Marketing Research and Analysis-II (Application Oriented)

Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

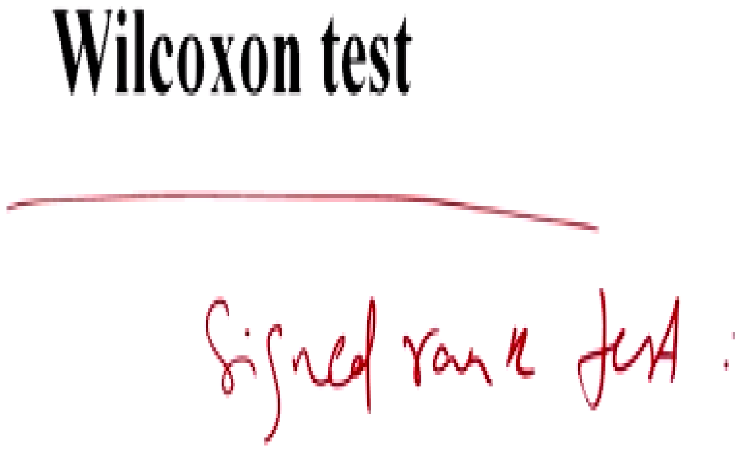
Lecture – 29
Non-Parametric Test - V

Welcome everyone to the class of Marketing Research and Analysis. In the earlier classes, we were continuing with the nonparametric test and we discussed that nonparametric test is one test which is used when the data is distribution free or the data does not follow normal distribution. Suppose the nonparametric test that we had done in the last class was like the Mann-Whitney U test which we compared with the independent sample t test with the parametric one and the Kruskal-Wallis test similarly like the one way analysis of variance.

Then we did the runs test which was used to check the randomness of factor, the occurrence of a variable or product. So we wanted to measure whether it is a systematic you know occurrence or a random occurrence and we said if it is systematic in occurrence, then we would have to take care of the problem. Then we also did the sign test in which we tried to check the hypothesis some of the hypothesis of a nonparametric measure, but we said that sign test is not a powerful test and it has own limitations and in comparison there is a more powerful test which is the Wilcoxon signed-rank test.

So today we will be dealing with it. One more test that we had done was the association test which is very close to the Pearson correlation in which a nonparametric ones you say the Spearman correlation. So the Spearman correlation was a nonparametric correlation method for the nonparametric values which we have adopted and found out how you should find out the association within 2 variables when the data is in a non-normal nature.

(Refer Slide Time: 02:32)

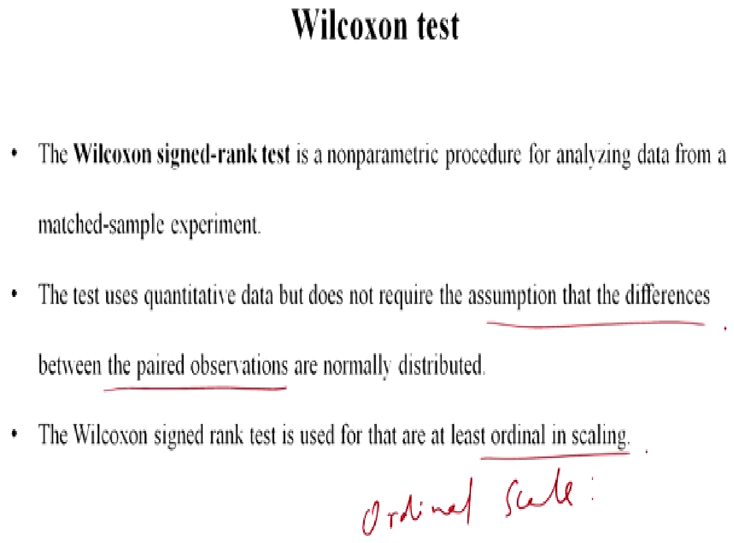


Wilcoxon test

Signed rank test :

So today we will start a lecture with the Wilcoxon test which is called the signed-rank test.

(Refer Slide Time: 02:37)



Wilcoxon test

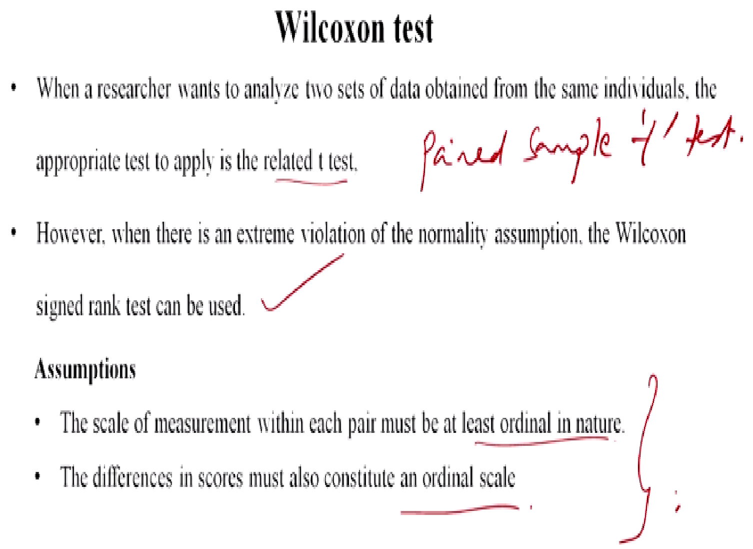
- The **Wilcoxon signed-rank test** is a nonparametric procedure for analyzing data from a matched-sample experiment.
- The test uses quantitative data but does not require the assumption that the differences between the paired observations are normally distributed.
- The Wilcoxon signed rank test is used for that are at least ordinal in scaling.

Ordinal Scale :

So Wilcoxon signed rank test is a very popular test. So we had just started in the last class. What it says is a nonparametric procedure for analyzing data from a matched sample experiment. So basically when there are 2 pairs of sample, you are using this test. So it uses quantitative data, but does not require the assumption, this is important that the differences between the paired observations are normally distributed, so it is not normally distributed that is why nonparametric.

The Wilcoxon signed-rank test is used for at least for that or at least ordinal in scaling so the data has to be measured in an ordinal scale. So then only it is eligible to use the Wilcoxon test.

(Refer Slide Time: 03:19)



Wilcoxon test

- When a researcher wants to analyze two sets of data obtained from the same individuals, the appropriate test to apply is the related t test. Paired Sample 't' test.
- However, when there is an extreme violation of the normality assumption, the Wilcoxon signed rank test can be used. ✓

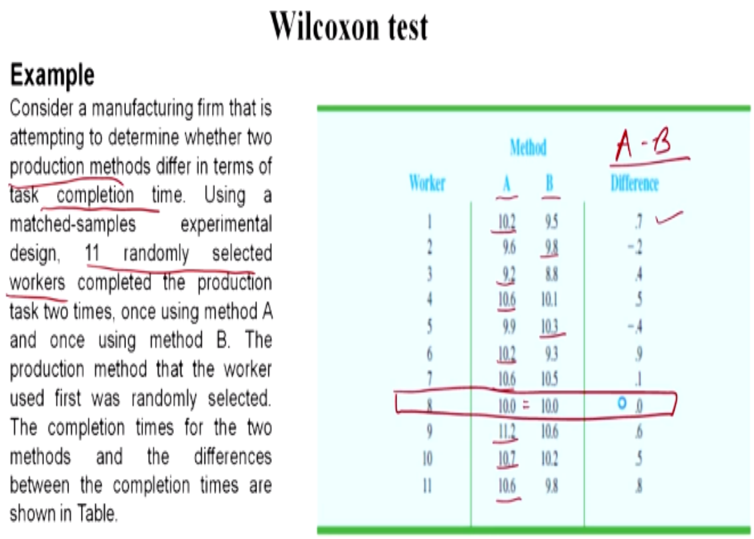
Assumptions

- The scale of measurement within each pair must be at least ordinal in nature.
- The differences in scores must also constitute an ordinal scale

So when a researcher, what is the condition. When a researcher wants to analyze 2 sets of data obtained from the same individual, the appropriate test is applied is a related t test, so which you says a paired sampled t test which we use in for a parametry. So here similarly, we use the Wilcoxon. However, when there is an extreme violation of the normality assumption, the Wilcoxon test can be used. So the assumptions of this test are the scale of measurement within each pair must be at least ordinal in nature.

The differences in scores must also constitute an ordinal scale. So these are the 2 assumptions which we have to follow before we start go for a Wilcoxon signed-rank test.

(Refer Slide Time: 04:05)



Wilcoxon test

Example

Consider a manufacturing firm that is attempting to determine whether two production methods differ in terms of task completion time. Using a matched-samples experimental design, 11 randomly selected workers completed the production task two times, once using method A and once using method B. The production method that the worker used first was randomly selected. The completion times for the two methods and the differences between the completion times are shown in Table.

