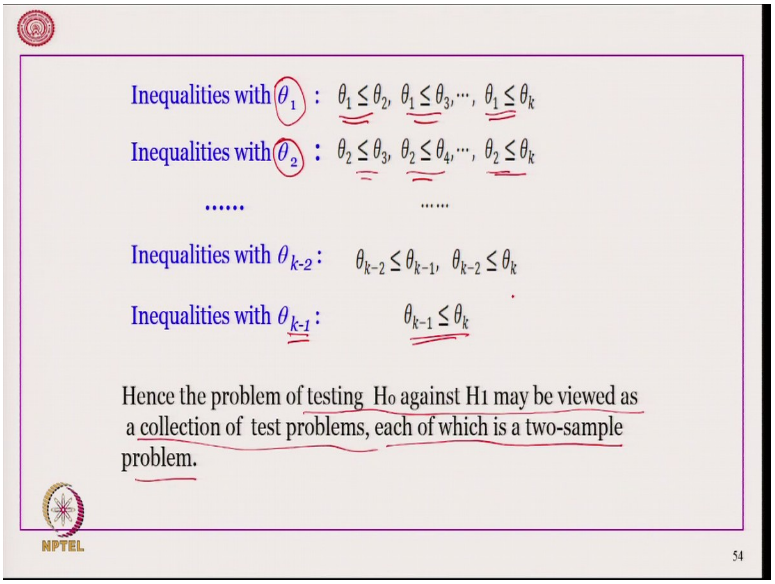
The above H1 is expanded in the following way with $1 + 2 + 3 + \ldots + k\text{-}1 = k^*(k\text{-}1)/2$ many inequalities of which at least one will be strict :

Here alternative is written as follows H1 is that $\theta 1 \le \theta 2 \le \theta 3 \ldots \le \theta k$

That means, we are looking at the location parameters in an increasing order of magnitude. Note that under the location model, when $\theta i \le \theta j$, the population corresponding to Fj is stochastically larger than the population corresponding to Fi. So, pictorially what you are looking at? Suppose, this is how F1 is distributed and similarly then F2 is distributed something like this, which is stochastically larger than that 1. Similarly, for F3 which may be somewhat like this.

Like that the distribution of the k populations, we are trying to check if the null hypothesis can be rejected against this type of alternative. These alternatives, now actually what we do we divide it into $k * (k - 1) / 2$ many inequalities of two populations.

(Refer Slide Time: 51:08)



Inequalities with θ_1 : $\theta_1 \le \theta_2,\ \theta_1 \le \theta_3, \cdots,\ \theta_1 \le \theta_k$

Inequalities with θ_2 : $\theta_2 \le \theta_3,\ \theta_2 \le \theta_4, \cdots,\ \theta_2 \le \theta_k$

…… ……

Inequalities with θ_{k-2} : $\theta_{k-2} \le \theta_{k-1},\ \theta_{k-2} \le \theta_k$

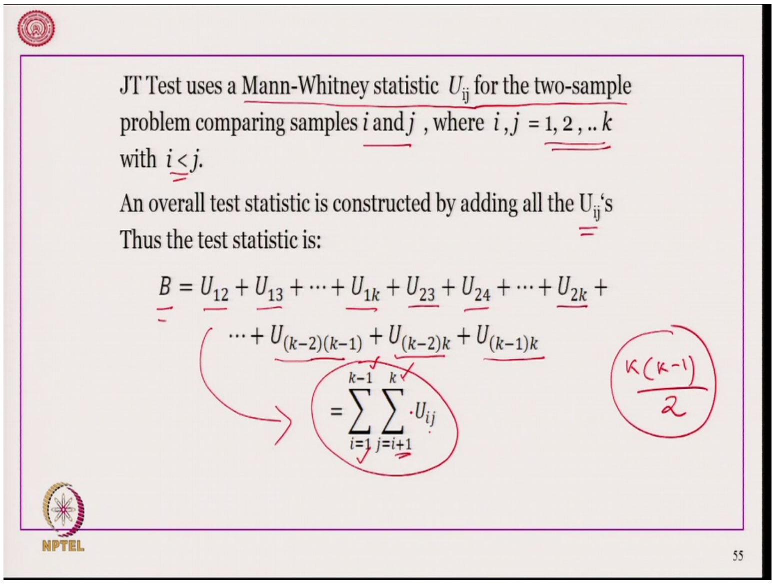
Inequalities with θ_{k-1} : $\theta_{k-1} \le \theta_k$

Hence the problem of testing H_0 against H_1 may be viewed as a collection of test problems, each of which is a two-sample problem.

