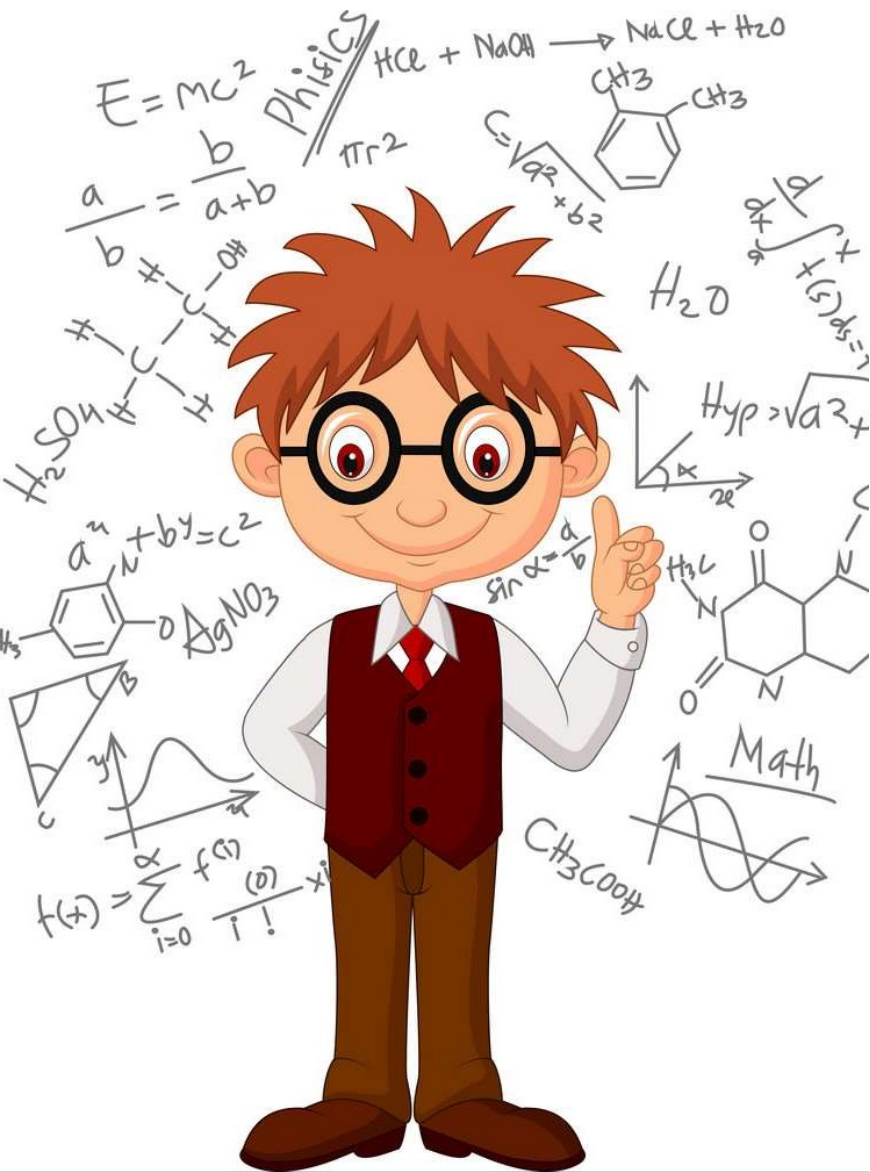


# Statistics for Data Analytics

Lecturer: Marina Iantorno

E-mail: [miantorno@cct.ie](mailto:miantorno@cct.ie)





In today's class we will cover:

- ☐ Kruskal-Wallis test
- ☐ U-Mann Whitman test

# Kruskal-Wallis Test



# Kruskal-Wallis

This test compares the median of more than one populations to see whether or not they are different.

The basic idea of Kruskal-Wallis is to collect a sample from each population, rank all the combine data from smallest to largest, and then look for pattern in how those ranks are distributed among the various samples.