	Method		A - B
Worker	A	B	Difference
1	10.2	9.5	.7 ✓
2	9.6	9.8	−.2
3	9.2	8.8	.4
4	10.6	10.1	.5
5	9.9	10.3	−.4
6	10.2	9.3	.9
7	10.6	10.5	.1
8	10.0 =	10.0	0 .0
9	11.2	10.6	.6
10	10.7	10.2	.5
11	10.6	9.8	.8

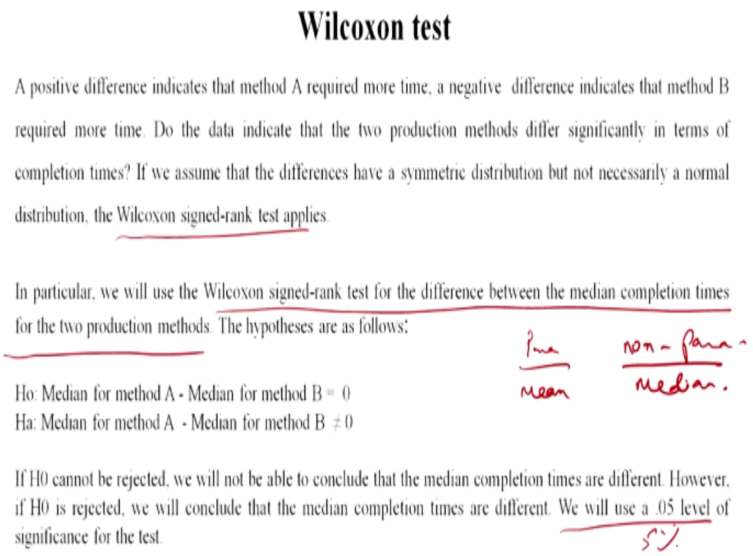
So this is the example we had just started. So this is a manufacturing firm which has got around 11 workers in this case. They are using 2 methods, they are method A and method B of producing. So what it says, so it says that consider a manufacturing firm that attempted to determine whether 2 production methods, method A and method B in differ in terms of the task completion time, so is there a difference in the task completion when somebody adopts method A versus the method B is to be found out.

So what they have done is, they have an 11 randomly selected workers were chosen and they were asked to do the work 2 times, once using method A and the other using method B. The production method the worker used first was randomly selected, so the worker was free to select any of the method of his choice. The completion time for the 2 methods and the differences between the completion times are shown below in this table. So you see for example if you take, it is unit less at the moment, so 10.2 seconds or 10.2 minutes you can take anyway.

So let us forget the units. So let us say the first worker took 10.2 let us say seconds by using methods A and the same worker has used 9.5 seconds, so that means he has taken less time by using the method B. Similarly worker 2 has used 9.6 seconds here and 9.8 seconds by using method B, so worker 2 is taking more time by using method B whereas worker 1 has taken more time with method A. Similarly the third worker has taken more time with A, fourth worker has taken again more time with A.

Fifth worker has taken more time with B, sixth again A, seventh A, eighth there is no difference, this is both are same, ninth if you see it is again A, tenth A, 11 A. So the difference in the timing is recorded here. So 10.2 – 9.5 is 0.7. So if it is positive that means we understand since it is A-B, so we understand the method A takes more time. So if it is minus, then we understand that the method B takes more time. So this is a simple thing that we have done and the difference has been found out to be like this.

(Refer Slide Time: 06:53)



Wilcoxon test

A positive difference indicates that method A required more time, a negative difference indicates that method B required more time. Do the data indicate that the two production methods differ significantly in terms of completion times? If we assume that the differences have a symmetric distribution but not necessarily a normal distribution, the Wilcoxon signed-rank test applies.

In particular, we will use the Wilcoxon signed-rank test for the difference between the median completion times for the two production methods. The hypotheses are as follows:

Para
Mean

non-Para-
Median.

Ho: Median for method A - Median for method B = 0
Ha: Median for method A - Median for method B ≠ 0

If H0 cannot be rejected, we will not be able to conclude that the median completion times are different. However, if H0 is rejected, we will conclude that the median completion times are different. We will use a .05 level of significance for the test.

5%.

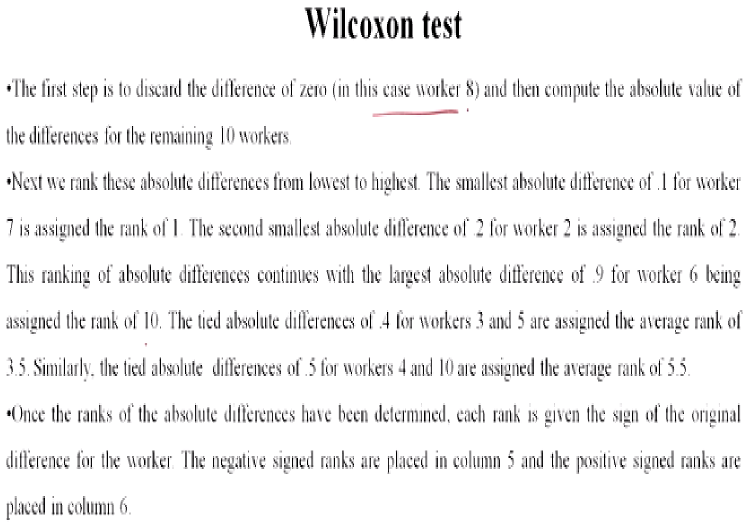
A positive difference indicates that method A required more time as I just said, a negative difference indicates method B required more time. Do the data indicate that the 2 production methods differ significantly in terms of their completion times, so this is of significant interest to the researcher and in fact there are several situations in life where you might have to compare 2 methods of production or 2 methods of teaching or you might have several experiments can be done with such kind of options.

If we assume that the differences have a symmetric distribution but not necessarily a normal distribution, the Wilcoxon rank test applies. So it is not a normal distribution we have said. In particular, the Wilcoxon signed-rank test for the difference between the median completion times for the 2 production methods, so what it does is, it does not as I have said earlier also in my earlier class in case of a parametric we use the mean, in case of a nonparametric we use mostly the median.

So here also when we make our hypothesis if you see, the null hypothesis what it says that the median time for completing the work using method A is equal to the median time of completing the work by using method B. So the median for method A minus median for method B is equal to 0, but that is not of the interest right. So any worker or any supervisor or any manager or any researcher, he is not interested, he wanted to see that there should be a difference, so what is the difference.

Alternate hypothesis states the median for method A is not equal to the median for method B okay. So if H0 the null hypothesis cannot be rejected, we will not be able to conclude that the median completion times are different. However, if alternate is accepted or H0 is rejected, we will conclude that the median completion times are difference, obviously. What we are doing, we are using a 0.5, 5% level of significance.

(Refer Slide Time: 09:13)



Wilcoxon test

•The first step is to discard the difference of zero (in this case worker 8) and then compute the absolute value of the differences for the remaining 10 workers.

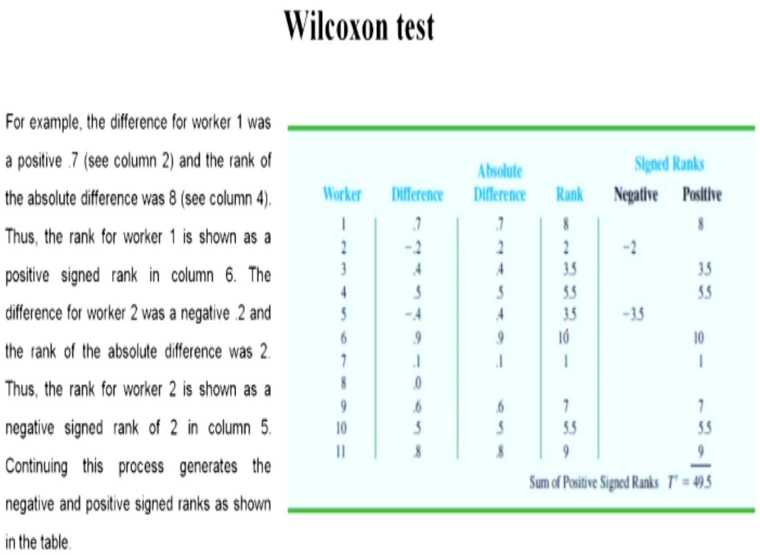
•Next we rank these absolute differences from lowest to highest. The smallest absolute difference of .1 for worker 7 is assigned the rank of 1. The second smallest absolute difference of .2 for worker 2 is assigned the rank of 2. This ranking of absolute differences continues with the largest absolute difference of .9 for worker 6 being assigned the rank of 10. The tied absolute differences of .4 for workers 3 and 5 are assigned the average rank of 3.5. Similarly, the tied absolute differences of .5 for workers 4 and 10 are assigned the average rank of 5.5.

