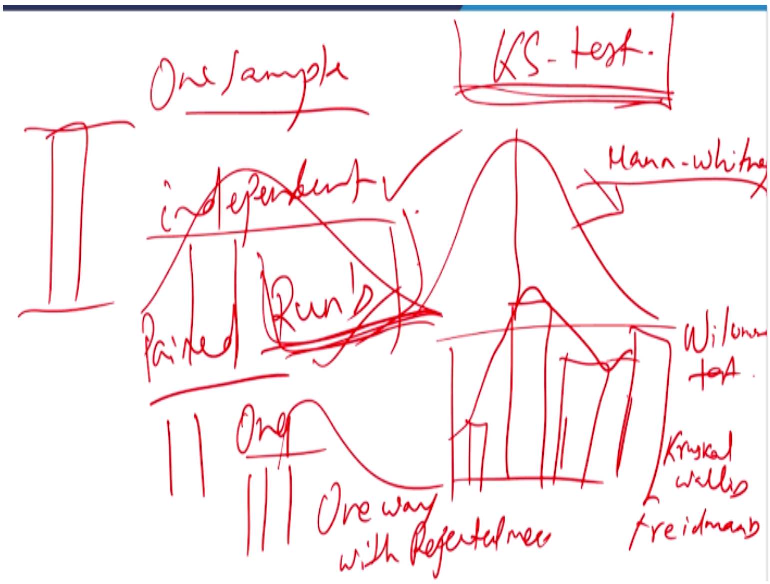
Marketing Research and Analysis-II
(Application Oriented)

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

Lecture – 32
Non-Parametric Test - VIII

Welcome friends to the course of Marketing Research and Analysis. Well we have been covering the nonparametric tests for the last few classes, around 6-7 classes. So we have seen why nonparametric tests are so useful, what is the importance of nonparametric test? And why they should be used in research studies? Basically in most of the classrooms I have seen teachers tend to do the statistical test based on the normal data, but we ignore the data which is not normal, so there nonparametric tests come of great importance, great value.

(Refer Slide Time: 01:18)



So we have seen for example when the data is not normal, you can convert the data to a normal data. So if the data is not normal, so if the data is skewed to the right or left whatever or it is a kurtotic data, then you can convert the data and make it look like a normal data. So what is the normal data mean, the central tendency the mean, median, mode lie on the same line, in the same center, but if suppose you convert the data into a normal data and you do it then well and fine.

But if the data is still not normal, is not behaving normally, it is still a skewed data, so maybe if you make a histogram, something the data it looks something like this. So if you see if this is the pattern of data, then you will not say if you take a path, then you cannot say this data is a normal data. So when the data is not normal, it could be for many reasons, one of them being the presence of outliers, so outliers can affect the data, so we have seen that, and when your data is a non-normal data, so you use nonparametric test and you have similar nonparametric tests as you have for parametric tests.

So for example when you have one group only, one sample, then we say one group for example, then we say we use a one sample test. So here on the other hand when you do a nonparametric test, you do a K-S test basically. So the one sample test helps you to find out whether the population and the sample they are connected or they are related or not similarly the K-S test also helps you to check normality and it helps you to find out whether the sample is part of the population or not.

You have the independent sample t test, so when you have sample t test for 2 groups, in the nonparametric you have a Mann-Whitney test similarly, Wilcoxon test. So the Mann-Whitney test exactly does the same work what the independent sample t test does but the point is here the data is not normal and here the data is a normal data. So these things we have seen, for example you a have a paired sample t test for a sample group but done may be twice. So to do that in this case, you have a Wilcoxon test.