# Kruskal-Wallis

Some characteristics of this test:

- We use it to detect whether three or more samples draw similar median values or not.
- Scale: ordinal (check Class 8).
- We do not need to assume normal distribution for the original variable.
- The samples must be independent (we cannot analyse the same observation twice).
- The scale must be the same in all the samples.
- We do not assume homogeneity.
- This test is unilateral (always to the right).
- This is the non-parametric version of ANOVA.

# Kruskal-Wallis

The CCT College wants to place a seminary for Data Analytics students of three different colleges. The director of the seminary wants to determine whether the three groups have similar knowledge. He decided to take an exam and if there is no difference between the results of the three colleges, he will place only one seminary for everybody. Three samples were randomly taken, and these are the results:

College A	51	32	17	69	86	62	96	
College B	14	31	68	87	20	28	77	97
College C	89	20	60	72	56	22		

At a 5% significance level, could only one seminary be in place for everybody?

# Kruskal-Wallis

## **Step 1:** Statement of the HT

$H_0$ : there is no difference in the knowledge level between the universities

$H_1$ : at least one of the universities draws differences in the knowledge level

# Kruskal-Wallis

## Step 2: Formula

$$H = \frac{12}{n(n+1)} \left[ \frac{(\sum R_1)^2}{n_1} + \frac{(\sum R_2)^2}{n_2} + \frac{(\sum R_3)^2}{n_3} \right] - 3(n+1)$$

R = rank

n = sample size

This test follows a  $\chi^2_{k-1}$

k = number of groups tested (in this case 3)



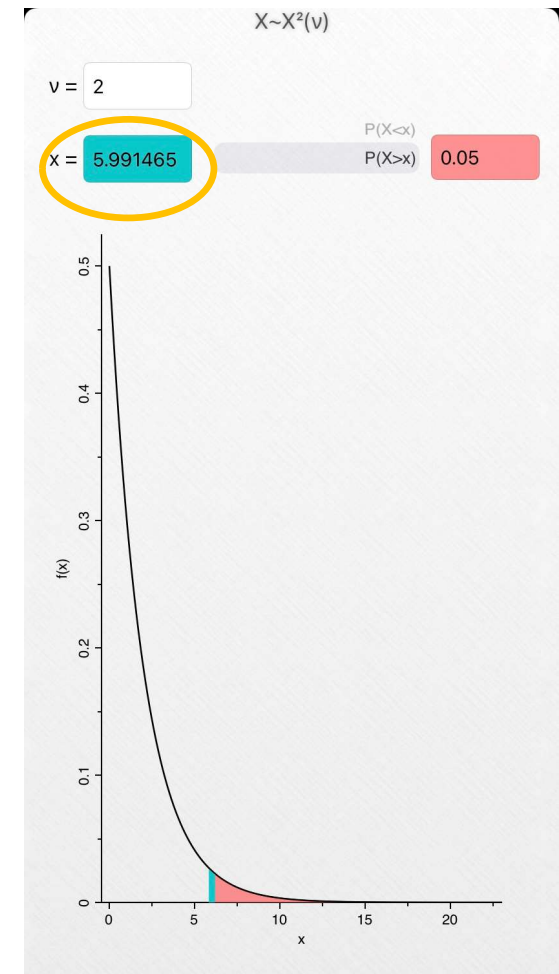
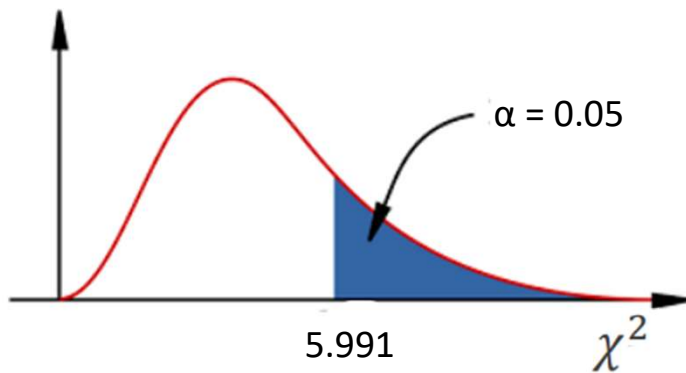
# Kruskal-Wallis