•Once the ranks of the absolute differences have been determined, each rank is given the sign of the original difference for the worker. The negative signed ranks are placed in column 5 and the positive signed ranks are placed in column 6.

The first step is to if you go back to the table, there was one time for the worker 8, can you see this and here the worker has taken similar time for both by using method A and method B. So when such a case occurs in this kind of a test in the Wilcoxon test, it is necessary to discard that particular element or that particular case. So the first step is to discard the difference of 0 in this case the worker 8 and then compute the absolute value of the differences for the remaining 10 workers okay.

Now what do we do next. The next is we rank these absolute differences from lowest to highest, so the one which is the lowest difference in time is the first rank, the highest difference in time is the last rank, the tenth rank in this case. So smallest the absolute difference of 0.1 for worker 7 is assigned the rank of 1. Similarly the second smallest absolute difference of 0.2 is assigned the rank of 2. This ranking of absolute differences continues with the largest absolute difference of 0.9 for worker 6, being assigned the rank of 10.

(Refer Slide Time: 10:24)



Wilcoxon test

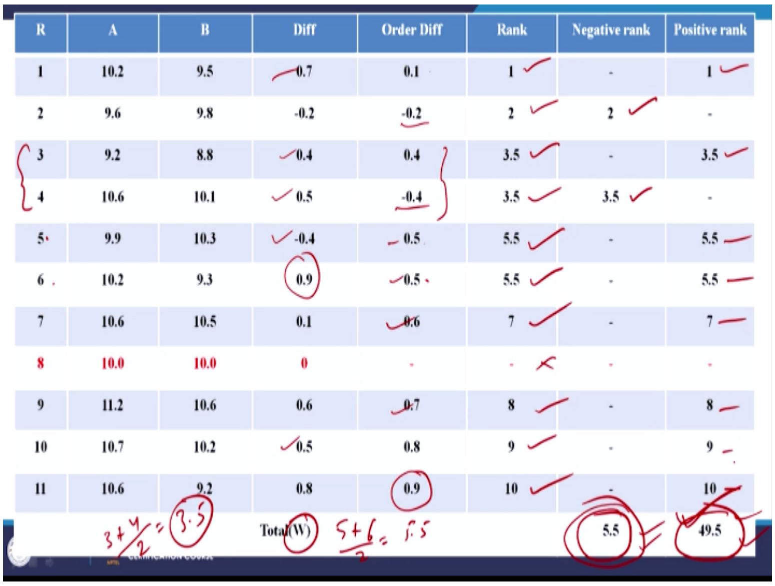
For example, the difference for worker 1 was a positive .7 (see column 2) and the rank of the absolute difference was 8 (see column 4). Thus, the rank for worker 1 is shown as a positive signed rank in column 6. The difference for worker 2 was a negative .2 and the rank of the absolute difference was 2. Thus, the rank for worker 2 is shown as a negative signed rank of 2 in column 5. Continuing this process generates the negative and positive signed ranks as shown in the table.

				Signed Ranks	
Worker	Difference	Absolute Difference	Rank	Negative	Positive
1	.7	.7	8		8
2	−.2	.2	2	−2	
3	.4	.4	3.5		3.5
4	.5	.5	5.5		5.5
5	−.4	.4	3.5	−3.5	
6	.9	.9	10		10
7	.1	.1	1		1
8	.0				
9	.6	.6	7		7
10	.5	.5	5.5		5.5
11	.8	.8	9		9

Sum of Positive Signed Ranks $T^{+} = 49.5$

So let us see this table. So for example the difference is here, so here the absolute difference we are making it irrespective of any sign. So it is all positive, so 0.7, 0.2, 0.4, 0.5, 0.4, 0.9, 0.1, 0.6, 0.5, 0.8. Now if you have to rank, let us say we have to rank. I think I have done in the next slide.

(Refer Slide Time: 10:46)



R	A	B	Diff	Order Diff	Rank	Negative rank	Positive rank
1	10.2	9.5	0.7	0.1	1 ✓	-	1 ✓
2	9.6	9.8	-0.2	-0.2	2 ✓	2 ✓	-
3	9.2	8.8	✓0.4	0.4	3.5 ✓	-	3.5 ✓
4	10.6	10.1	✓0.5	-0.4	3.5 ✓	3.5 ✓	-
5	9.9	10.3	✓-0.4	0.5	5.5 ✓	-	5.5
6	10.2	9.3	0.9	0.5	5.5 ✓	-	5.5
7	10.6	10.5	0.1	0.6	7 ✓	-	7
8	10.0	10.0	0	-	- ✗	-	-
9	11.2	10.6	0.6	0.7	8 ✓	-	8
10	10.7	10.2	✓0.5	0.8	9 ✓	-	9
11	10.6	9.2	0.8	0.9	10 ✓	-	10
	$\frac{3+4}{2} =$	3.5	Total(W)	$\frac{5+6}{2} = 5.5$		5.5 ✓	49.5 ✓

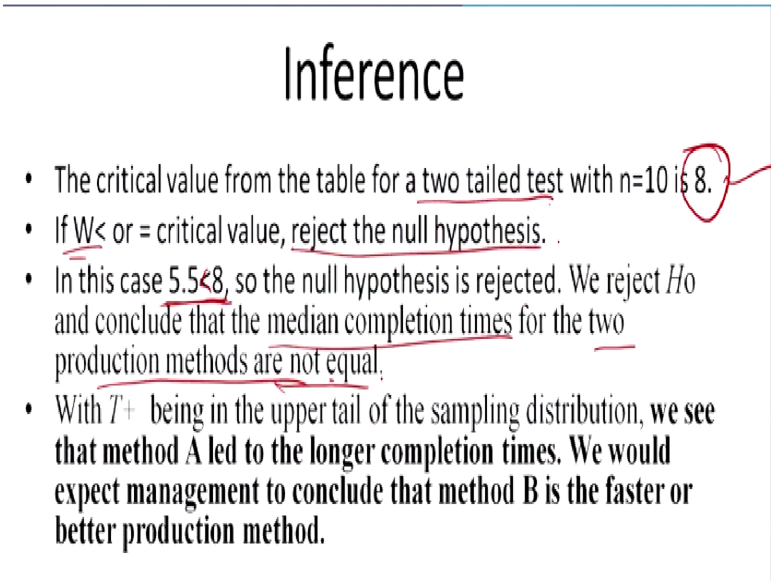
Yeah I have done better here. So if you look at the order difference first of all which is the lowest out of the lot, now 0.1, we will only consider the absolute differences, so 0.1. Then what is the next 0.2, so we are keeping the with a negative, no issues. Then 0.4, then again there is a 0.4, this is one, this is one right, so again a 0.4, then 0.5, then 0.5 again, then 0.6, 0.6 next, then 0.7, this is the one. So this is discarded, so it is not to be counted. Then 0.8, so this is coming here right and finally 0.9 the difference with the worker 6 is coming here.

So this is the difference arranged in an order. Now let us give a rank, now how do we give a rank? So the first thing is the first one gets a first rank, the minimum difference. The second one gets a second rank, but now you see there are 2 people at 0.4, taking the absolute difference as 0.4, they are in the position third and fourth, so here the ranking will become now 3+4/2, I will write it here 3+4/2 = 3.5, so both get a 3.5, 3.5. The next is 5.5, the next is 5.5, let us see so what is happening here.

Now if you look at this value what is coming up, which number is this, so 0.5, 0.5, so 5 and 6, so 5+6/2 so that is equal to again 5.5, and there are for 2 cases, so 5.5, 5.5. So the next is once you have done with it, then the 0.6 which gets a seventh rank. So 1, 2, 3, 3.5, 3.5, this is 3 and 4, 5 and 6, then 7. Then this is discarded, then eighth rank, then the ninth rank, and the finally the tenth rank. So all these ranks have been given now. Now the question is, now let us see which of them are fallen to the negative category.

So this one has a negative value, this one has a negative value, these two. So the second rank and the fourth rank, they have a negative value. Now if you look at the positive rank similarly it is 1, 3, 5, 6, 7, 8, 9, 10. So if I add the ranks, for this if I add which I am saying is W, the total weight is 5.5 for the negative ranks and 49.5 for the positive ranks. Now what, now what will you do, how will you proceed from here, let us see. So what is the inference done.