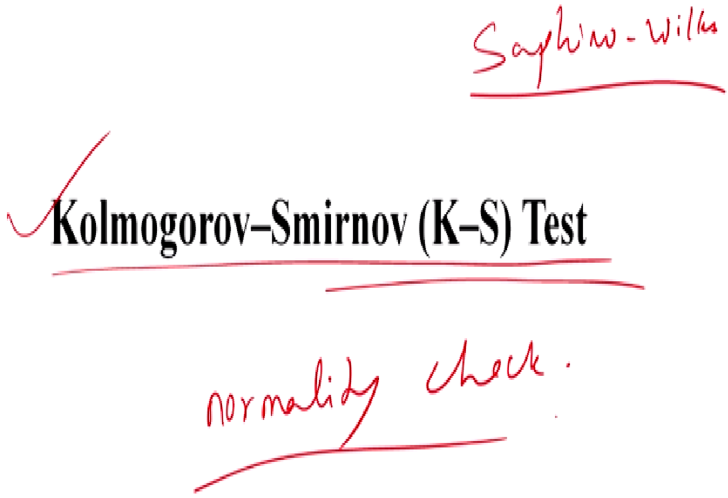
So the Wilcoxon test helps you to do the same thing. When you have more than 2 groups, we said it is a case of 3 groups, so it is a one-way ANOVA. So instead of a one-way ANOVA for the nonparametric test, we have a Kruskal-Wallis. Then we said what happens if my data has repeated measures, that means the same respondents are being repeated twice, thrice, four times, so the sample remains the same. So what will I do in this case? So you have a one-way ANOVA with repeated measures. Similarly you have a Friedman’s test for this.

So there are several things which you can see making it similar to the parametric ones, the nonparametric and the parametric. So one such test out of what we have done for example we have done the chi-square test which is a very powerful test, the McNemar test, the Mann-Whitney, the Kruskal-Wallis, the sign test, the runs test. The runs test is a test which I had already said, so since I am recapping, I am trying to tell you again.

So the runs test was at test where we are trying to check the randomness, the occurrence of a variable, if the variable or product or anything that is occurring in a systematic manner or it is coming in a random manner. Suppose it is coming in a random manner, then it is okay, no problem, but if a systematic occurrence is there, then we would try to understand why this is happening systematically, is it because of some problem in the machine or some problem in understanding, what is happening here.

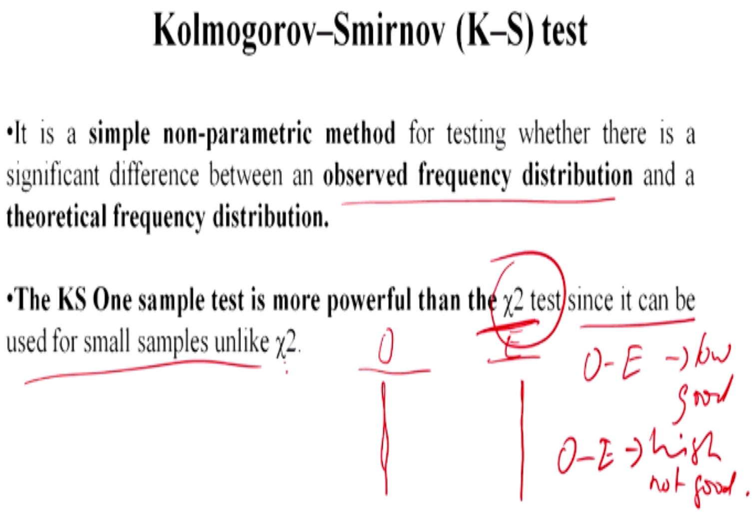
So these are some of the tests which are very, very helpful, at least the for example the runs test is very helpful in a quality control case. So when the researcher is trying to do a quality control test in that kind, the runs test is very important. So today what I will do is we will wind up with one of the last test that I should have done earlier also, but no issues we will do it now.

(Refer Slide Time: 05:47)



So this is the Kolmogorov-Smirnov, K-S test. So this test you must have seen if you are working with the SPSS or any database. Then you see in software packages, this is always attached with the Shapiro-Wilk test, you will find it with Shapiro-Wilk test. Now these tests are used for normality checking. So what it basically does is it helps you to check the normality of the data, tells you whether the data is coming from the same population or not, the sample is coming from the same population or not let us say.

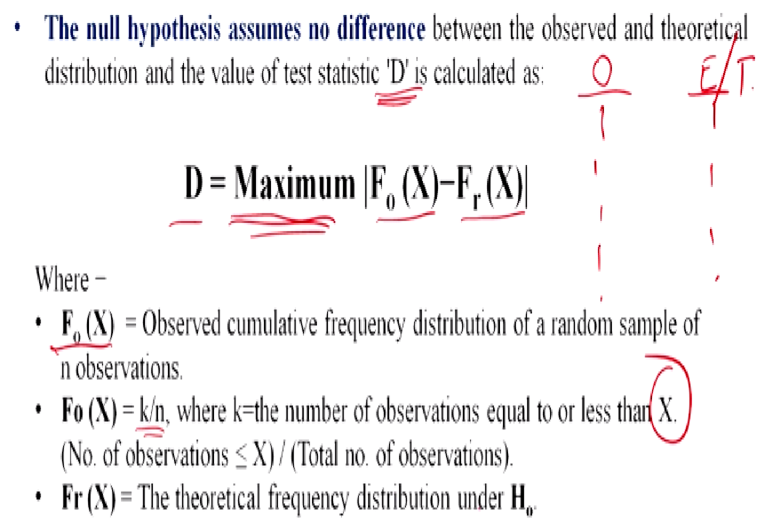
(Refer Slide Time: 06:32)