## Step 3:

➤ Table: Chi-Square

➤  $v = k - 1 = 3 - 1 = 2$

➤  $\alpha = 0.05$



# Kruskal-Wallis

## Step 4: Decision Rule

We reject  $H_0$  if  $H \geq 5.991$

We accept  $H_0$  if  $H < 5.991$



# Kruskal-Wallis

## **Step 5:** Calculation of H

- We will create a table in which the scores will be allocated from the lowest to the greatest as they were only one sample (we will not divide by groups).
- $\mathbb{N}$  will refer to the natural numbers (each cell will have its number).
- $R^*$  will be the rank we will assign to each score of the table.
- The green cells are optional, but they are useful to check if we are doing the correct procedure.

# Kruskal-Wallis

## Step 5: Calculation of H

Scores	14	17	20	20	22	28	31	32	51	56	60	62	68	69	72	77	86	87	89	96	97	Total
N	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	231
R*	1	2	3.5	3.5	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	231



If we are doing everything correctly, these two cells should result the same

When we have the same score, we cannot assign different ranks since they are the same values, and we could not make a difference. In this cases we calculate an average between the ranks.

In this case the first 20 would be rank 3 and the second 20 would be rank 4, then we do

$$R^* = (3+4)/2 = 3.5$$

# Kruskal-Wallis

## Step 5: Calculation of H

Now, we will create the tables for each college and allocating to each score the  $R^*$  obtained on the previous procedure

								Total
<b>A</b>	51	32	17	69	86	62	96	
<b>R<sub>1</sub></b>	9	8	2	14	17	12	20	<b>82</b>

								Total
<b>C</b>	89	20	60	72	56	22		
<b>R<sub>3</sub></b>	19	3,5	11	15	10	5		<b>63,5</b>

									Total
<b>B</b>	14	31	68	87	20	28	77	97	
<b>R<sub>2</sub></b>	1	7	13	18	3,5	6	16	21	<b>85,5</b>

We obtained here the following values for the formula

$$\Sigma R_1 = 82$$

$$\Sigma R_2 = 85,5$$

$$\Sigma R_3 = 63,5$$

\*As a control  $\longrightarrow \Sigma (\Sigma R_i) = \Sigma N$

$$82 + 85,5 + 63,5 = 231$$

# Kruskal-Wallis

## Step 5: Calculation of H

$$n_1 = 7$$

$$n_2 = 8$$

$$n_3 = 6$$

$$n = n_1 + n_2 + n_3$$

$$n = 21$$

$$\Sigma R_1 = 82$$

$$\Sigma R_2 = 85,5$$

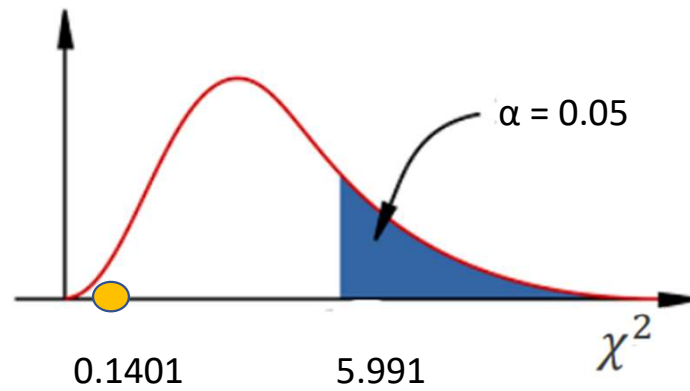
$$\Sigma R_3 = 63,5$$

$$H = \frac{12}{n(n+1)} \left[ \frac{(\Sigma R_1)^2}{n_1} + \frac{(\Sigma R_2)^2}{n_2} + \frac{(\Sigma R_3)^2}{n_3} \right] - 3(n+1)$$

$$H = \frac{12}{21 \cdot 22} \left( \frac{82^2}{7} + \frac{85,5^2}{8} + \frac{63,5^2}{6} \right) - 3 \cdot 22 = \boxed{0,14011}$$

# Kruskal-Wallis

**Step 6:** Result of the HT



We accept  $H_0$ .

