Marketing Research and Analysis-II
(Application Oriented)
Prof. Jogendra Kumar Nayak
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture - 25
Non-Parametric Test – I

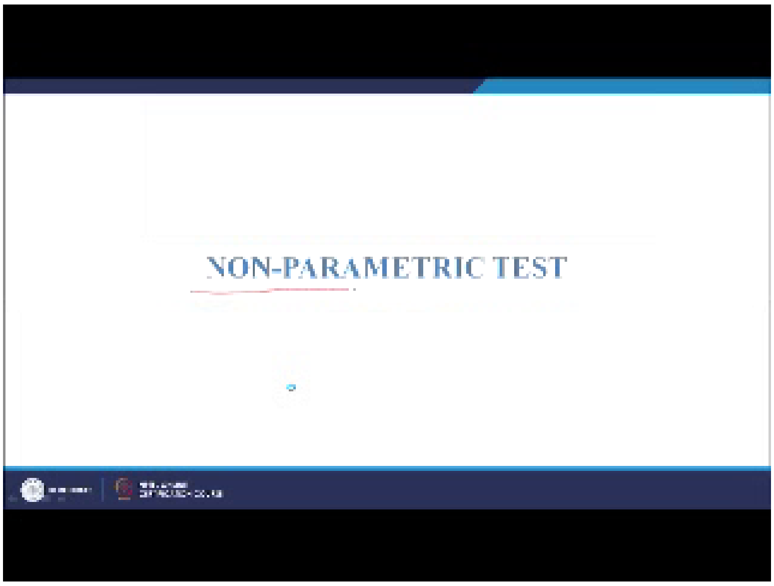
Warm welcome to everyone! Welcome to our lecture series of the course Marketing Research and Analysis. Last class we have discussed about data purification if you remember, in which we discussed about how to correct your data and proceed further to you and for your analysis. In that context we have discussed about three important problems regarding data purification and cleaning.

That is one was the handling of the missing data and where I explained how you should be scientifically handling the missing data and going for it so that you did not be putting in the data, forge fitting the data, or you know just putting in some data for the heck of it. Second case was the case of outliers in which I explained you the role of outliers and how you should handle outliers and what should be the logical explanation behind outliers.

Third thing was we discussed about normality. Covering these things and linearity also we covered. So we covered basically these things and then we said if the data is behaving properly and is fitting all those parameters then the researcher should move further for statistical analysis. However, whenever you correct the data for example you tried to check for the normality.

And you see and you tried to transform the data and try to convert the non normal data into a normal data that means a skewed data into a normal data. In such a condition it is well and good if the data convert data itself to a normal one but what happens if data does not become you know normal in nature or the distribution is still not becoming normal? So in such a condition we do not have an option left with us rather we have to use new technique or use new techniques which are very important and they are called the non-parametric test.

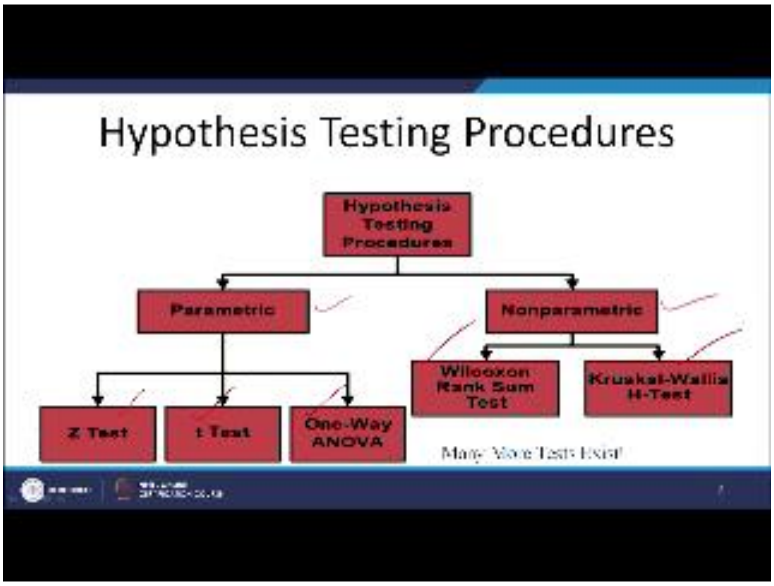
(Refer Slide Time: 02:32)



Non-parametric as the name suggests it means that the distribution of the data is not normal in nature. So it is not normal in nature so the normality assumption is invalid in such a condition, the normal statistical test are not applicable. If you do any statistical analysis which follows the normal distribution like a t z f test then it would be wrong.

So let us understand that whenever a data is not a normal data and then we have to use non parametric test. So this is what I am trying to say.

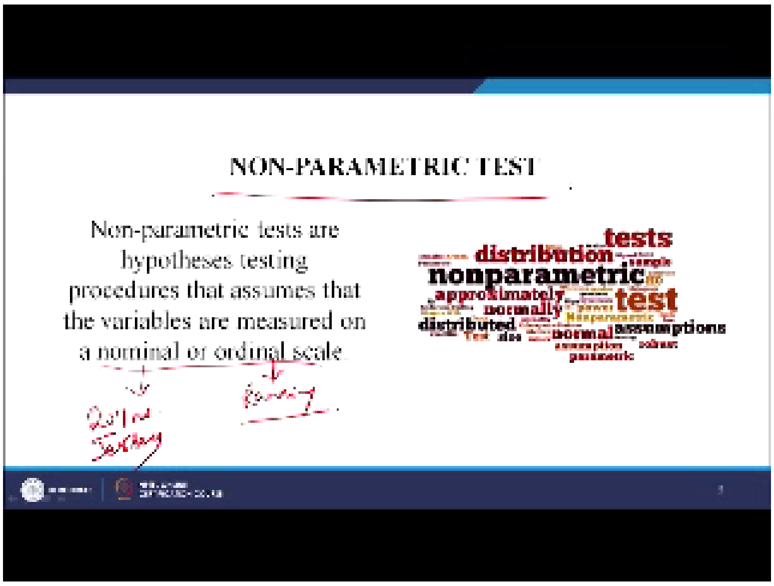
(Refer Slide Time: 03:08)



The hypothesis testing procedure is basically divided into two parts: one, when the data is parametric or normal in nature and the other is the non parametric or non normal in nature. So for this you have test like that t, Z, one way anova and for the other hand for the non parametric test your test like the Wilcoxon rank sum test, Kruskal Wallis test, Mann Whitney U test, ranking you know Sign test.

Several tests are there. And we will be trying to cover. So this part we had introduced as a part of the hypothesis testing. But this is the way we will be going deep into this right in the in the further, future classes. Today we are going to today and then maybe in the next classes we are going to deal with the non parametric test which, usually are not taught much normal classes.