(Refer Slide Time: 13:48)



Inference

- The critical value from the table for a two tailed test with n=10 is 8.
- If W< or = critical value, reject the null hypothesis. .
- In this case 5.5<8, so the null hypothesis is rejected. We reject *Ho* and conclude that the median completion times for the two production methods are not equal.
- With *T+* being in the upper tail of the sampling distribution, **we see that method A led to the longer completion times. We would expect management to conclude that method B is the faster or better production method.**

The critical value from the table for a two-tailed test with n = 10. Now what is this n, so the n is here 10 because one person has been discarded, so remaining is 10. So I will show you, I have brought the table also just I can show you or you can just follow the Wilcoxon table where it says with a particular number of sample size, what is the value. So the critical value for a two-tailed test, this is obviously a two-tailed test, because it is not equal to case, so it can fall anywhere to the right or left we do not know, is 8 actually.

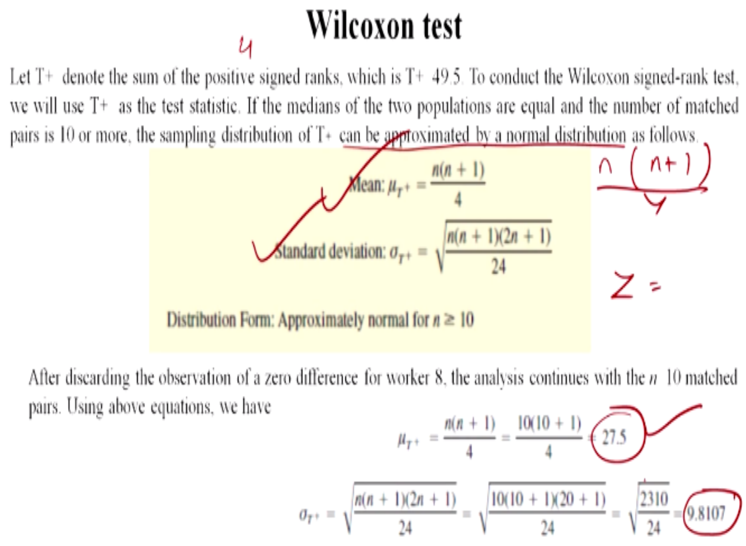
So if W that means the weight, this weight 1, 2, this weight is less than or equal to the critical value reject the null hypothesis. Now the question is here we are saying W, but here we having two W's, in fact one for the negative rank one for the positive rank, so which one would you select or keep in during your study. So always keep the one which is the lowest out of the 2, so in this case which is lower, the 5.5 and 49.5, so obviously the 5.5 is the lower rank.

So we will compare this 5.5 with the value of 8. In this case, 5.5 which is the value is lesser than the critical value of 8 for a sample size of 10. So the null hypothesis is rejected. We reject H0 and conclude that the median completion times for the 2 production methods are not equal. Let us see that is what we have written, let us go back. So median for this is rejected, so this is accepted. So median for method A and median for method B is not equal, that is what we were saying, are not equal.

So now which then takes more time, so let us go to this. With T+ the time being in the upper tail of the sampling distribution, we see that, now let us go back, now here it is 49.5, so the rank, now this is obviously if you look at the time, it takes more time. So the T+ being the upper tail, we see that the method A led to the longer completion times, just go back, so if you if you look at the difference in which side if you look at the difference positive, then we say is A is bigger, if it is negative B is bigger.

So in in this case, we see this is more of positives, so that means what. If you look at this differences and from here we see the positive rank is more, so that is why we say obviously the A has taken more time than the B. So the method A, when the workers are using method A, they are consuming more time than the method B. So we would expect management to conclude that method B is the faster or better production method and A takes more time than B, this is method.

(Refer Slide Time: 16:58)



Wilcoxon test

4

Let T+ denote the sum of the positive signed ranks, which is T+ 49.5. To conduct the Wilcoxon signed-rank test, we will use T+ as the test statistic. If the medians of the two populations are equal and the number of matched pairs is 10 or more, the sampling distribution of T+ can be approximated by a normal distribution as follows.

Mean: $\mu_{T^+} = \frac{n(n+1)}{4}$

$\frac{n\,(n+1)}{4}$

Standard deviation: $\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}}$

$Z =$

Distribution Form: Approximately normal for $n \geq 10$

After discarding the observation of a zero difference for worker 8, the analysis continues with the n 10 matched pairs. Using above equations, we have

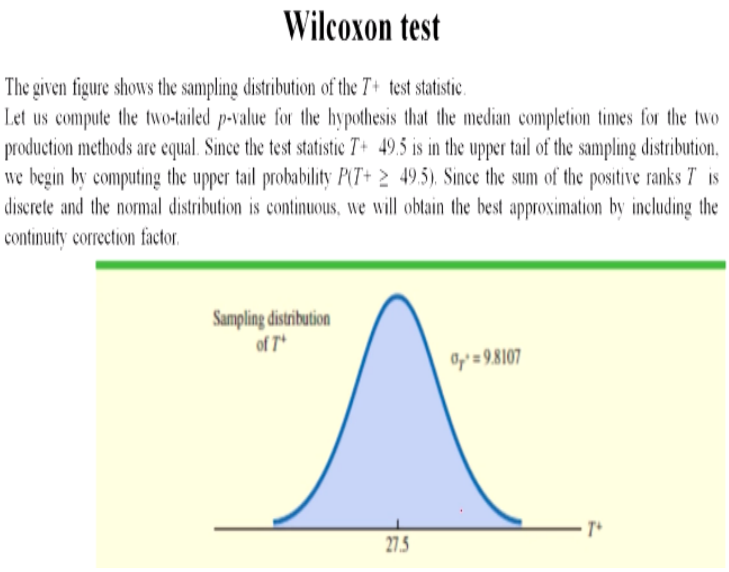
$$\mu_{T^+} = \frac{n(n+1)}{4} = \frac{10(10+1)}{4} = \boxed{27.5}$$

$$\sigma_{T^+} = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{10(10+1)(20+1)}{24}} = \sqrt{\frac{2310}{24}} = \boxed{9.8107}$$

There is another way of also doing the same problem which I have said as method 2 and I will show you how to do that the Wilcoxon signed-rank test. Now let us see T+ denote that sum of the positive signed ranks which is 49.5. To conduct the test, we use the test statistics. If the median of the 2 populations are equal and the number of the matched pairs is 10 or more, the sampling distribution can be approximated by a normal distribution. So in our case it is 10, so how do you calculate the mean in such a condition.

The formula is n x n+1/4, this is the mean you calculate, this is the standard deviation. So because we required to calculate the z, so to calculate the z now you require the mean and the standard error. So let us see what is the mean, the mean has been 27.5. So n is 10. So 10 x 10+1/4 gives to 27.5. What is the standard error. The standard error is n x n+1 x 2n+1/24 which comes to 9.8107 okay.

(Refer Slide Time: 18:08)

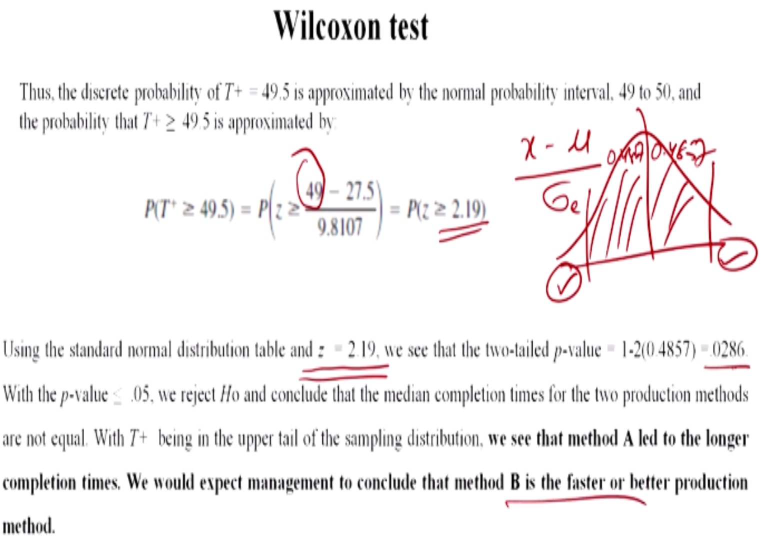


Wilcoxon test

The given figure shows the sampling distribution of the $T+$ test statistic.

Let us compute the two-tailed p-value for the hypothesis that the median completion times for the two production methods are equal. Since the test statistic $T+$ 49.5 is in the upper tail of the sampling distribution, we begin by computing the upper tail probability $P(T+ \geq 49.5)$. Since the sum of the positive ranks T is discrete and the normal distribution is continuous, we will obtain the best approximation by including the continuity correction factor.