So what is this K-S test? It is simple nonparametric method for testing whether there is a significant difference between an observed frequency distribution and a theoretical frequency distribution. I have repeated it many times that whenever you have a observed set of data and you have an expected set of data, so then the data if there is a lot of difference between the 2 sets, then what happens, then we will assume that a model is a weak model or is not a fitting model, but if the observed frequency and expected frequency is close, then we will say the model is actually behaving like what it should be doing.

So the difference, the lower the difference, the better it is, if it is low it is a good thing, but if it is high, then not good. So the K-S One sample test is a more powerful than the chi-square since we are talking about nonparametric tests and chi-square deemed one of the most powerful, nonparametric follows the chi-square distribution, which is a right tailed distribution, right skewed distribution. Since it can be used, why it is powerful because it can be used for small samples which chi-square does not permit to sometimes. So unlike the chi-square, this test is a test which can be used for small samples.

(Refer Slide Time: 08:00)



Now what is the null hypothesis in this case? The null hypothesis assumes that no difference exists between the observed and the theoretical distribution. Now what do I mean by the observed, I just said you have some observed values, you have some expected or the theoretical, so this is what you expect. So this value that comes, let us say whatever values you get now this it says that there is no difference, we are saying no difference.

Now just imagine if there is no difference, that means what, it is the observed and the theoretical are more or less the same, okay fine, and what we calculate is the observed and the theoretical distribution value of test statistics which is given as D. D is calculated as

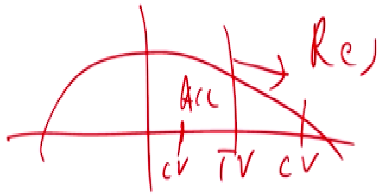
D = Maximum $|F_o (X)-F_r (X)|$

So F_0X or F_0X = observed cumulative frequency distribution of a random sample of n obviations, so there are n observations.

Now what is the frequency distribution of this random sample of n observations is called my F_0X. So F_0X is given by k/n where is what, the k is the number of observations equal to or less than X and n is my total number of observations and what is FrX, the theoretical frequency distribution, so this is the expected under H0.

(Refer Slide Time: 09:42)

- The critical value of D is found from the K-S table values for one sample test.
- **Acceptance Criteria:** If calculated value is less than critical value accept null hypothesis.
- **Rejection Criteria:** If calculated value is greater than table value reject null hypothesis.



So the critical value of D the difference is found from the K-S table values for a one sample test, you will see that. Now the acceptance criteria is same again like what you have to do, when your calculated value is more than the table value, it goes away from the table value. Then if the calculated value is less than the critical value, we accept, that means what if you remember if this is my let us say table value and this is my calculated value, this is going out, so this is a rejection case.

But if the calculated value falls here, then it is an acceptance case oaky. So acceptance criteria is if calculated value is less than the critical value or table value, accept the null hypothesis. If calculated value is greater than the table value, we reject the null hypothesis okay.

(Refer Slide Time: 10:42)

Problem Statement:

In a study done from various streams of a college 60 students, with equal number of students drawn from each stream, are we interviewed and their intention to join the Drama Club of college was noted.

	B.Sc.	B.A.	B.Com	M.A.	M.Com
No. in each class	5	9	11	16	19

It was expected that 12 students from each class would join the Drama Club. Using the K-S test to find if there is any difference among student classes with regard to their intention of joining the Drama Club.

Now let us take an example and understand the K-S test okay. This is a study done for 60 college students of various streams okay with equal number of students drawn from each stream, they are interviewed and their intention to join the Drama Club of the college was noted. So number in each class from B.Sc, B.Sc. is bachelor of science 5 students came, bachelor of arts 9 students, bachelor of commerce 11 students, master of arts 16 students, master of commerce 19 students.