We were doing it in the following way. We were looking at $\theta_1 \le \theta_2$, $\theta_1 \le \theta_3$ … $\theta_1 \le \theta_k$. So, these are the inequalities corresponding to θ_1. With respect to θ_2, what we are trying to check? If $\theta_2 \le \theta_3$, $\theta_2 \le \theta_4$ … $\theta_2 \le \theta_k$. Finally, corresponding to θ_{k-1}, you are just taking $\theta_{k-1} \le \theta_k$. That is what we ever say that each time you are just considering only two populations, and we are trying to compare the allocation parameters.

Here is the problem of testing Ho against H1 may be viewed as a collection of test problems, each of which is a two sample problem. I hope that concept is clear. And of course, we have to understand that at least one of the inequalities have to be strict, because if all inequalities become equality, then it is nothing but the null hypothesis.

(Refer Slide Time: 52:22)



JT Test uses a Mann-Whitney statistic U_{ij} for the two-sample problem comparing samples i and j , where $i, j = 1, 2, .. k$ with $i < j$.

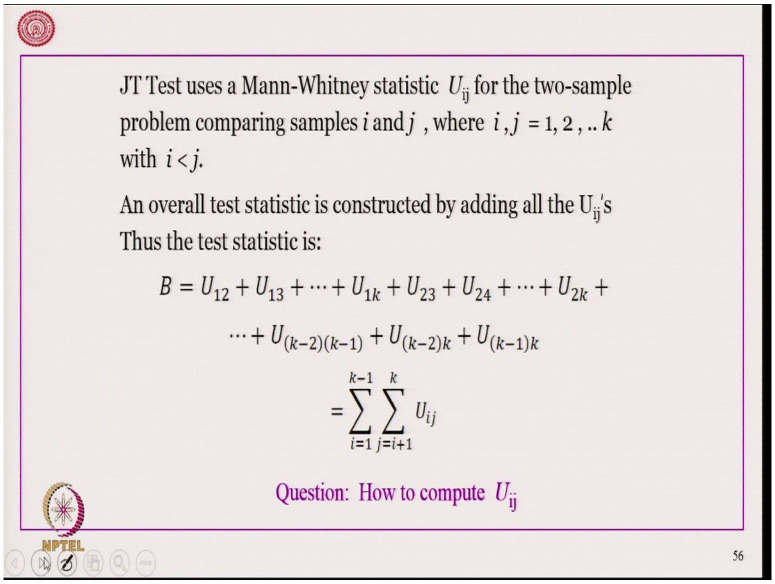
An overall test statistic is constructed by adding all the U_{ij}'s

Thus the test statistic is:

$$B = U_{12} + U_{13} + \cdots + U_{1k} + U_{23} + U_{24} + \cdots + U_{2k} + \cdots + U_{(k-2)(k-1)} + U_{(k-2)k} + U_{(k-1)k}$$

$$= \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij}$$

$$\frac{K(K-1)}{2}$$



JT Test uses a Mann-Whitney statistic U_{ij} for the two-sample problem comparing samples i and j , where $i, j = 1, 2, .. k$ with $i < j$.

An overall test statistic is constructed by adding all the U_{ij}'s

Thus the test statistic is:

$$B = U_{12} + U_{13} + \cdots + U_{1k} + U_{23} + U_{24} + \cdots + U_{2k} + \cdots + U_{(k-2)(k-1)} + U_{(k-2)k} + U_{(k-1)k}$$

$$= \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} U_{ij}$$

Question: How to compute U_{ij}

Now, JT test uses Mann-Whitney statistic Uij, for the 2-sample problem comparing the samples i and j, where i and j belong to 1 to k with i < j. The overall test statistic therefore, is constructed by adding the Uij in the following way, this is equal to

$$U_{12} + U_{13} + \cdots + U_{1k} + U_{23} + U_{24} + \cdots + U_{2k} + \cdots + U_{(k-2)(k-1)} + U_{(k-2)k} + U_{(k-1)k}$$