(Refer Slide Time: 18:12)



Wilcoxon test

Thus, the discrete probability of $T+ = 49.5$ is approximated by the normal probability interval, 49 to 50, and the probability that $T+ \geq 49.5$ is approximated by:

$$P(T^+ \geq 49.5) = P\left(z \geq \frac{49 - 27.5}{9.8107}\right) = P(z \geq 2.19)$$

$$\frac{x - \mu}{\sigma_e}$$

Using the standard normal distribution table and $z = 2.19$, we see that the two-tailed p-value = 1-2(0.4857) = .0286.

With the p-value $\leq$ .05, we reject Ho and conclude that the median completion times for the two production methods are not equal. With $T+$ being in the upper tail of the sampling distribution, **we see that method A led to the longer completion times. We would expect management to conclude that method B is the faster or better production method.**

So we have the 2 values right. Now taking these 2 values, we can calculate you see the what you say the discrete probability of T is equal to 49.5 is approximated by the normal probability interval 49 to 50, that means it lies between 49 to 50, and it is approximated by P is equal to so that means you are calculating the z here, z is greater than equal to 49, approximation value has been taken the lower one, 49. So 49-27.5, so this is as good as if you remember x –mu/standard error.

So this is coming to be z is greater than equal to somewhere around 2.19. So if z is equal to 2.19, what does it mean. So let us see, so it is a two-tailed test, so we see that the two tailed p-value is equal to how much now the area under the curve when z = 2.19 is 0.4857, so this area is 0.4857 for this side, 0.4857 for this side. So what is remaining. So 1-2 x obviously we had multiplied two side, so 1- this much comes, 0.286. So this plus this value, this value and this value is coming to be point 0.0286 and we have taken a significance value of 5%.

So now this p-value and comparing with the significance value we have taken, now if you compare we see that this calculated value is lesser than the significance value. In this case since it is lesser, so we know from beginning when it is lesser, we will reject the null hypothesis. So we reject H0 and conclude that the median completion times for the 2 production methods are not equal, the same thing we had got earlier also. So we see the method A led to the longer completion times and you would expect management to conclude that method B is the faster or better production method.

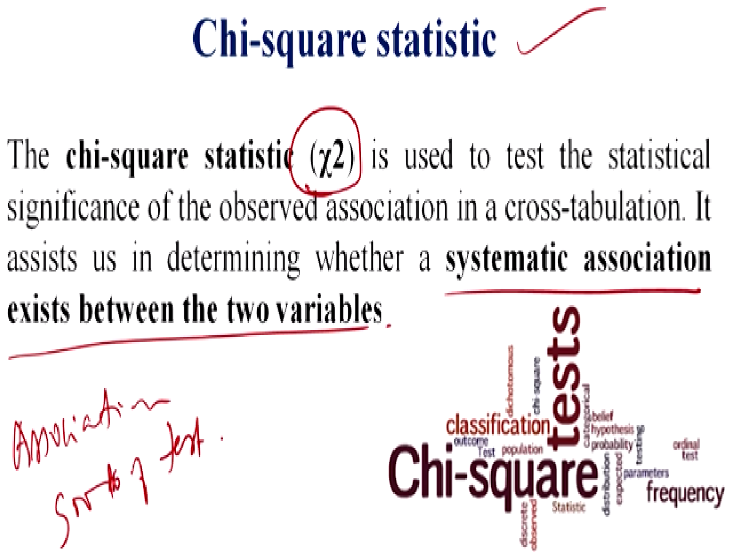
So I hope that Wilcoxon signed-rank test is very clear to you because both the methods you can utilize and understand and make an inference. So this is the one thing that we have done and it is a very powerful test much better than the sign test, which is not so powerful, in comparison Wilcoxon test is a much powerful test. I will show you how to run this test also on a software, may be little later on.

(Refer Slide Time: 20:50)

Chi-square Test

So the next is what we start with is another test called the chi-square test. Now what is this chi-square test. This chi-square test is another nonparametric test which has a very very large utility and it is utilized in order to compare data which are collected in a nominal manner, in nominal way. So let me just see if the data is available, no, we have not, I will show you later on.

(Refer Slide Time: 21:21)



Chi-square statistic

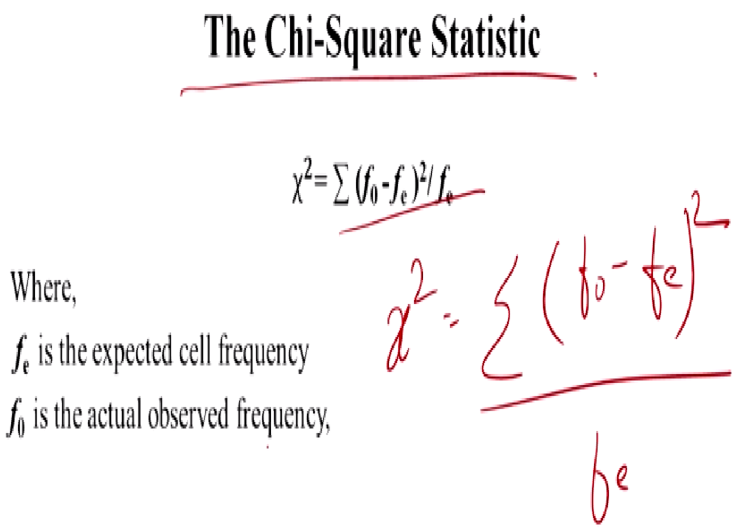
The **chi-square statistic** ($\chi 2$) is used to test the statistical significance of the observed association in a cross-tabulation. It assists us in determining whether a **systematic association exists between the two variables**.

Association
Signf test.

So what is this chi-square statistic or chi-square test. The chi square test is denoted as chi-square is used to test the statistical significance of the observed association in a cross-tabulation. Now in a cross-tabulation, we have seen many a times the use of cross-tabulation is very high, but then it does not tell us whether the association is significant or not. So in a chi-square statistic or test, we use it to check whether the data we have got or the results we have got can be termed as significant or not significant.

It assists us in determining whether a systematic association exists between the 2 variables. Chi-square statistics is a very powerful technique because it tells about 2 important things, one is the association and the other is the goodness of it, which I will explain both the things in this lecture, goodness-of-fit test okay.

(Refer Slide Time: 22:23)



The Chi-Square Statistic

$$\chi^2 = \sum (f_0 - f_e)^2 / f_e$$

Where,

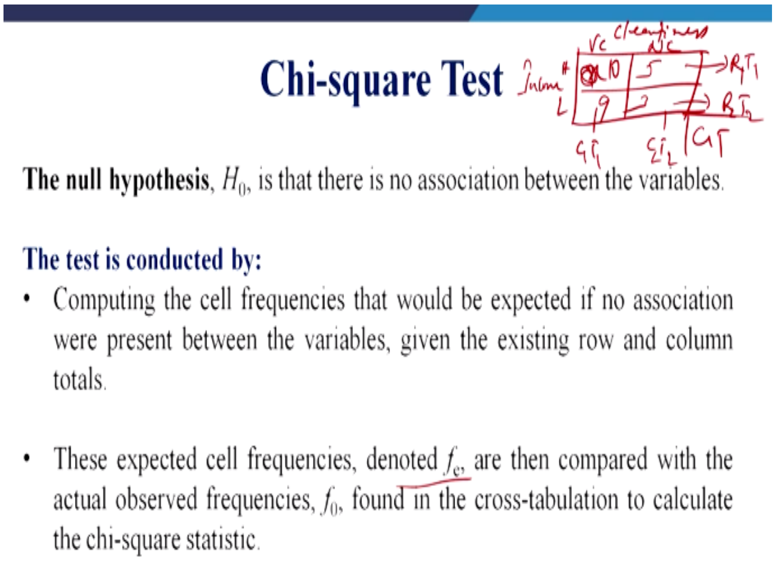
f_e is the expected cell frequency

f_0 is the actual observed frequency,

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

So what is this chi-square statistic. The chi-square statistic is given by the formula. So chi-square = summation f observe that means observed frequency minus expected frequency square divided by the expected frequency, this is the formula. So fe is the expected cell frequency, fo is the actual observed frequency.

(Refer Slide Time: 22:48)



Chi-square Test

Income	Cleanliness VC	NC	
H	10	5	→ RT_1
L	19	2	→ RT_2
	CT_1	CT_2	GT

The null hypothesis, H_0, is that there is no association between the variables.