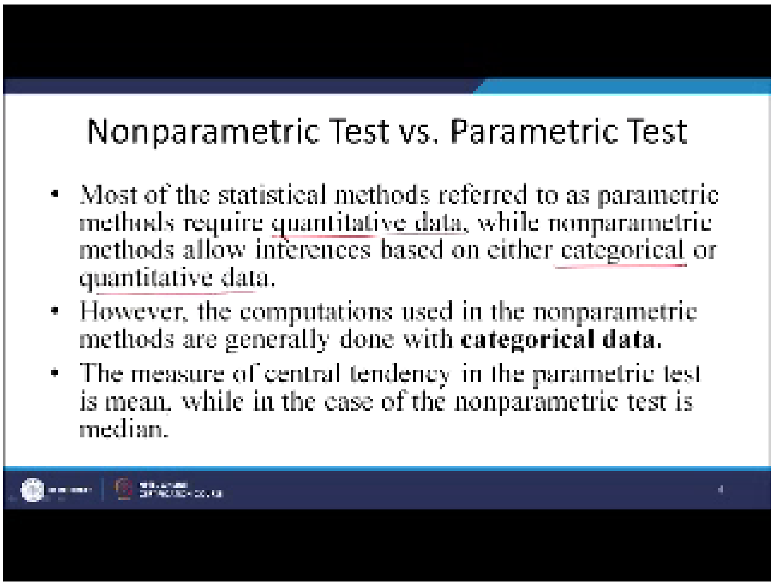
(Refer Slide Time: 04:01)



So what is this non-parametric test? Non parametric test are hypothesis testing procedures that assumes that the variables are measured on a nominal or ordinal scale. So what is this saying is that the variables are measured not in an interval, ratio scale, but it is measured in a nominal or ordinal. I hope you remember what is this nominal or ordinal. And nominal scale was something which represents something.

For example the roll number of a student right. The jersey number jersey number of a player right. These were in the nominal data. Ordinal data are those data which follows the order. So it is for example ranking data. So in such kind of problems when you are using a representation like a nominal or ordinal scale in such condition, the data does not follow a parametric you know condition. So in such a condition we are using automatically a non parametric test.

(Refer Slide Time: 05:08)

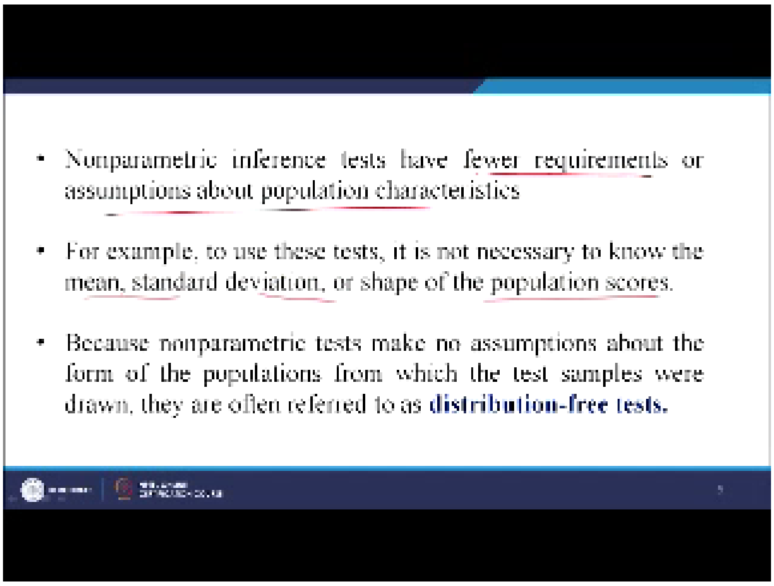


Non parametric test versus parametric test: Let us see some of the differences. Most of the statistical methods referred to as parametric methods require quantitative date. But you might not have been having all the time data which are quantitative in nature. You might be having some kind of data which categorical in nature while nonparametric data allows inferences base on categorical or quantitative, both.

So suppose for example I am trying to see the categories of people in terms of income, gender for example so these are some of the data which I have. Now how do I use these? These data are not going to follow any normal distribution. However, the computations used in the non parametric methods are generally done with categorical data as I said just now. The measure of Central tendency, this is very important.

The measure of Central tendency in the parametric test is mean which we generally use. But the mean is not the right one in the non parametric test because of the conditions of the violations of the normality conditions. And we therefore, we use another one which is the median which is more robust to outliers and other problems, right. So here, when the data is a non normal when you are doing a non parametric test the measure of Central tendency we use is the median.

(Refer Slide Time: 06:48)

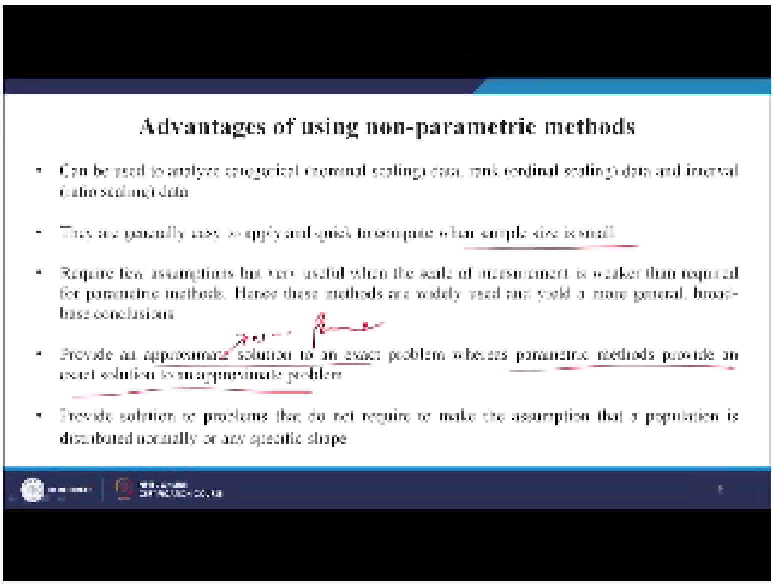


Nonparametric test, inference test have few requirements or assumptions about population character. They are not bothered when you are doing a non parametric test the good thing is that you are having very few requirements right or assumptions about the population characteristics which is mandatory or is requisite in case of a parametric test ok.

For example to use this test it is not necessary to know the mean the standard deviation or even shape of the population scores because nonparametric test make no assumptions about the form of the population it is also from which the test samples were drawn, they are often referred to as distribution free test. That means what? There is no particular distribution that is required right.

We do not we are not concerned about the distribution. So we are saying it a distribution free test. Many a time you see that, when you have collected a raw data, from the field, the data might be following a certain distribution or might not be following. If it is following a certain distribution then fine. You can go for any statistical the test. But suppose it is not showing any particular distribution then you have an option to go for a non parametric test. Now let us see the advantages.