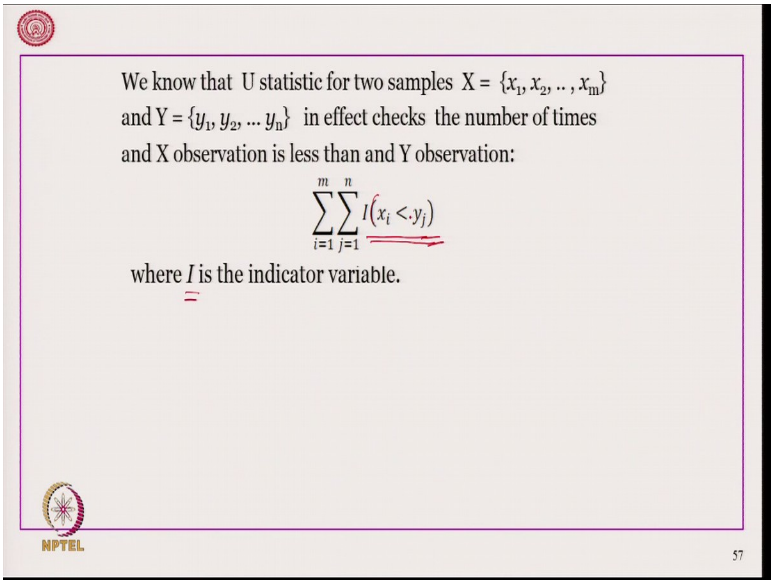
So, since we have k*(k -1) / 2 many different pairs, therefore we have to compare for all of them and we have to obtain the corresponding Mann Whitney statistic.

Therefore, the final statistic is this sum of this

$$\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} U_{ij}$$

As you can easily understand that this expression can be written neatly concisely like this. Question is how to compute Uij?

(Refer Slide Time: 53:50)



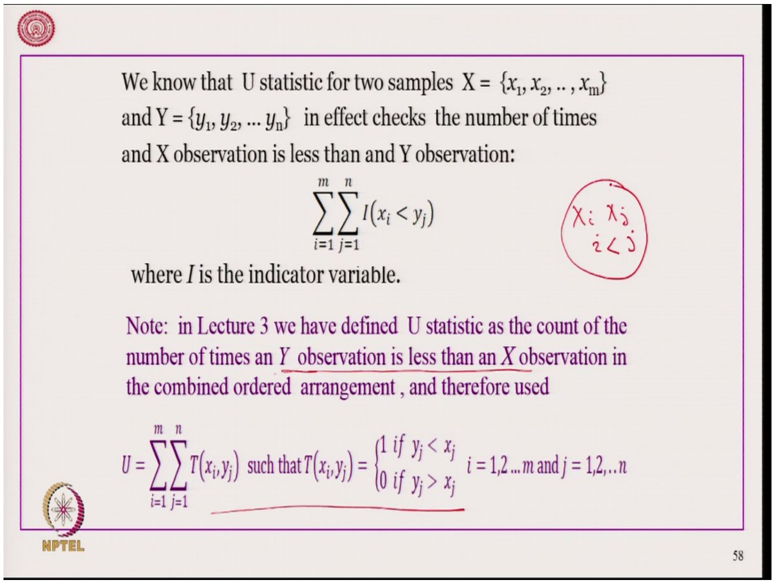
We know that U statistic for two samples X = $\{x_1, x_2, .., x_m\}$ and Y = $\{y_1, y_2, ... y_n\}$ in effect checks the number of times and X observation is less than and Y observation:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} I\underline{(x_i <. y_j)}$$

where $\underline{I}$ is the indicator variable.

We know that the U statistic for 2 samples X and Y in effect checks the number of times an X observation is less than Y observation that is i = 1 to m, j = 1 to n indicator of Xi is less than Yj. Therefore, you are looking at all pairs the Xi and Yj and we are checking how many times Xi is less than Yj, i is indicator variable that means that if this is true, then i = 1 otherwise i = 0.