The test is conducted by:

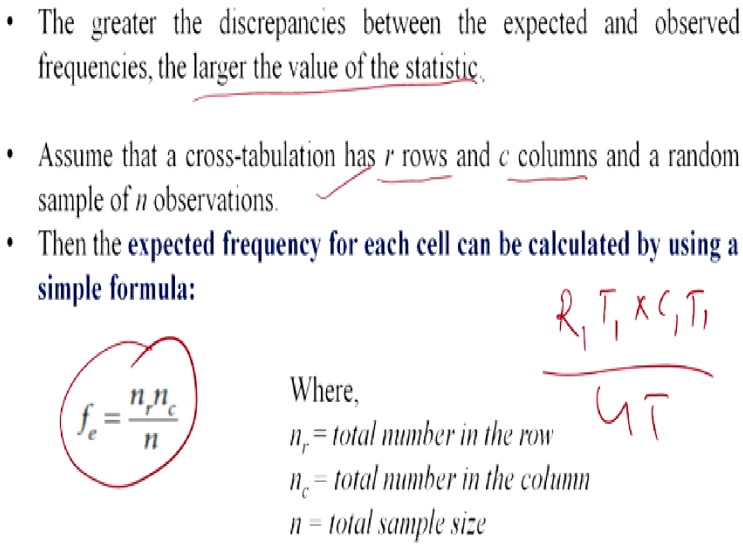
- Computing the cell frequencies that would be expected if no association were present between the variables, given the existing row and column totals.
- These expected cell frequencies, denoted f_e, are then compared with the actual observed frequencies, f_0, found in the cross-tabulation to calculate the chi-square statistic.

The null hypothesis is that there is no association between the variables. So there is no association between the variables, this is the null hypothesis, and what is the alternative, obviously the alternate will be that there is association between the variables, but we do not know what is the association like but there is an association. This test is conducted by computing the cell frequencies that would be expected if no associations are present between the variables.

Given the existing row and column totals, so suppose we say something looks like the income of people versus the cleanliness okay. So high income, low income, let us say people can be very clean, hygiene oriented or not so clean. So now when these are the row total, this is the column total, and you have a grand total out here. So what it says is this expected cell frequencies denoted as fe are then compared with actual observed frequencies. So you have an observed frequency.

Now how many people are actually from the high income group and who are very clean let us say x or we let us give number 10, high income group not so clean let us say is 5, low income group very clean is let us say 9, low income group not so clean is 3 let us say. Now from here is any association between cleanliness and income that was what we need to find out. So it helps to calculate the chi-square statistic.

(Refer Slide Time: 24:37)



- The greater the discrepancies between the expected and observed frequencies, the larger the value of the statistic.
- Assume that a cross-tabulation has r rows and c columns and a random sample of n observations.
- Then the **expected frequency for each cell can be calculated by using a simple formula:**

$$\boxed{f_e = \frac{n_r n_c}{n}}$$

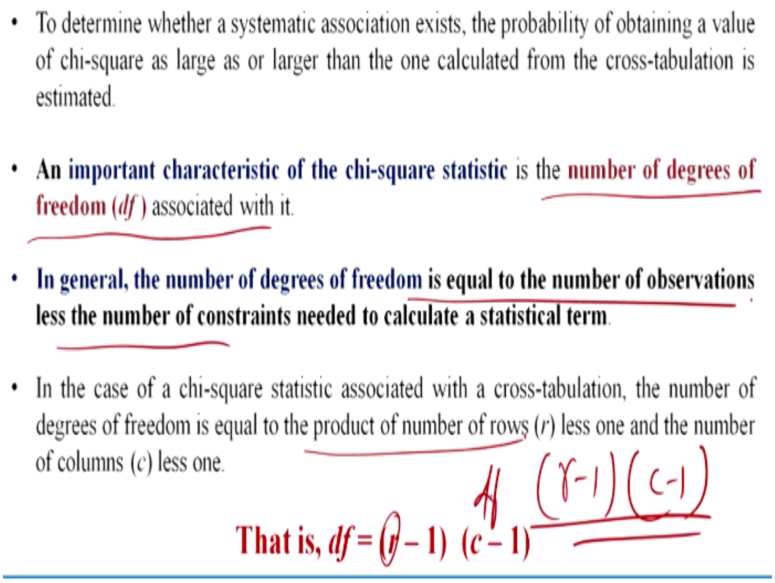
$$\frac{R_1T_1 \times C_1T_1}{GT}$$

Where,
n_r = *total number in the row*
n_c = *total number in the column*
n = *total sample size*

The greater the discrepancies between the expected and the observed frequency, so higher the difference between the expected frequency and the observed frequency, the larger is the value of the statistic and it is not a very desirable thing. Assume that a cross-tabulation has r rows and c columns and a random sample of n observations, then the expected frequency for each cell is calculated as like this.

So how is it calculated. Now let us say this R1T1, R2T2, C1T1, V2T2 and this grand total. Now what you do is you write it like this R1T1 x C1T1/grand total, so where nr is the total number in the row, total number in the column nc, and total sample size is the n.

(Refer Slide Time: 25:33)



- To determine whether a systematic association exists, the probability of obtaining a value of chi-square as large as or larger than the one calculated from the cross-tabulation is estimated.
- An important characteristic of the chi-square statistic is the number of degrees of freedom (df) associated with it.
- In general, the number of degrees of freedom is equal to the number of observations less the number of constraints needed to calculate a statistical term.
- In the case of a chi-square statistic associated with a cross-tabulation, the number of degrees of freedom is equal to the product of number of rows (r) less one and the number of columns (c) less one.

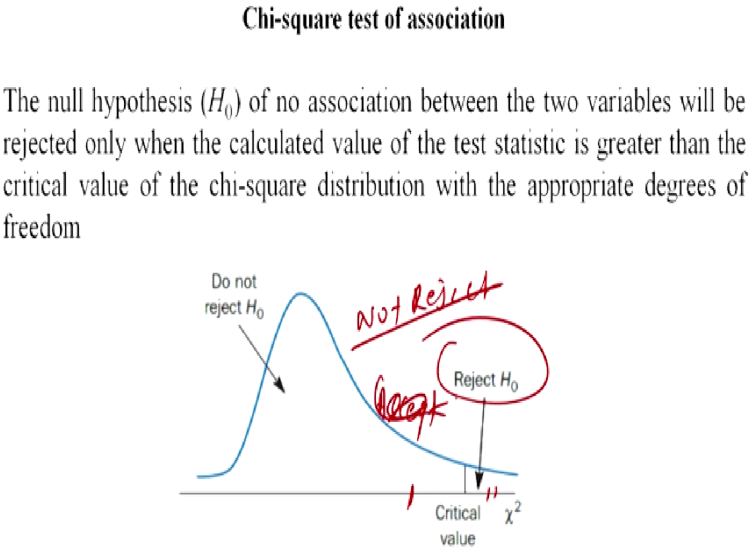
That is, $df = (r - 1)\ (c - 1)$

$df\ (r-1)(c-1)$

To determine whether a systematic association exists, the probability of obtaining a value of chi-square as large or larger in the one calculated is estimated. The important characteristic of the chi-square statistic is the number of degrees of freedom associated. Now generally we know that the number of degree is written is equal to the number of observations less by the number of constraints.

So generally it is n-1 in many cases, so n is the sample size minus one, but here it is slightly different. In this case of a chi-square statistic, the number of degrees of freedom is equal to the product of the number of rows r less one and the number of columns less one, so r-1 x c-1, this is our degree of freedom.

(Refer Slide Time: 26:23)

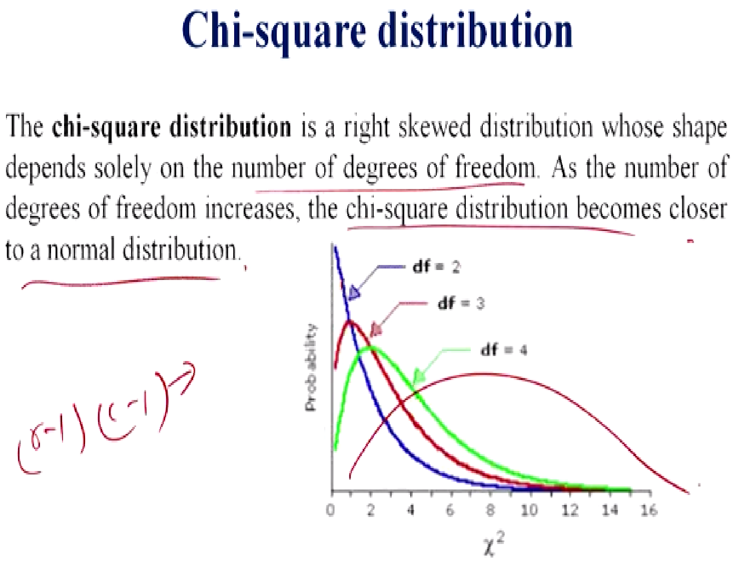


Chi-square test of association

The null hypothesis (H_0) of no association between the two variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-square distribution with the appropriate degrees of freedom

The null hypothesis of no association between the 2 variables will be rejected only when the calculated value of the test statistic is greater than the critical value of the chi-square distribution. What is it saying? The null hypothesis of no association between the 2 variables will be rejected only when the calculated value is greater. So your critical value is here, if your calculated value falls here, then you reject, but suppose it falls here, then you cannot reject, accept is not correct, but saying not reject is much preferable.