(Refer Slide Time: 07:54)



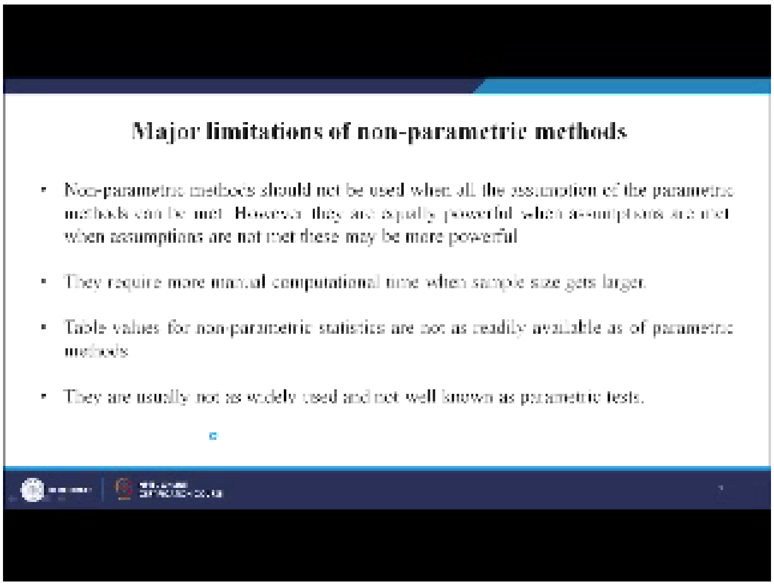
It can be used to analyse categorical data, rank data and intervals data also. Now why interval I am saying? This interval data can also be converted into an ordinal data. Let us say for example, you have certain let us say for example the data of income of people. You have let us say the data of income of people. You know the exact income. Now you can put into it into three categories.

Let us say category 1 is up to 1 million from category 2 is from 1million let us say 3 millions from above 3 million it is category 3.So you can create those categories. They are generally used to apply and quick to compute when sample size is small. One problem with you know the parametric data is that the sample size has to be large. If you have a very small sample size then the parameters cannot test be done properly?

It requests few assumptions but very useful when the scale of measurement is weaker than required. So we said about the nominal and ordinal scales which are the elementary scales and at the interval and ratio are much better scales. It provides an approximate solution to an exact problem whereas parametric methods provide an exact solution. So the difference between parametric and nonparametric is that no doubt there is no doubt that parametric methods are more robust and more accurate.

But sometimes you do not have a choice you do not have an option so you need to go for a rather approximate solution and thus the nonparametric comes into play right. It provides solutions to problems that do not required to make the assumption that a population is distributed normally for any specific shape which I have already explain now. What are the limitations?

(Refer Slide Time: 09:41)



Limitations just now I said. One is that it gives you an approximate solution. It should not be used when all the assumption of the parametric methods can be met. If you are data is meeting the parametric assumptions of normality, linearity, outliers and everything then you should be using parametric test because that is more robust right. If they are not matching then only you should go for a non-parametric.

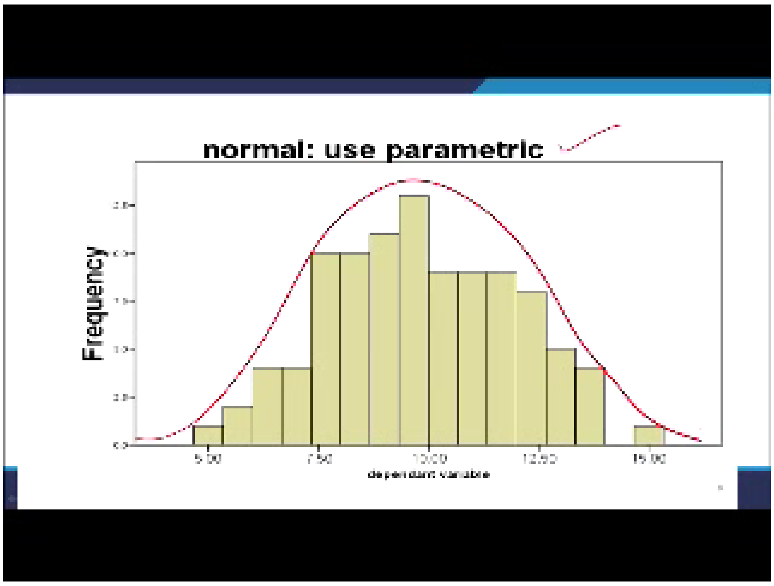
However there are equally powerful when assumptions are met. When assumptions are not met these may be more powerful. When assumptions are met non parametric methods are more powerful. They require manual computational time when sample size gets larger. So one of the limitations is it requires more manual computation when sample size is becoming larger and larger, ok.

The table values are not as readily available as a parametric test. Parametric test is used in the table values that table value area under the curve for example the t value and all these things which are generally not available. They usually not as widely used and not well known as the parametric test so if you even remember your own class, may be a faculty must have taught you the parametric test quite well.

And might be I am not saying this but might be there might have ignored the non parametric test because we generally it is not given much important but that is not correct because it is only about the violations of the normality or the assumptions of the normal distribution. So if something is why is not matching meeting the assumptions, then automatically have to go for non parametric.

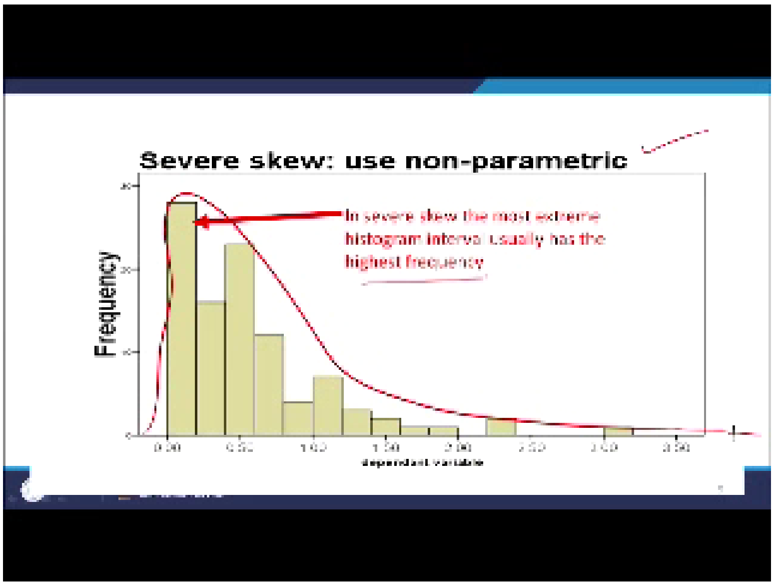
So you should be learning you should have the knowledge of the non-parametric test. So let us see this is a case you know that if it is normal.

(Refer Slide Time: 11:22)



Suppose the data seems to be normal correct. It is equally dispersed and distributed. So in such a condition I am using a parametric method. So you see the distribution is quite normal. If I draw if I draw a normal distribution curve it looks exactly correct.