But it was expected that 12 students from each class would join the Drama Club, but actual figure is the observed frequency is not same, it is 5, 9, 11, 15, 19. So this is more than the twelve, this is much less than the twelve. So can we infer something?

(Refer Slide Time: 11:27)

Solution

H_0: There is no difference among students of different streams with respect to their intention of joining the drama club.

We develop the cumulative frequencies for observed and theoretical distributions.

Now what is the null hypothesis in this case? There is no difference among students of different streams with respect to the intention of joining the drama club. So as you had expected that they all would be equal 12, 12, 12, 12, 12, then the observed is different but then your null hypothesis says there is no difference, the observed and expected are the same, and what is the alternative, obviously there is a difference between the observed and expected frequency. So we develop the cumulative frequencies for observed and theoretical distributions.

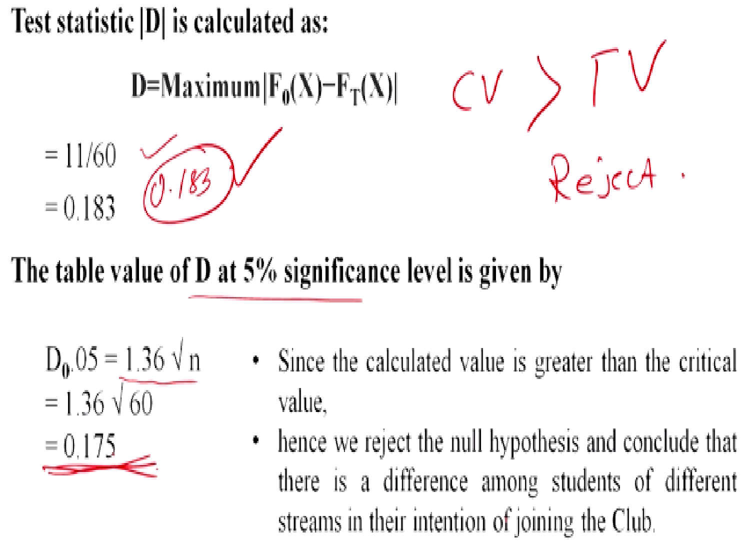
(Refer Slide Time: 12:04)

Streams	No. of students interested in joining		$F_O(X)$	$F_T(X)$	$\|F_O(X)-F_T(X)\|$
	Observed (O)	Theoretical (T)			
B.Sc.	5	12	5/60	12/60	7/60
B.A.	9	12	14/60	24/60	10/60
B.COM.	11	12	25/60	36/60	11/60
M.A.	16	12	41/60	48/60	7/60
M.COM.	19	12	60/60	60/60	0
Total	n= 60				

Now let us go. So the streams are B.Sc., B.A., B.Com., M.A., M.Com. Number of students interested in joining. So O is my observed frequency, so we have it already. So n is equal to how much 60, 5+9+11+16+19 is 60. Theoretically, I expected at least 12 will come from each class. So I said 12, why 60/5 = 12; so 12, 12, 12, 12, 12. Now what is my observed frequency now, the cumulative frequency, the first is 5/60, so 5 upon 60 + this, so (5/60) + (9/60) is (14/60), then +(11/60), so that comes (25/60), then I have + (16/60) right, so that gives me how much 25+16 is 41/60, and lastly this is 60/60, so we have got this.

Now what is my theoretical distribution at 12/60, second theoretical cumulative, next is 24, 12+12+12+12+12, so 12, 24, 36, 48, and 60 okay. So what is the difference? So if we subtract it, (12/60) – (5/60) = 7/60; (24/60) – (14/60) this value, 24/60 minus the observed, so this one minus this one right these 2, so (24 – 14), 10; (36 – 25), 11/60; (48 – 41), 7/60; 60/(60-60) is 0. So we got the difference. Now you see which is the highest difference? Where do you get the highest difference, the highest difference comes in the B.Com, case okay.

(Refer Slide Time: 14:16)



So this test is calculated as

D = Maximum $|F_0 (X) - F_r (X)|$

so that gives us 11/60, so this is equal 0.183, so you can divide it and find out. Now the table value of D at 5% level of significance, now significance level you can choose up to that is up to your choice. This 5% level of significance is given by how much, now

= 1.36 √n, now 1.36 let us say is what you are getting, so this value is coming to be how much, 0.175.