(Refer Slide Time: 54:30)



We know that U statistic for two samples X = $\{x_1, x_2, .., x_m\}$ and Y = $\{y_1, y_2, ... y_n\}$ in effect checks the number of times and X observation is less than and Y observation:

$$\sum_{i=1}^{m}\sum_{j=1}^{n} I(x_i < y_j)$$

$X_i \; X_j$
$i < j$

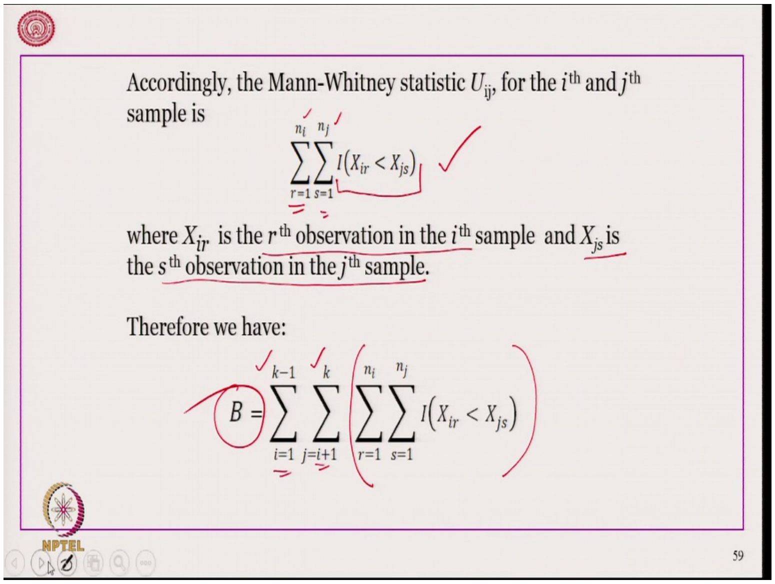
where I is the indicator variable.

Note: in Lecture 3 we have defined U statistic as the count of the number of times an Y observation is less than an X observation in the combined ordered arrangement, and therefore used

$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} T(x_i, y_j) \text{ such that } T(x_i, y_j) = \begin{cases} 1 \text{ if } y_j < x_j \\ 0 \text{ if } y_j > x_j \end{cases} \quad i = 1,2...m \text{ and } j = 1,2,..n$$

So, I am giving a slightly different notation. So that if you do not get confused, I am telling you that in lecture 3 we have defined U statistic as the count of number of times that Y observation is less than X. So, that is what we have done there and therefore we got this formula. If you want you can go and check with your lecture notes 3, but here the difference is that they are ordered that means they went on talking about Xi and Xj basically i < j. So this ordering we are maintaining to give it you convenient notation for JT test.

(Refer Slide Time: 55:13)



Accordingly, the Mann-Whitney statistic U_{ij}, for the i^{th} and j^{th} sample is

$$\sum_{r=1}^{n_i}\sum_{s=1}^{n_j} I(X_{ir} < X_{js})$$

where X_{ir} is the r^{th} observation in the i^{th} sample and X_{js} is the s^{th} observation in the j^{th} sample.

Therefore we have:

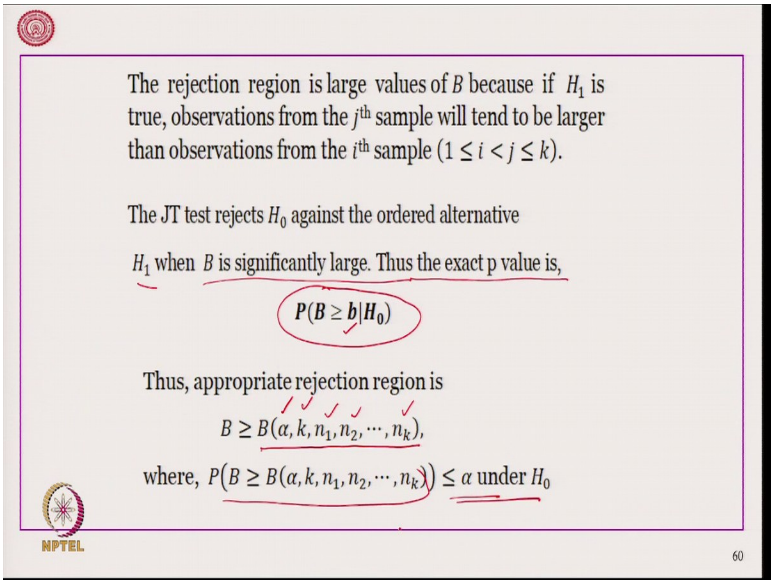
$$\boxed{B} = \sum_{i=1}^{k-1}\sum_{j=i+1}^{k}\left(\sum_{r=1}^{n_i}\sum_{s=1}^{n_j} I(X_{ir} < X_{js})\right)$$

