-5-

Non – parametric methods

5.1 Introduction

In all situations involving testing hypotheses encountered so far, some assumption had to be made. For example, to apply the *t* – test on small samples, one has to assume that the population is normally distributed. When this assumption is not valid or if we know nothing about the population distribution, the application of the *t* – test will not necessarily yield correct inference about the tested hypothesis.

Also, in some cases we are confronted with non – numerical data or data in the form of ordinal or interval variables which cannot be analyzed using the tests described in the previous chapters, these being only applicable to ratio variables.

5.2 Tests for independent samples

Non – parametric tests are either applied on independent or dependent samples. In the former case, samples are drawn from different populations while in the second there is some dependence between the populations.

For example, when we test the hypothesis that the mean grades of students from two different populations are equal, this represents a case of independent samples. On the other hand if we test the grades of a number of students in two different exams, then there will be some correlation between their grades and the samples can no more considered to be independent.

5.2.1 The sign test

This is one of the earliest tests devised to test hypotheses about the median of a population *M* following sampling. Let the null hypothesis be:

H_o: $M = \xi$

Sample data are given the sign + if they are higher than ξ and – if they are lower. If the statistic value of any specimen is ξ, it is discarded from the sample. Then, the number of + signs and that of – signs is calculated. Let *n* denote the number of + signs. The ratio n / N is then calculated, where *N* is the total number of considered specimens. The proportion corresponding to the null hypothesis is $\pi = 0.5$.

The problem then becomes one of testing a hypothesis regarding a proportion. Such cases were discussed in Chapter 7. For a sample size *n*

≥ 20, a normal distribution can be assumed. Otherwise, a cumulative binomial distribution is used.

Example 5.1

A factory producing resins has purchased a new production line. Resin cubes pipes produced by an old line used to have a median strength of 25 MPa. 14 specimens of the new lines showed the following values for strength:

N°	1	2	3	4	5	6	7	8	9	10	11	12	13	14
σ MPa	28	25	26	27	27	24	26	26	24	23	27	26	24	27

Verify the hypothesis that the newly purchased line produced stronger resins (at 0.05 significance level)

Solution:
The null hypothesis is: H_o: $M = 25$
The alternate hypothesis is: H_1: $M > 25$

Discarding the value 25 from the set we get 9 values with + sign and 4 with – sign.
The null hypothesis is rejected (and the alternate hypothesis accepted) if we have too few negative signs. This occurs if the probability of negative signs is < 0.05.
The corresponding probability of negative sign as calculated by cumulative binomial distribution P($x \leq 4$) with probability 0.5 gives 0.13. For a significance level of 0.05 (or even 0.1), the null hypothesis is accepted, that is no significant change in median value has taken place.
If the number of negative signs was 3, we would have obtained P($x \leq 3$) = 0.0461 and the alternate hypothesis would have been accepted.

5.2.2 The Mann – Whitney test

This test involves the comparison between two mean or median values for two independent populations. The null hypothesis would then be: H_o: $M_1 = M_2$

Let a sample of size n_1 be drawn from the first population and a sample of size n_2 from the second. Their values are placed in ascending order regardless their origin and a rank is assigned to each value. Let R_1 be the sum of the ranks of readings originating from the first sample.

The upper limit of the region of acceptance is then calculated from the following equation:

$$UL = \frac{n_1.(n_1 + n_2 + 1)}{2} + z_{crit} \times \sqrt{\frac{n_1.n_2.(n_1 + n_2 + 1)}{12}} \qquad (5.1)$$

The lower limit is calculated from:

$$LL = n_1 \times (n_1 + n_2 + 1) - UL \qquad (5.2)$$

Where z_{crit} is the critical value corresponding to the specified confidence level for two tailed normal distribution.
If the sum of the ranks of the first set lies between these limits, then H_0 is accepted.

Example 5.2
Nine pieces of flint were collected for a simple experiment, four from region A and five from region B. Hardness was judged by rubbing two pieces of flint together and observing how each was damaged. The one having the least damage was judged harder. Using this method all nine pieces of flint were tested against each other, allowing them to be rank ordered from softest (rank 1) to hardest (rank 9).

Region	A	A	A	B	A	B	B	B	B
Rank	1	2	3	4	5	6	7	8	9

Test whether the flints from the two regions have the same mean.

Solution:

This is a typical case of non – parametric testing since no numerical values were assigned to the statistic under consideration (hardness) but rather relative rank.

The null hypothesis is: H_o: $\mu_1 = \mu_2$

And the alternate hypothesis is: H_1: $\mu_1 \neq \mu_2$

The data are already ranked so that we can directly write:

$n_1 = 4$ and $n_2 = 5$

The value of $R_1 = 1 + 2 + 3 + 5 = 11$

We now calculate the limits of the acceptance region from equations (5.1) and (5.2) for a confidence level = 0.95:

$$UL = \frac{4\times(4+5+1)}{2} + 1.96\times\sqrt{\frac{4\times5\times(4+5+1)}{12}} = 28$$

$$LL = 4\times(4+5+1) - 28 = 12$$

Since $R_1 = 11$ does not lie within the acceptance region, then the null hypothesis is rejected and we can conclude with 95% confidence that the two types of flint don't have the same mean hardness.