So since now this calculated value is 0.175, so this means what this is bigger or this is bigger, this is bigger right. So, calculated value is bigger than the table value. So what should you do in this case, so since it is bigger it is coming away out of it, so you reject the null. So since the calculated value is greater than the table value, hence we reject the null hypothesis and conclude that there is a difference among students of different streams in their intention of joining the club. So this is what you do, it helps you to find out.

Now I will show you one more advantage of the K-S test. The K-S test is also used for normality checking. So I will tell you how to do that. So in SPSS you see that **(Video Starts: 15:46)** this is the case where what you have taken is there are few patients, diabetic patients, and we have taken 2 groups of diabetic patients, one 0 who are in some control and the other is 1 who are in no control, so they are eating whatever they like, they do whatever they like, so they are free.

The controlled group is 1, they are under medication, they are doing some exercise and everything. The palpitation, so the doctors feel that because of the presence of diabetes, the palpitation, the heartbeat will change, will be more if it is higher diabetic. So what we have done is we have taken the palpitation levels. So now we want to check whether this is following a normal distribution or not. So the K-S test helps you in checking normality. So how do you do that?

So go to analyze, now this is a nonparametric test, so now let us go to the legacy, you can come to one-sample K-S. Now when you take the one-sample K-S, automatically you see we have taken the palpitation for the diabetes. So what we go to options, so this is the descriptive we want, so now the normal, this is the normal, so we go for the statistics we are finding here is you see the mean is 41. The palpitation mean for all controlled and uncontrolled is 41.13, the standard deviation is 11, the maximum and minimum values are this.

Now from the one sample, so if you look at this value it is giving a significant value of 0.138. Now when we say it is 0.138, what do you mean, that mean the null hypothesis is to be accepted or rejected, the null hypothesis is to be accepted because it is bigger than or greater than 0.05. So, if it is greater than 0.05, we have to accept the null hypothesis. Then what is the null hypothesis in our case, the null hypothesis that there is no difference between the data, so if there is no difference the dataset, then we will say it is same.

The dataset is more or less the same, and if it is more or less the same, it is like if you remember there is a test called Levene's test for normality in which homogeneity of variance in which we assume that the homogeneity of variance between 2 groups if you are testing the homogeneity of variance, the null hypothesis is to accepted and not rejected. So normally when we talk about any hypothesis testing, we generally infer that the null hypothesis should be rejected, but when you talk about homogeneity of variance, you say that the null hypothesis should be accepted, why?

Because you want to test 2 groups and say that these 2 groups are comparable because they have a similar kind of a variance. Similarly here when you talk about the significance if it is high, more than 0.05, that we means we assume that this data is a normal data and not normal, so the normality assumption is accepted. So that is why in the K-S test it should be a more than 0.05 case if it is a 5% level of significance, so that is coming true. So this tells us that the data is a normal test.

So the K-S test is largely used for checking this normality and thus it is a very useful test and you can do it yourself with some new dataset you can take it as and run it **(Video Ends:19:16).**

(Refer Slide Time: 19:17)

A summary of hypothesis testing

So we have understood all the tests now I just want to make a recap of the summary of the hypothesis testing and we will wind up here. So this is something that I want to tell you and you can use it later on

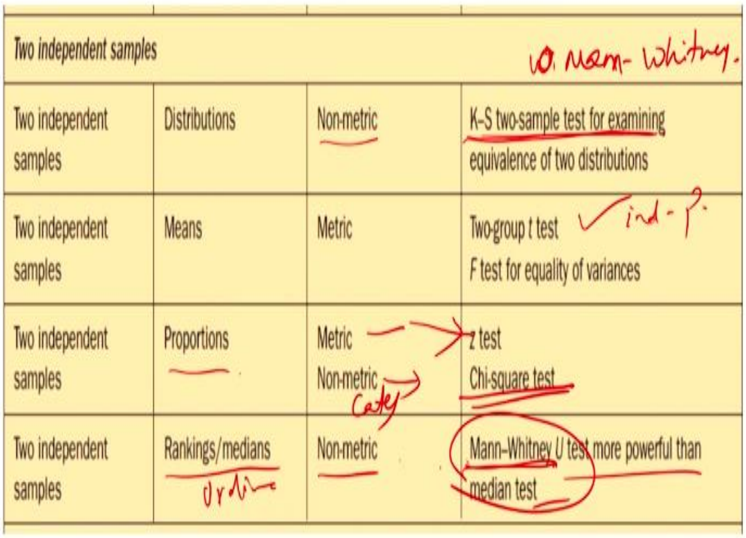
(Refer Slide Time: 19:30)