# Kruskal-Wallis

## **Step 7:** Interpretation

*At a 5% of significance level, we could say that the knowledge level of the three colleges is similar, therefore, it would be possible to place one seminary for all the students at the same time.*



# U-Mann Whitman test

# U-Mann Whitman test


## **Objectives and characteristics of this test:**

- This test is another non-parametric test that wants to confirm whether or not there are differences between the median values of two populations.
- Scale: ordinal.
- It is not necessary to assume normal distribution for the original variable.
- We analyse two independent samples.
- The unit of measurement has to be the same for both samples.
- We do not need to assume homogeneity between the variables.

# U-Mann Whitman test

## Objectives and characteristics of this test:

➤ The alternative Hypothesis could be:

 < if we want to determine if the values of one sample are less than the other one

> If we want to determine if the values of one sample are greater than the other one

≠ if we want to determine whether there are differences or not

➤ This is the non-parametric version of the t-test for two populations means.

➤ This is also known as the Wilcoxon sum of rank test.

# U-Mann Whitman test

You work for a company that produces two brands of detergents: Super and Best. Before deciding the investment in advertising you need to know if there are differences in the acceptance between one and another one. The company gives a free trial to 25 people for the Super detergent and to 22 people for the detergent Best and asked them to qualify the performance from 1 to 10. You can find the results on the file “detergent.xlsx”. Try to find if one brand needs more publicity than the other one using 5% level of significance.

# U-Mann Whitman test

## **Step 1:** Hypothesis Statement

H<sub>0</sub>: There is no difference in the performance of the detergents

H<sub>1</sub>: There is a difference in the performance of the detergents

\*For example, if you wanted to test whether Super is performing better than Best, the performance should be greater, this would be your test:

H<sub>0</sub>: There is no difference in the performance of the detergents

H<sub>1</sub>: Super performance is greater than Best

# U-Mann Whitman test

## Step 2: Formula

$$U^* = \text{Mín} (U ; U')$$

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

$$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2$$

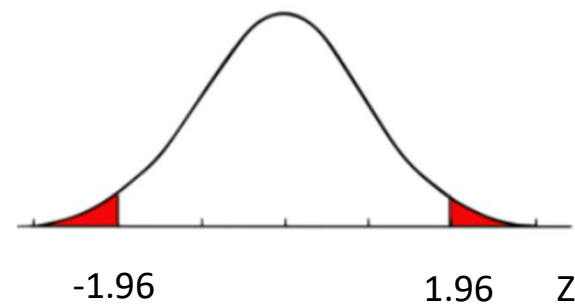
When  $n_1 > 10$  and  $n_2 > 10$  we approximate this to the Normal distribution and use this formula

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

# U-Mann Whitman test

## Step 3: Critical values

- Table: Normal
- Sign of  $H_1$ :  $\neq$
- $\alpha = 0.05$
- $\alpha/2 = 0.025$

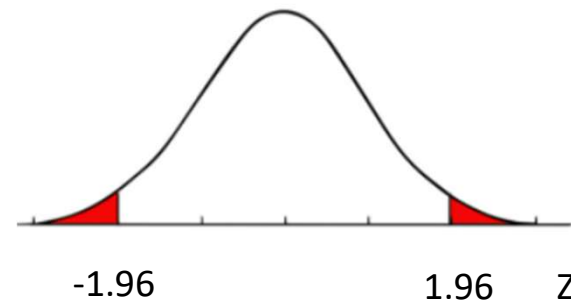


# U-Mann Whitman test

## Step 4: Decision Rule

We reject  $H_0$  if  $Z \leq -1.96$  or  $Z \geq 1.96$

I accept  $H_0$  if  $-1.96 < Z < 1.96$





# U-Mann Whitman test

## Step 5: Calculation of Z

We have to create a table that contains both brands and their performances.

H	I	
Detergent	Performance	
Super	4	{
Super	10	
Super	10	
Super	9	
Super	7	
Super	10	
Super	9	
Super	8	
Super	8	
Best	6	{
Best	6	
Best	4	
Best	4	
Best	10	
Best	10	
Best	9	
Best	7	
Best	10	

# U-Mann Whitman test

## Step 5: Calculation of Z

We sort the data from the smallest to the largest applying filter.