(Refer Slide Time: 27:02)



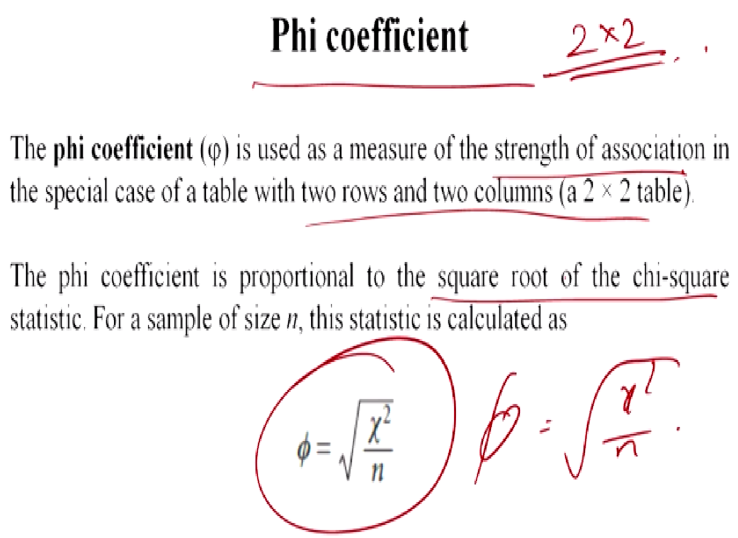
Chi-square distribution

The **chi-square distribution** is a right skewed distribution whose shape depends solely on the number of degrees of freedom. As the number of degrees of freedom increases, the chi-square distribution becomes closer to a normal distribution.

$(r-1)(c-1) \rightarrow$

Now how does this chi-square distribution look like? The chi-square distribution is actually a very skewed distribution to the right. Now it is right skewed distribution whose shape depends solely on the number of degrees of freedom. As the degrees of freedom increases, so when r-1 x c-1 goes on increasing, this distribution becomes more like a normal distribution. So you see when the degree of freedom is 2 this case, when degree of freedom is 3, when degree of freedom is 4, you can see moving more towards the positive side. So now slowly it will tend to towards become the normal distribution.

(Refer Slide Time: 27:44)



Phi coefficient

2×2

The **phi coefficient** (φ) is used as a measure of the strength of association in the special case of a table with two rows and two columns (a 2 × 2 table).

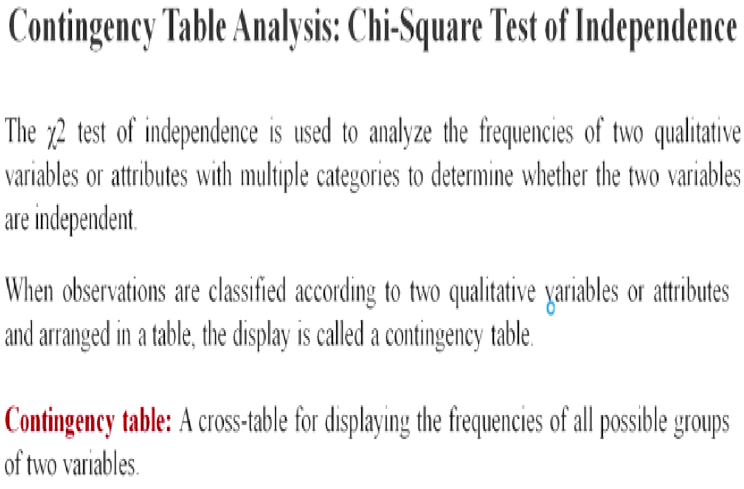
The phi coefficient is proportional to the square root of the chi-square statistic. For a sample of size *n*, this statistic is calculated as

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

What is this phi coefficient when we calculate chi-square. The phi coefficient is a measure of the strength of association in a special case of a table of 2 rows and 2 columns, in a special case of a chi-square where you have only 2 x 2 matrix. The phi coefficient is proportional to the square root of the chi-square, now that is why the formula is this. It is phi = root over chi-square/n, so it is the square root of the chi-square statistic.

(Refer Slide Time: 28:19)



Contingency Table Analysis: Chi-Square Test of Independence

The $\chi 2$ test of independence is used to analyze the frequencies of two qualitative variables or attributes with multiple categories to determine whether the two variables are independent.

When observations are classified according to two qualitative variables or attributes and arranged in a table, the display is called a contingency table.

Contingency table: A cross-table for displaying the frequencies of all possible groups of two variables.

Now what is this contingency table, I will just show you. The contingency table something looks like this. So there is the variable A and the variable B. So you see when observations are classified according to 2 qualitative variables or attributes, the display is called a contingency table. So whenever 2, in our case income and cleanliness let us say are 2 qualitative variables which are measured may be in a nominal way. So then in that case this table is called a contingency. It is a cross table for displaying the frequencies of all possible groups of 2 variables.

(Refer Slide Time: 28:21)

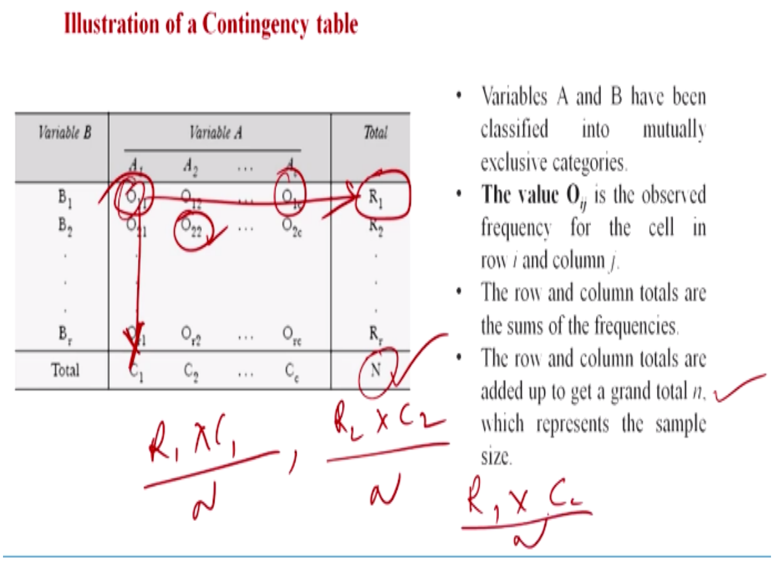


Illustration of a Contingency table

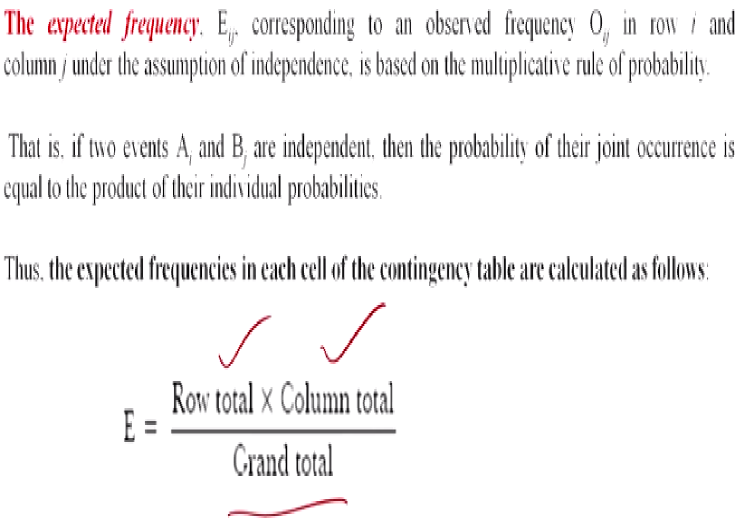
Variable B	*Variable A*				*Total*
	A_1	A_2	...	A_c	
B_1	O_{11}	O_{12}	...	O_{1c}	R_1
B_2	O_{21}	O_{22}	...	O_{2c}	R_2
.					.
.					.
.					.
B_r	O_{r1}	O_{r2}	...	O_{rc}	R_r
Total	C_1	C_2	...	C_c	N

- Variables A and B have been classified into mutually exclusive categories.
- **The value O_{ij}** is the observed frequency for the cell in row *i* and column *j*.
- The row and column totals are the sums of the frequencies.
- The row and column totals are added up to get a grand total *n*, ✓ which represents the sample size.