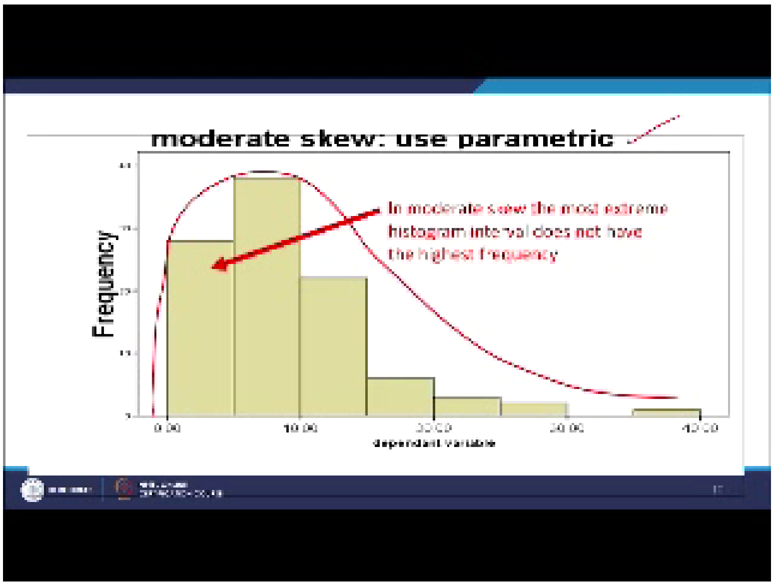
(Refer Slide Time: 03:08)



Now this is, let us see this data. How you, how does this look like? So it looks like this? So it is severely skewed. Now skewed to the positive or negative? This is a positive skew because it is moving towards the positive right? The positive skew so when is there is a skewness, serious skewness, either whichever side that does not matter, you use that non parametric methods.

So by this time I have already explained in the last, last class you know in the earlier classes that how do you should take the normality and other things right the skewness of the data. In a serious skew, the most extreme Instagram interval usually have the highest frequency so this is skewed correct.

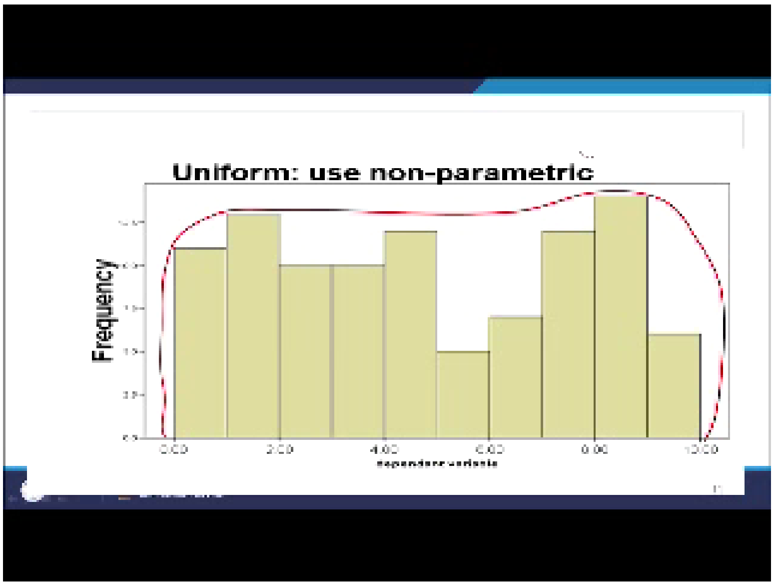
(Refer Slide Time: 12:23)



Now this is moderately skewed. Now this, this is little better than the last one right. This is moderately skewed. Now you can use parametric method why? It says if in case if your data is really skewed then you should go for a non parametric method. But however if it is slightly away moving away from the normality and it is although it is skewed but not very much skewed then, in that case it is all it is better to use transform the data convert into normal data and use a parametric test.

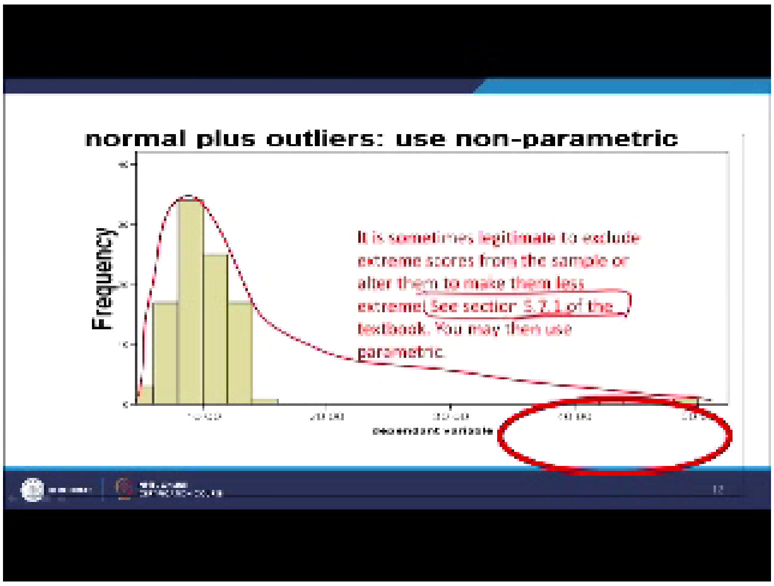
In moderate skewness the most extreme instagram you see this one interval does not have the highest frequency. So the frequency is moving towards this side the largest frequency is rather here is not exactly the extreme end, uniform.

(Refer Slide Time: 13:08)



Suppose this is interesting if your data has this kind of a shape. Now when I am having more of a flat, it is a flat distribution, so in this kind of distribution it is a complete violation of the normal distribution or the bell's curve. So, you should use a nonparametric method again ok.

(Refer Slide Time: 13:31)

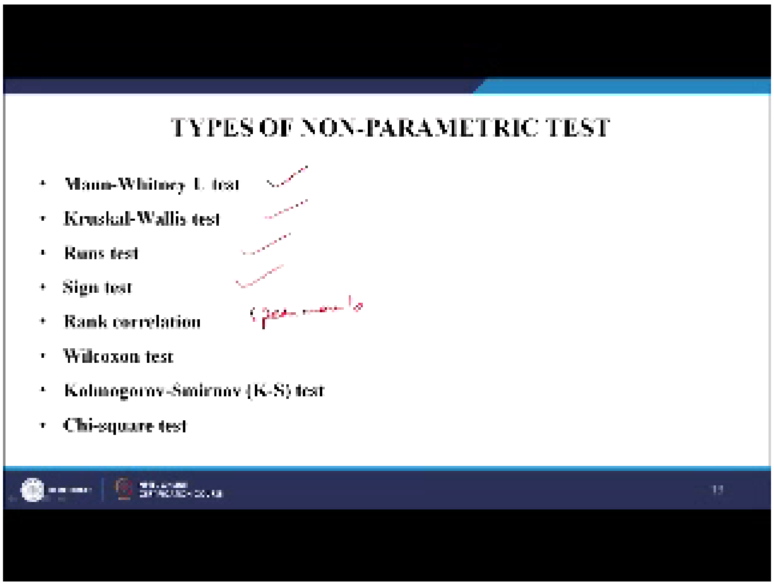


Now normal plus outliers this data is let us say I not even normal and it is also has large number of outliers. Can you see this? So this is something and the outliers are also existent. So when an outliers existent, automatically we know that the median is the most powerful Central tendency that we use and in such conditions again nonparametric test are the right method.

So it is sometimes legitimate to exclude extreme scores from the sample or alter them to make them less extreme you can do that as outliers as I already explained. You can thus then use that this from the book. So this section you can avoid it. So you may then use the parametric method, ok. So once you see that there is a legitimate score from the sample.

Either you alter them otherwise you remove them or do whatever you can rationally logically and then use the parametric methods. Otherwise in such a condition the nonparametric method is the right method.