Accordingly, the Mann Whitney statistic Uij for ith and jth sample is r is equal to 1 to ni s is equal to 1 to nj. Because these are the number of samples for the ith and jth population respectively. And we are looking at the corresponding value whether it is less than or not. Where Xir is rth observation in ith sample and Xjs is the sth observation in the jth sample. Therefore, the finally what we get

$$B = \sum_{i=1}^{k-1} \sum_{j=i+1}^{k} \sum_{r=1}^{n_i} \sum_{s=1}^{n_j} I(X_{ir} < X_{js})$$

this we have already seen and this part is replacing Uij from here and these complicated expression gives us the value of B the statistic.

(Refer Slide Time: 56:08)



The rejection region is large values of B because if H_1 is true, observations from the j^{th} sample will tend to be larger than observations from the i^{th} sample $(1 \le i < j \le k)$.

The JT test rejects H_0 against the ordered alternative H_1 when B is significantly large. Thus the exact p value is,

$$\boxed{P(B \ge b | H_0)}$$

Thus, appropriate rejection region is

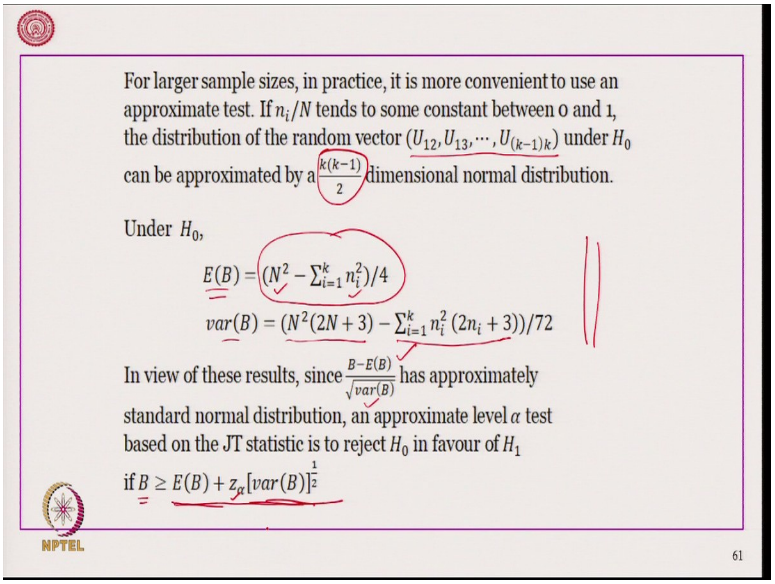
$$B \ge B(\alpha, k, n_1, n_2, \cdots, n_k),$$

where, $P(B \ge B(\alpha, k, n_1, n_2, \cdots, n_k)) \le \alpha$ under H_0

The rejection region is large values of B, because if H1 is true observations from the jth sample will tend to be larger than observations from the ith sample the JT test rejects Ho against our ordered alternative H1 when B is significantly large does that exact p value is coming out to be probability of B is greater than equal to some small b under H naught.

Therefore, the appropriate rejection region is going to be B is greater than equal to some value which comes as a function of alpha k and also in 1 into nk such that the probability that B is greater than equal to this value has to be less than or equal to alpha under Ho.

(Refer Slide Time: 57:05)



For larger sample sizes, in practice, it is more convenient to use an approximate test. If n_i/N tends to some constant between 0 and 1, the distribution of the random vector $(U_{12}, U_{13}, \cdots, U_{(k-1)k})$ under H_0 can be approximated by a $\frac{k(k-1)}{2}$ dimensional normal distribution.