Deterge	Performan
Super	4
Super	4
Super	4
Best	4
Best	4
Best	4
Best	4
Super	5
Super	5
Super	6
Super	6

# U-Mann Whitman test

## Step 5: Calculation of Z

We will add another column next to the Performance with the ranks. Remember that if we have repeated values we have to calculate an average for the rank, therefore we will use the Rank Average function.

The first cell on the numerical column

H	I	J	K	L	M	N
Deterge	Performan	Rank				
Super	4	=RANK.AVG(I2,\$I\$2:\$I\$48,				
Super	4	RANK.AVG(number,ref, [order])				
Super	4					
Best	4					
Best	4					

Order is always 1 (ascendent)

(...) 0 - Descending  
(...) 1 - Ascending

The range will be all the values on the numerical column

# U-Mann Whitman test

## Step 5: Calculation of Z

We need to do the sum of the ranks. We will use here the function SumIF.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	Deterge	Performan	Rank										
2	Super	2	1		Super	n1 =	25		R1 =	=SUMIF(\$A\$2:\$A\$48,E2,C2:C48)			
3	Super	3	2		Best	n2 =	22		R2 =	SUMIF(range, criteria, [sum_range])			
4	Super	4	5										
5	Best	4	5										
6	Best	4	5										

Categorical variables

Rank

Brand

# U-Mann Whitman test

## Step 5: Calculation of Z

We can calculate U and U'. Replace the values according to the formula.

	E	F	G	H	I	J	K	L	M	N	O	P
Super	n1 =	25			R1 =	601.5		U =	=25*22+(25*(25+1)/2) - 601.5			
Best	n2 =	22			R2 =	526.5		U' =				

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \sum R_1$$

$$U' = n_1 n_2 + \frac{n_2(n_2+1)}{2} - \sum R_2$$

	E	F	G	H	I	J	K	L	M	N	O	P
Super	n1 =	25			R1 =	601.5		U =	273.5			
Best	n2 =	22			R2 =	526.5		U' =	=25*22+(22*(22+1)/2)-526.5			

U = 273.5

U' = 276.5

MIN(U ; U') = 273.5

# U-Mann Whitman test

## Step 5: Calculation of Z

At this point we have everything to calculate Z

$$Z = \frac{U - \frac{n_1 n_2}{2}}{\sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}}$$

We check the formula in step 2

$$Z = \frac{273.5 - \frac{25 \cdot 22}{2}}{\sqrt{\frac{25 \cdot 22 (25 + 22 + 1)}{12}}} = -0.0319$$

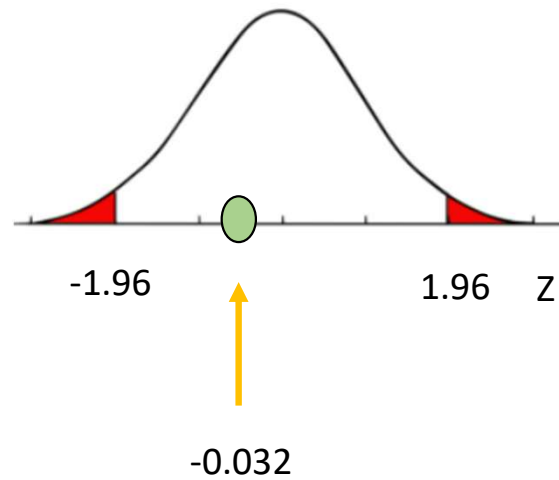
We replace the values in the formula

Z =	<code>= (273.5 - 25*22/2)/SQRT(25*22*(25+22+1)/12)</code>
-----	---

Or we replace the values in Excel

# U-Mann Whitman test

## Step 6: Test Result



We accept  $H_0$ .

# U-Mann Whitman test

## **Step 7:** Interpretation

*According to the test and at a 5% of significance level, we can say that there is no evidence to say that there are differences in the consumer's preferences between the brands*



**THAT'S ALL FOR TODAY**

**THANK YOU**