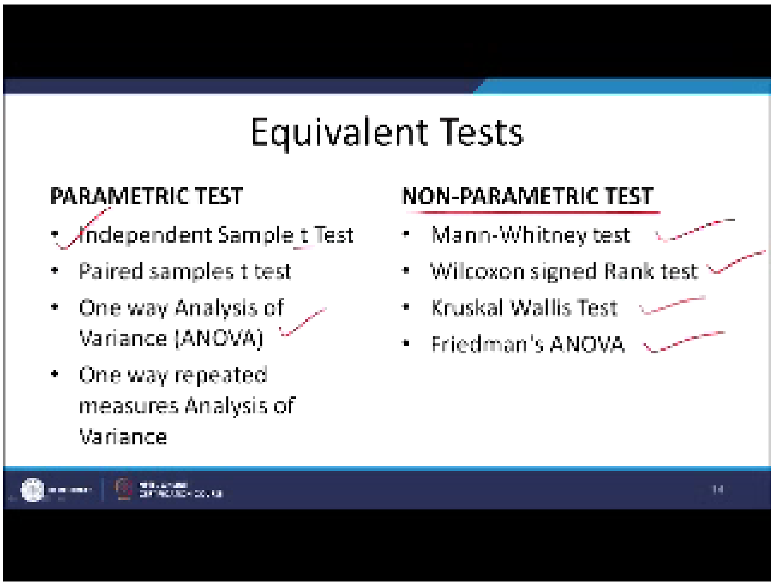
(Refer Slide Time: 14:40)



Now we are coming to we have understood when to use your non parametric methods and how they are in which conditions they are important. But now we will see some now we will see some of the test. So some of the tests for example the Mann Whitney U test which is very popular and it is like you know to, is an independent sample T test. For example the Kruskal Wallis test, runs test, Sign test, Rank correlation.

So here we will use the spearman correlation very popular Spearman's correlation Wilcoxon test and K-S test and Chi square test. So these are some of the nonparametric tests that are widely utilised under the non-parametric heads. So these are some of the equivalents tests which I tried to explain so that you will feel more comfortable.

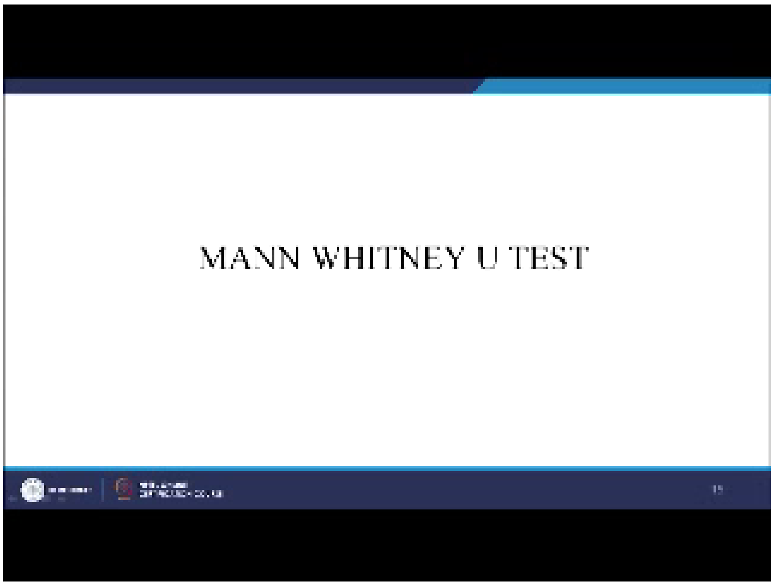
(Refer Slide Time: 15:31)



Parametric tests for example you have in the parametric test where the data is normally distributed, independent sample T test is there correspondingly for you have non parametric data you have a Mann Whitney U test. This is a when you have a paired sample t-test in case of a normal distribution in a parametric test paired sample t test against that you have here Wilcoxon sign rank test, ok.

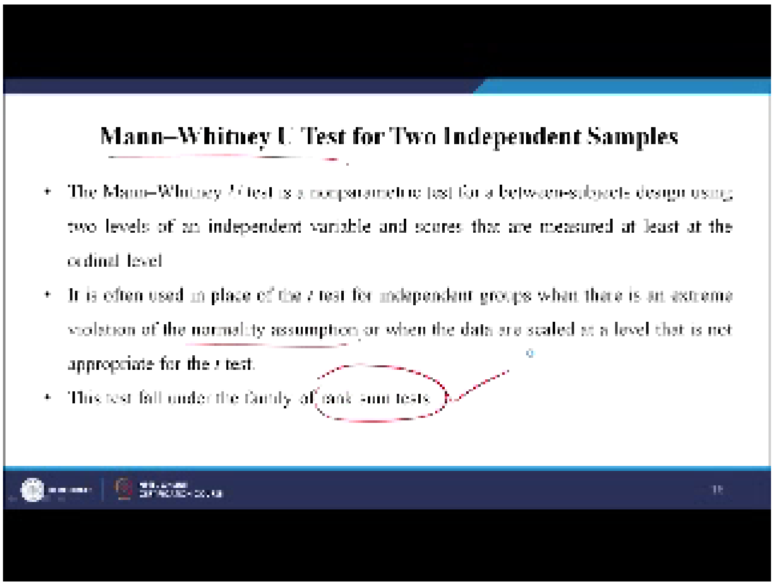
There you had a one way analysis of variance here you have a Kruskal Wallis test, there you had one be repeated measures analysis of variance where the repetition is allowed, here you have said Fredman's ANOVA. These are some of the equivalent tests. This is now to understand that the parametric or non parametric are entirely not different. They are there is a lot of similarity. But because of the conditions of normality and other problems you have to use this non-parametric test. So let us start with the first one.

(Refer Slide Time: 16:28)



Mann Whitney U Test or which is correspondingly similar to the independent, independent sample T test.

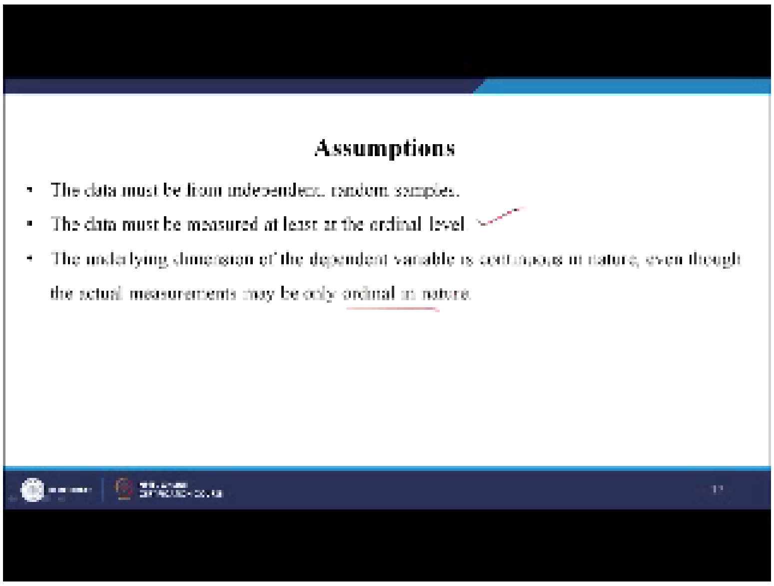
(Refer Slide Time: 16:37)



So the Mann Whitney U test is a non parametric test for a between subject design using two levels of an independent variable. There are two levels. For example, let us see you want to know that there are two groups of people male and female and their respective let us say interest for soft drinks for example. So if I want to compare how much are males spending on soft drinks and how much are females spending on soft drinks, I want to test.