Under H_0,

$$E(B) = (N^2 - \textstyle\sum_{i=1}^{k} n_i^2)/4$$

$$var(B) = (N^2(2N+3) - \textstyle\sum_{i=1}^{k} n_i^2\,(2n_i+3))/72$$

In view of these results, since $\frac{B-E(B)}{\sqrt{var(B)}}$ has approximately standard normal distribution, an approximate level α test based on the JT statistic is to reject H_0 in favour of H_1 if $B \geq E(B) + z_\alpha [var(B)]^{\frac{1}{2}}$

However, for larger samples, in practice, it is more convenient to use an approximate test if ni by capital N tends to be some constant between 0 to 1 the distribution of the random vector U1,2 U1,3 up to Uk-1,k under Ho can be approximated by a k*(k- 1)/2 dimensional normal distribution. So, it is a multivariate normal type of distribution and the expected value of B is coming out to be

$$(N^2 - \textstyle\sum_{i=1}^{k} n_i^2)/4$$

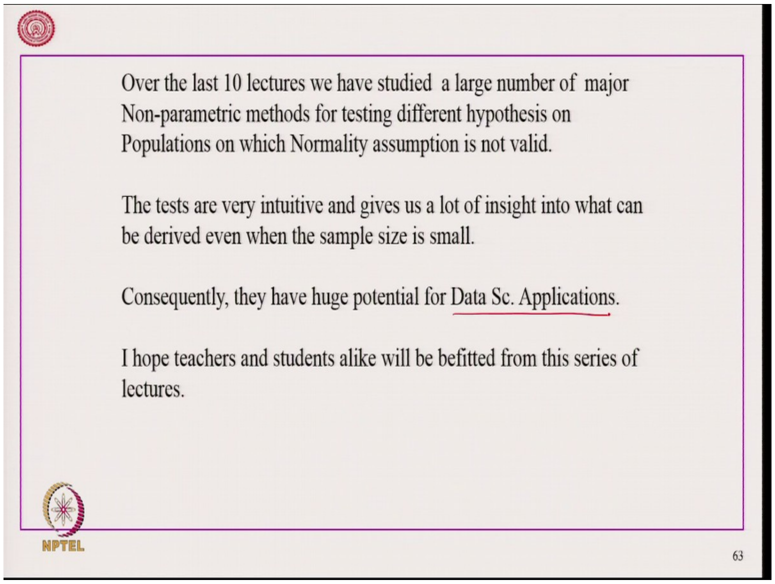
and variance of these coming out to be

$$(N^2(2N+3) - \textstyle\sum_{i=1}^{k} n_i^2\,(2n_i+3))/72$$

We are not going to derive these here that is no scope. However, we know that when value is large, then B minus expected value of B over square root of variance of B can be considered to be a normal approximation, where expected value of B and variance of B can be found out without going into the detail simply from the values of capital N and small a ni is equal to 1 to k, right. Once we are going to use a normal test, we shall check if B is greater than equal to the expected value of B plus Zα which comes from the normal distribution multiplied by the standard deviation that is variance of B to the power half that is the square root of the variance.

So, that is how we actually conduct the test to check if we can accept or reject the null hypothesis.

(Refer Slide Time: 59:01)



Over the last 10 lectures we have studied a large number of major Non-parametric methods for testing different hypothesis on Populations on which Normality assumption is not valid.

The tests are very intuitive and gives us a lot of insight into what can be derived even when the sample size is small.

Consequently, they have huge potential for Data Sc. Applications.

I hope teachers and students alike will be befitted from this series of lectures.

Okay friends, with that, I come to an end. Over the last 10 lectures, we have studied a large number of major nonparametric methods for testing different hypotheses on populations on which normality assumptions is not valid. The tests are very intuitive, and gives us a lot of insight into what can be derived when the sample size is small.

Consequently, they have huge potential for data science applications. And I hope that all those who attended whether a teacher or a student they will be benefited from this series of lectures and this is my sincere respect to all those great beings who have thought so nicely and who could get so much out of small sample size that is truly amazing and truly inspiring for all of us. So, this series of lectures is from my side as a homage to all these great minds.

Finally, this series could not have been possible without you students really attending and understanding the rather complicated subject. I could make this possible because of a lot of help, particularly from NPTEL people of IIT Delhi, NPTEL of IIT Madras, And last, but not the least, my teaching assistant, my student Kushagri Tandon. who understood the concept in a very short period and helped me throughout in preparation of the lectures and framing the questions for all of you. I hope you solve the problems and get the beauty of nonparametric statistics. Thank you.