Sample	*Application*	*Level of scaling*	*Test/comments*
One sample			
One sample	Distributions	Non-metric	K-S and chi-square for goodness of fit Runs test for randomness Binomial test for goodness of fit for dichotomous variables
One sample	Means	Metric	*t* test, if variance is unknown *z* test, if variance is known
One sample	Proportions	Metric	*z* test

So what happens, when the sample is one sample, the application the distribution is given to you, the level of scaling and everything is given to you, now what are the tests that you use? A case of one sample and the scale is nonmetric, so you use a K-S, just one that now which you did the Kolmogorov-Smirnov test and the chi-square for goodness-of-it or you can also do a runs test for randomness, a binomial test with similar to a McNemar test also, so you can do that but when it is a one sample and the data is in a metric, so what you do is a t test.

If the standard deviation and variance is unknown to you, then you do a t test; if it is known, you do a z test, but in case of one sample and it is a metric, you do a z test. So just have a comparison and check the difference between the parametric and nonparametric.

(Refer Slide Time: 20:22)



Two independent samples			
Two independent samples	Distributions	Non-metric	K-S two-sample test for examining equivalence of two distributions
Two independent samples	Means	Metric	Two-group *t* test *F* test for equality of variances
Two independent samples	Proportions	Metric Non-metric	*z* test Chi-square test
Two independent samples	Rankings/medians	Non-metric	Mann-Whitney *U* test more powerful than median test

In case of a 2 independent sample and it is nonmetric, 2 independent sample that means 2 different groups are there, so we will use a K-S two-sample test or we can also use the Mann-Whitney test. Suppose you have 2 independent samples and it is metric, that means it is a more or less like a continuous, so use a two-group test or a it is independent sample t test. Now 2 two independent samples but they are in proportions, so it could be metric or nonmetric you can see.

So if it is metric you have a z test, but if it is nonmetric that means categorical, then use a chi-square distribution or a chi-square test. You have 2 independent samples and they are in ordinal ranking and it is nonmetric in nature, so we use the Mann-Whitney U test which is the most powerful than the normal median test. So these are some of the tests that we are talking about.

So just see this difference you can make out, now here it is ordinal and here sometimes you use the K-S two-sample test or you can use the Mann-Whitney, but in this case if it is 2 independent samples, so you mostly use a Mann-Whitney test which is a more powerful test.

(Refer Slide Time: 21:56)

Paired samples			
Paired samples	Means	Metric	Paired t test
Paired samples	Proportions	Non-metric	McNemar test for binary variables Chi-square test
Paired samples	Rankings/medians	Non-metric	Wilcoxon matched-pairs ranked-signs test more powerful than sign test

The last, if it is a paired sample and it is a metric, then you have a paired t test. If it is a paired sample but they are nonmetric, then we talked about the McNemar test for binary variables or the chi-square test. Paired samples again ordinal, so what you do, nonmetric is the Wilcoxon. Some of the things which I have even not preset here suppose for example as I said it is 3 groups for example, 3 samples and it is metric, then ANOVA.

One sample let us say but thrice done that means repeated and it is metric, then one-way analysis of variance with repeated measures, but suppose we have a one sample let us say and done repeatedly, repeated but it is nonmetric in nature, in such a condition we use a Friedman's test. So why I am explaining this to you is that you need to be very clear that you should not feel you know awkward that if my data is not normal, what will I do? In fact in the real world, many a times we see that the data might not follow a normal distribution, many many many times it should not follows a normal distribution.

In that case, it does not mean you will stop progressing or moving ahead, instead of that you could use a nonparametric test for your benefit, but only thing is you need to understand in which case which test is optimum or the right test to be used. If you understand this, then your job would be easy and you can very well publish your papers, you can use it in your thesis writing, and you can do everything that is possible in a parametric test that you can do it here.

I hope I have tried to clear the nonparametric test part to you, I hope it is clear to you, but if you have any doubts, you can surely put in your questions and ask and I will try to answer them and as much as possible. Because nonparametric tests generally I have seen people avoiding, this is my practical experience, so but that does not matter. If the more you do it, it becomes simpler and more easier to you and then you will feel much easier. Well, I hope this knowledge comes to your benefit and you use it well. Thank you so much. I wish you all the luck.