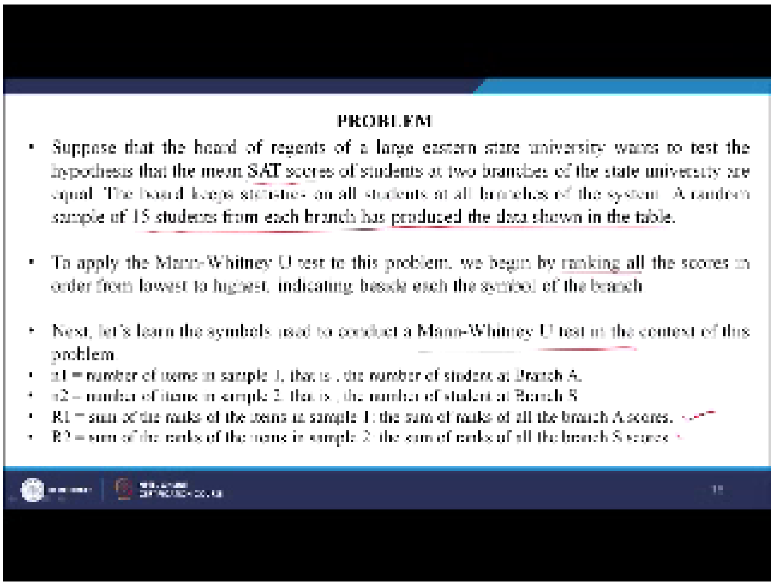
By chance I found the data collected is absolutely not following a normality pattern, normal distribution then, I have no option but I need to go for a non parametric test under which I am using the Mann Whitney U test. It is often used in place of the t tests for Independence groups where there is as an issue violation of the normality. This tests falls under the family of rank sum test. Now this important is this particular Mann Whitney u test you know is one of the, test that falls under the rank sum test.

(Refer Slide Time: 17:41)



There are few more also. I will be explaining one by one. What are the assumptions? The data must be from Independent random samples. The data must be from independent random samples. It is measured at least at the ordinary level if you measure in the interval also, no issues, no issues at all because it can be always taken as ordinal. The underlying dimension of the dependent variable is continuous in nature even though the actual measurement maybe only ordinal. So these are some of the assumptions of the Mann Whitney test. Let us see a problem.

(Refer Slide Time: 18:09)



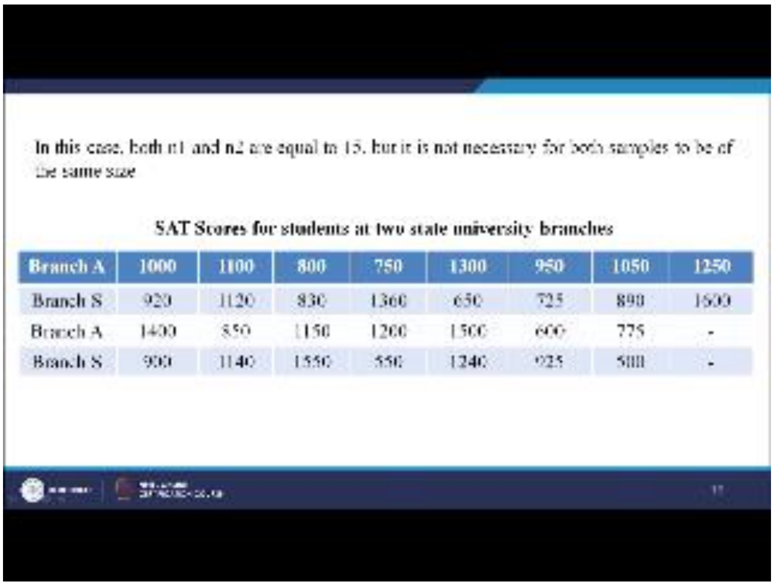
Let us try to solve a problem and understand I will be, I will explain it through a problem and also on the other side of the software. Suppose that the board of Regents of a large eastern state university wants to test the hypothesis that the mean SAT scores of students at two branches of the state university are equal. The board keep statistics on all students at all branches of the system.

A random sample of 15 students from each branch has produced data as shown in the table. So remember generally we have seen that whenever you are using a nonparametric test that the sample size is also much lesser right. However, if you have sample size goes on increasing given the non parametric test tends to move towards the normality right. And then normal a pattern can be used normal parametric test can be utilised.

To apply the Mann Whitney U test to this problem, we begin how to fix time. We begin by ranking because you have heard in the last slide that you it says that it falls under the rank sum test ranking all the scores in the order from lowest to highest, right. Let us learn the symbols. What symbols do you use here? N1, number of items in sample one. So there are two groups so the number of items in sample one group one.

So the number of student in Branch A and 2 so the number of students in sample 2, that is the student at branch S, second branch. R1 is a sum of the ranks of the items in sample one so some of ranks of all the branches of A scores similarly, R2, sum of the ranks of the items in sample 2, so sum of the ranks in all the branches, of the branch S score. So these are four things which we require.

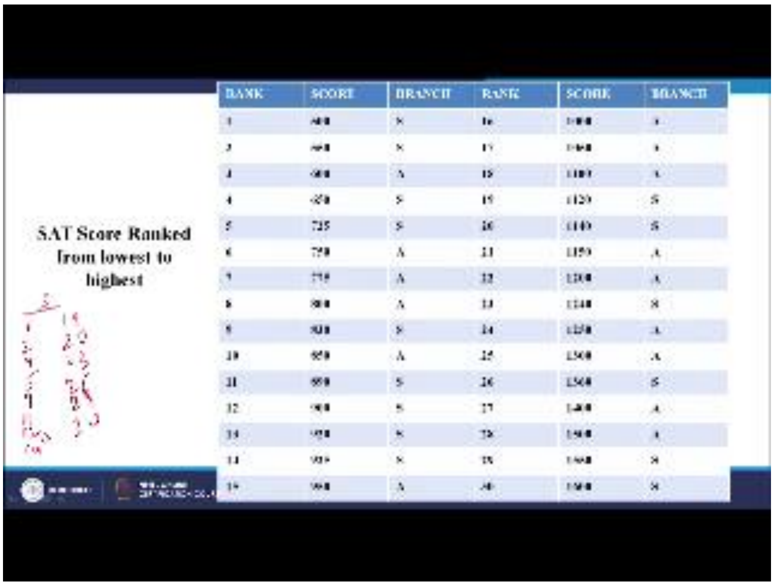
(Refer Slide Time: 19:58)



Branch A	1000	1100	800	750	1300	950	1050	1250
Branch S	920	1120	830	1360	650	725	890	1600
Branch A	1400	850	1150	1200	1500	600	775	-
Branch S	900	1140	1550	550	1240	925	500	-