$\frac{R_1 \times C_1}{N}$, $\frac{R_2 \times C_2}{N}$ $\frac{R_1 \times C_2}{N}$

So this is what I was talking about. So R1, R2, R3, you have Rs let us say, C1, C2 and goes on and this is the grand total. Variable A and B have been classified into mutually exclusive categories. The value O is the observed frequency for the cell in row i and column j. So let us say this is first row first column, so you have observed frequency, similarly for every one you have. The row and column totals are the sum of the frequencies, this is what I was talking about. The row and column totals added up to get a grand total of N, which represents the sample size.

(Refer Slide Time: 29:36)



The *expected frequency*, E_{ij}, corresponding to an observed frequency O_{ij} in row *i* and column *j* under the assumption of independence, is based on the multiplicative rule of probability.

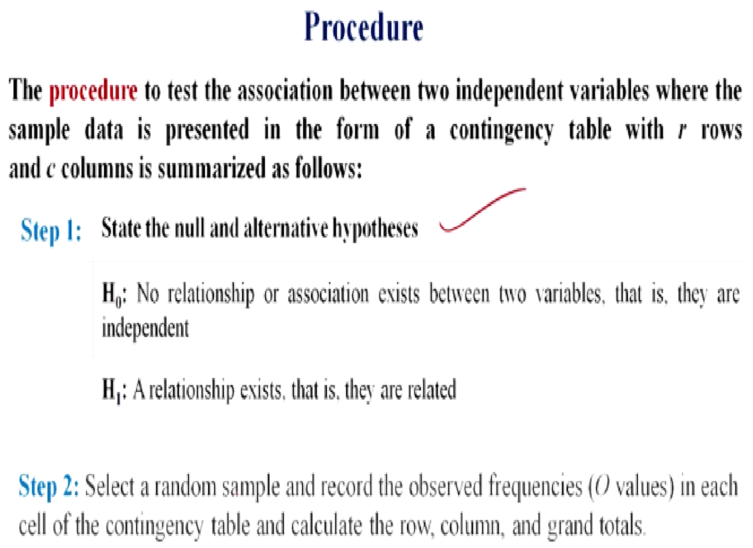
That is, if two events A_i and B_j are independent, then the probability of their joint occurrence is equal to the product of their individual probabilities.

Thus, the expected frequencies in each cell of the contingency table are calculated as follows:

$$E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Now the expected frequency is calculated like this. So row total x column total/grand total. So in this case for example if I want to measure the cell frequency of this one expected, so I will not take any other row but only this row, the corresponding row and the corresponding column, so R1 x C1/N. For let us say this case, it will be R2 x C2/N. Suppose it would have been this one, let us say this case, so it is R1 x Cc/N, so it goes on, you can calculate.

(Refer Slide Time: 30:23)



Procedure

The procedure to test the association between two independent variables where the sample data is presented in the form of a contingency table with r rows and c columns is summarized as follows:

Step 1: State the null and alternative hypotheses

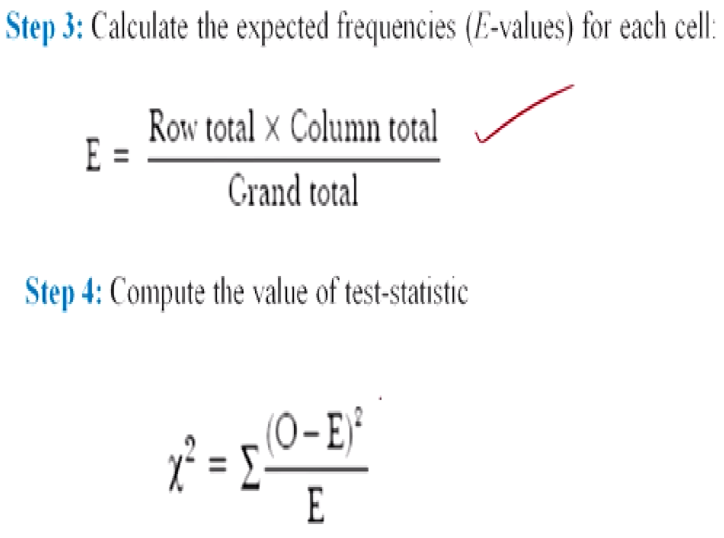
H_0: No relationship or association exists between two variables, that is, they are independent

H_1: A relationship exists, that is, they are related

Step 2: Select a random sample and record the observed frequencies (O values) in each cell of the contingency table and calculate the row, column, and grand totals.

The procedure to test the association between two independent variables where the sample data is presented in the form of contingency table is summarized as follows. First state the null and alternate hypothesis. What is this, no relationship or association exists between 2 variables. The second alternate is a relationship exists, that they are related. Now select a random sample and calculate the expected frequencies.

(Refer Slide Time: 30:51)



Step 3: Calculate the expected frequencies (E-values) for each cell:

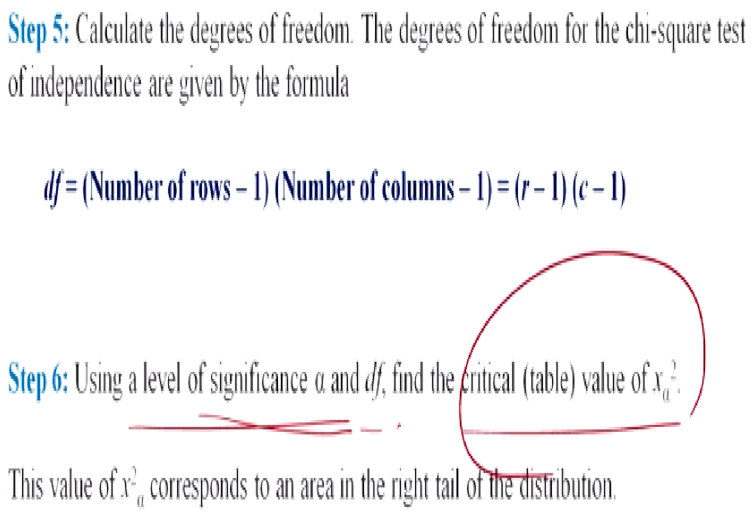
$$E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

Step 4: Compute the value of test-statistic

$$\chi^2 = \sum\frac{(O-E)^2}{E}$$

So the expected frequencies have been calculated for each cell and then we compute the chi-square value.

(Refer Slide Time: 31:01)



Step 5: Calculate the degrees of freedom. The degrees of freedom for the chi-square test of independence are given by the formula

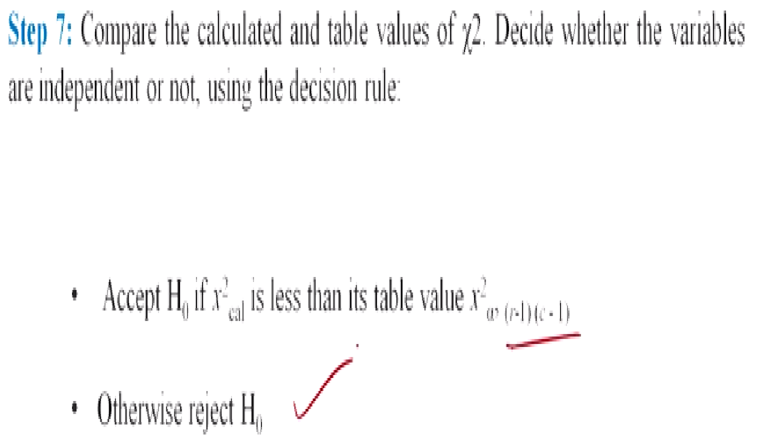
$$df = (\text{Number of rows} - 1)\,(\text{Number of columns} - 1) = (r - 1)\,(c - 1)$$

Step 6: Using a level of significance α and df, find the critical (table) value of x_α^2.

This value of x^2_α corresponds to an area in the right tail of the distribution.

Then after you calculate the chi-square value, you calculate the degrees of freedom. So we have seen how to do that, so number of rows -1 and number of columns -1. So now we use a level of significance and degree of freedom, we find the critical value for chi-square and now from here once we have calculated the critical value and we have the actual value or the calculated value, then we can compare and say whether the hypothesis is to be rejected or null hypothesis to be rejected or not rejected.

(Refer Slide Time: 31:37)



Step 7: Compare the calculated and table values of $\chi 2$. Decide whether the variables are independent or not, using the decision rule:

- Accept H_0 if x^2_{cal} is less than its table value $x^2_{\alpha\,(r-1)(c-1)}$
- Otherwise reject H_0

So finally when you compare the calculated and table values of chi-square decide whether the variables are independent or not using the decision rule accept H0 if chi-square calculated is less than the table value, chi-square alpha which is as per the degrees of freedom, otherwise reject H0. Well, I will wind up here. In the next session while continue with this, we will start with an example, we will solve the problem from chi-square statistics and then we will get into some other nonparametric techniques, a few more are left, and then we will wind up and I will try show you on the software also and we will continue with it. Thank you so much.