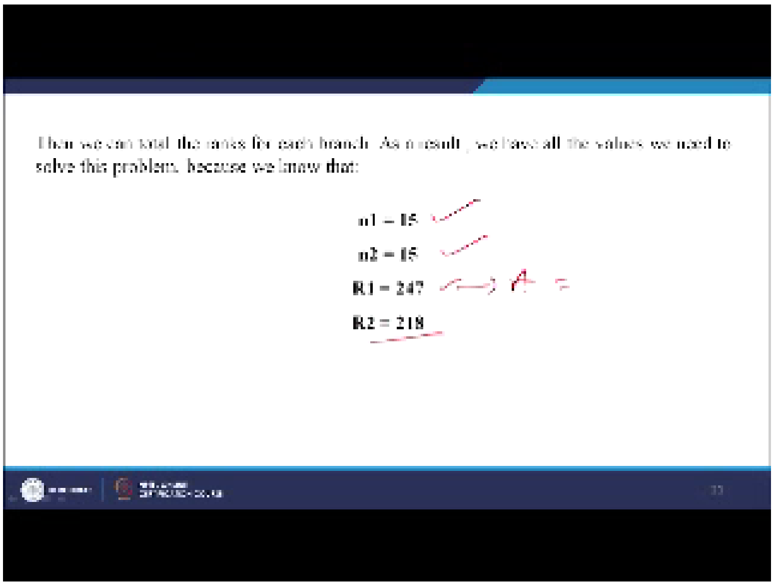
Let us take this. In this case both N1 and N2 are equal to 15. How many A are there? A - 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15. Similarly, if you look at in N2 branch S, how many are there? 1 2 3 4 5 6 7 8 and this is also 15. 15 ,15 both right.

(Refer Slide Time: 20:27)



Now we have now put in the ranks. Now if you see which is the lowest value out here? This is the lowest value which comes under the branch S. So the first one, first one, rank one is 500, Rank 2 is 550, rank 3 is 600, rank 4 is 650 and it goes on. Correspondingly we are writing in which order which branch are they coming? So 500 was in S, 550 was in s, 600 was in A ok. So we have given you know rank for, of the different scores and the branches ok.

(Refer Slide Time: 21:03)

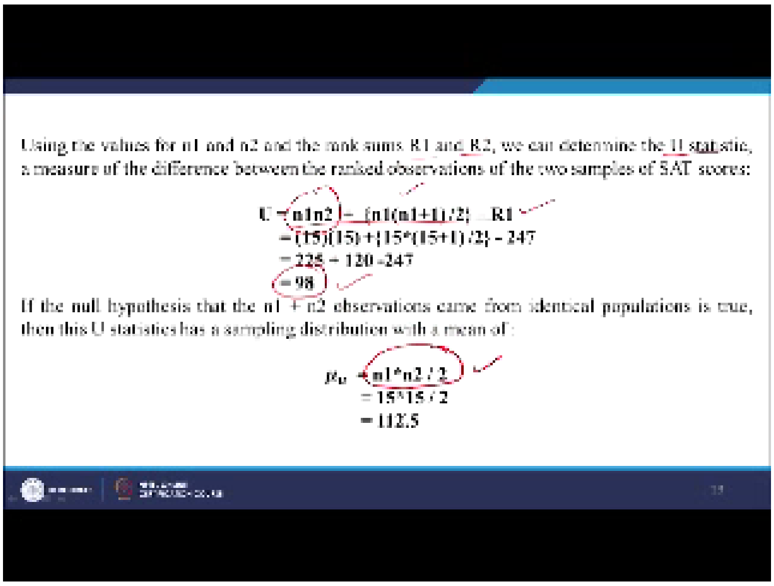


So once we have done this from lowest to highest you start from lowest to highest. So now what we can do is we can total the ranks for each branch. So you have done the branches, now branch A and S, now kindly add it up. As a result we have all the values we need to solve this problem. So we have n1 =15, n2 = 15. And R1 is 247. Now how does it come? Let us see here again.

So R1 is for which branch? Let us say it let us say count for this one for S. For S, what are the ranks? 2 then 4 then 5 then 9, 11 then 12 then 14 then 19 then 20, 23, 26, 29 and 30 ok if I add all of them, if our all of them all the total then how many is coming 30 59 65 85 88 108 128 137 147 151161 163 173 and 174 184 193 198 202 202 204 205. So 13 is also missed. 13 was also there. So, + 13. So if I had so it is almost coming to 218. 205 + 13 = 218.

Similarly if I add for the all the branch a it is 247. Now after doing this right 218, 247 has been done. So you can see this right. Now what?

(Refer Slide Time: 23:03)



Now using the values for n1 and n2 and the rank sums R1 and R2 we determine the U statistics, right a measure of difference between the ranked observations of the two samples of SAT scores.

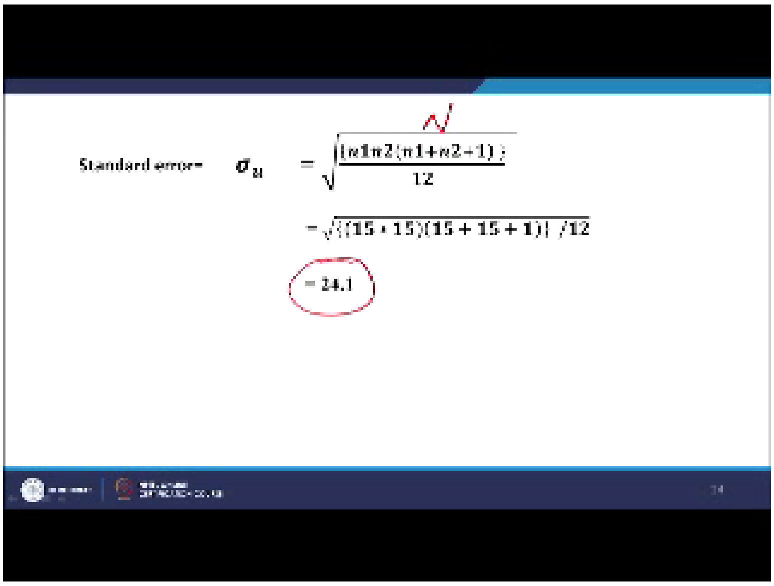
$$U = n1*n2 - \{n1(n1+1)/2\} - R1$$

So if I, if I do this, now you might be thinking to what if I take n2 not n1? You can do it and you will see that there is not much of a difference.

The only thing is this remains the same now this one will become n2 into n2 plus one and this is becomes minus R2. So that's the only difference that will happen. But you see the end result will be more or less the same. So when we calculate days it is coming 98 right U is equal to 98. So if the null hypothesis that the n1 plus n2 observations came from identical populations is true.

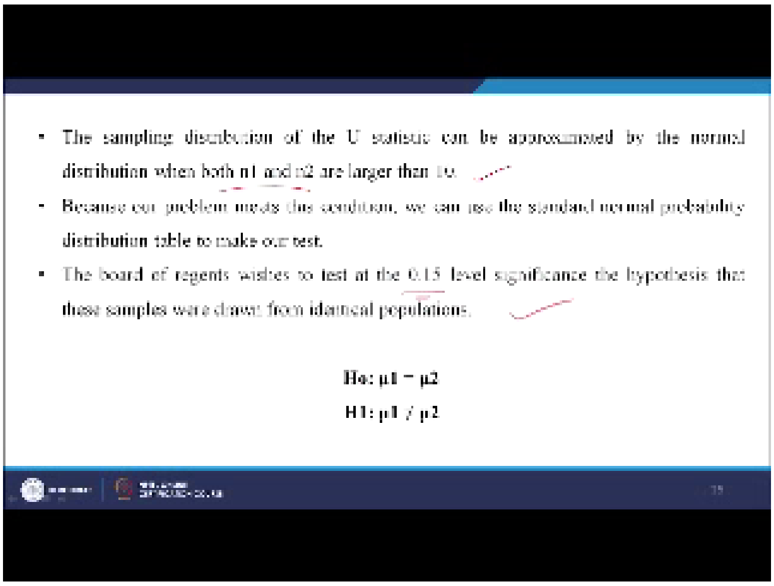
So what is the null hypothesis? That both the populations is, there is no difference between population right? The mean of the populations, so other two samples that the mean of 2 samples is the same. That means they originate from the same population. So the mean is in this case is measured as n1 into n2 divided by 2. So this has come 112.5.

(Refer Slide Time: 24:22)



The standard error is given as n1 into n2 divided by n1 plus n2 which also can be written as like this n1 + 12. So this is 24.1. So you have got now U you have got Mu mean, the standard error right.

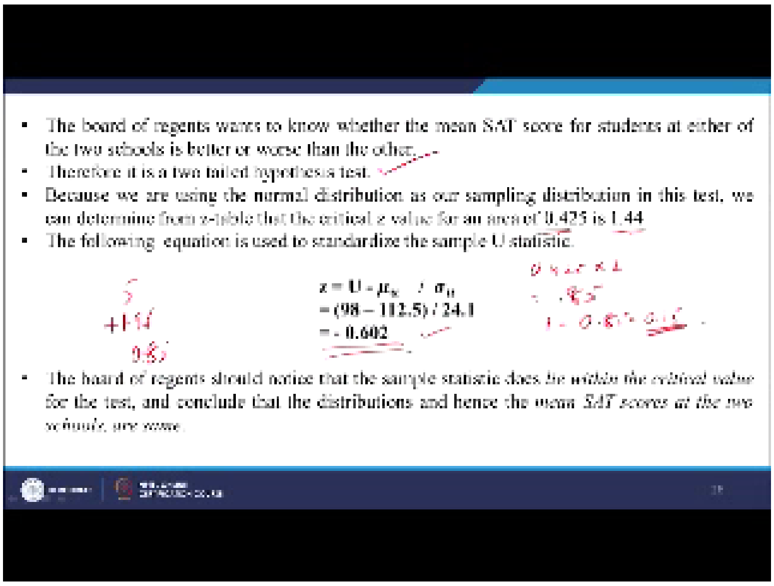
(Refer Slide Time: 24:40)



Now the sampling distribution of the U statistics can be approximated by the normal distribution when both n1 and n2 are larger than 10. Now this is the minimum value 10 if it is more than 10 you can use a normal distribution. Because our problems meet this condition because you are 15 we used the standard normal probability distribution to make our test. The board of regents wishes to test at the 15% level of significance.

The hypothesis is that the samples are drawn from identical population. In the first, what is my null hypothesis? The null hypothesis Mu1 is equal to Mu. There is no difference. What is my alternate hypothesis? There is a difference between the Mu1 and Mu2.

(Refer Slide Time: 25:23)

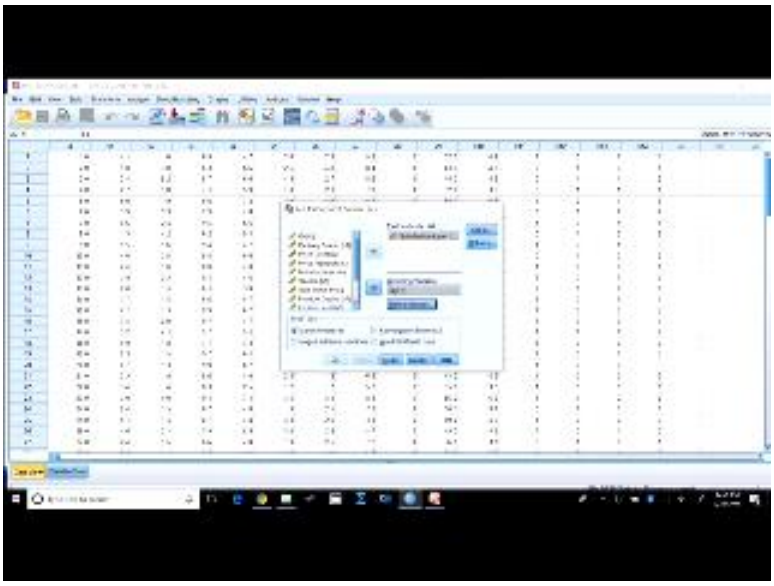


The board of regents wants to know whether the mean scores of the students is either of the schools is better or worse than the other. So since it is it is better or worse. It can be in this side of that side. So it can be it is a 2 tail test. We determine from the Z Table that the critical value Z value for an area of 0.425, 0.425 is 1.44. Why 0.425? 0.425, 0.425 into 2 is equal to how much? 0. 85 correct so this is 1 - 0.85 is equal to 0. 15, this is the level of significance.

So this remaining 0.85 was divided into two parts, right. So Z is equal to U – Mu divided by standard error. So that comes 0.602. So now Z value is this, what do I interpret, right? The board of Regents should notice the sample statistics just lie within the critical value right for the test and conclude that the distributions and hence the means scores at the two schools are not different because it lies within the limit of plus 1.96 or generally for 5%.

For 15% it is 0.85. So 0.85 let us say so now when it lying within the limit within the boundary we cannot reject the null hypothesis right. So and we say conclude that there is no difference between the two population means ok. This is how you do the Mann Whitney U test I will show you through a SPSS output also.

(Refer Slide Time: 27:09)



Now how do I do this? Let us go to analyse. Now it is a species has already given you in or nonparametric test 1 dialogue box for this. So go to Legacy. Now you see there are Chi square, Binomial, runs, one sample, independent sample, K sample, Tool related sample K related samples, many things are there. So what we are talking about is the independent sample.

So I want to see difference between the two independent sample groups? So what I am taking. For example I want to take the satisfaction level. Is there any difference in the satisfaction level between the two groups? So now what I can do is I can take firm size as my group variable. Now define the groups firm size, where is firm size? So you see firm size is 0 and 1. So this is zero here and this is one.

So there are two values so I will insert 0 and 1 so once you do this right and then you press OK then it will give you the result for the Mann Whitney U test. Well, what I will do is I will wind up here. So in the next lecture I will explain you how to do the Mann Whitney U test, SPSS software also. So I hope today the classes ok clear to you.

And you have understood what it means by the non parametric test and the first one we started is Mann Whitney u test is similar to the Independence sample T test and how it is to be done I was showing and may be, we will continue in the next class. Thank you so much.