Example 5.3

A new board took charge in a factory. One year later, 18 of the factory personnel were asked to express their opinion about the performance of this board by assigning marks from 0 to 5. Out of these 10 were from the engineering and 8 from the financial departments. The following table shows the results:

Eng.	2	1	3	3	4	2	0	4	4	5
Fin.	4	3	2	4	1	1	2	1		

At a 5% significance level, would you consider that the two groups showed the same level of satisfaction towards the performance of the new board?

Solution:

This is a case of ordinal values since the marks assigned cannot be treated quantitatively. Also, in this example appears data having equal ranks, known as **ties.**

The null hypothesis is: H_o: $\mu_1 = \mu_2$

And the alternate hypothesis is: H_1: $\mu_1 \neq \mu_2$

We first rank the data in an ascending way. The shaded values represent the data of the first group:

We have: $n_1 = 10$ and $n_2 = 8$

	0	1	1	1	1	2	2	2	2	3	3	3	4	4	4	4	4	5
Rank	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18

We note that several ranks are repeated. In that case we assign to each the mean values between these ranks. For example, the mean value of the four entries of rank 1 will be: (2+3+4+5) / 4 = 3.5 and so on.
The new table reads:

	0	1	1	1	1	2	2	2	2	3	3	3	4	4	4	4	4	5
Rank	1	3.5	3.5	3.5	3.5	7.5	7.5	7.5	7.5	11	11	11	15	15	15	15	15	18

The value of $R_1 = 1 + 3.5 + 7.5\times2 + 11\times2 + 15\times3 + 18 = 104.5$

We now calculate the limits of the acceptance region from equations (5.1) and (5.2) for a significance level = 0.05:

$$UL = \frac{10\times(10+8+1)}{2} + 1.96\times\sqrt{\frac{10\times8\times(8+10+1)}{12}} = 117$$

$$LL = 10\times(10+8+1) - 117 = 73$$

Since R = 104.5 lies within the acceptance region, then the null hypothesis is accepted and we can conclude with 95% confidence that the two sectors (engineering and financial) have comparable opinions about the performance of the new board.

5.2.3 The Kruskal - Wallis test

This test is an extension of the Mann – Whitney test involving the comparison between several means or median values. If we are in presence of c samples chosen from c different populations then the null hypothesis would be: H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \ldots = \mu_c$
While the alternate hypothesis would be that at least one mean (or median) is different.

This test can also be used for ratio values. In this respect it represents a non – parametric alternative to analysis of variance that does not involve any assumptions about the distribution of population. However this test assumes that all groups tested from which samples were drawn have similar distributions and that the measurement scale is at least ordinal. The following steps are undertaken:

Let the samples be represented by the following table:

Sample 1	**Sample 2**	**Sample 3**	**Sample *i***	**Sample *c***
X_{11}	X_{21}	X_{31}	X_{i1}	X_{c1}
X_{12}	X_{22}	X_{32}	X_{i2}	X_{c2}
X_{13}	X_{23}	X_{33}	X_{i3}	X_{c3}
X_{14}	X_{24}	X_{34}	X_{i4}	X_{c4}
…..	**…..**	**…..**	**…..**	**…..**
X_{1n}	X_{2n}	X_{3n}	X_{4n}	X_{cn}

- Let N be the total number of observations
- Rank all the observations by ascending order and let R be the sum of all ranks. In case of tied values an average value is taken.
- Calculate the variance of ranks from the formula:

$$s^2 = \frac{1}{N-1}.[\sum_{ranks} R_{ij}^2 - \frac{N.(N+1)^2}{4}] \quad (5.3)$$

- The test statistic is:

$$T = \frac{1}{s^2}.\left[\sum_{i=1}^{k} \frac{(\sum_{j=1}^{n} R_{ij})^2}{n_i} - \frac{N.(N+1)^2}{4}\right] \quad (5.4)$$

- The acceptance region is the interval $\left(\chi^2_{1-\alpha}, \chi^2_{\alpha}\right)$ for *d.f.* = $c - 1$

Example 5.4

The GPAs of 24 students from 4 different departments were compared as shown in the following table. At a 0.05 significance level would you accept the hypothesis that there is no appreciable difference between the mean GPAs at the different departments?

A	B	C	D
2.55	3.45	1.55	2.05
1.97	2.87	2.58	1.85
2.87	2.33	3.25	2.54
1.50	3.85	1.85	3.25
2.56	3.05	1.65	3.90
1.65	2.55	2.87	2.85

Solution:

Since the variable investigated is of the interval type, then the problem can be solved in two ways: Either by using ANOVA or by the Kruskal – Wallis method.

The null hypothesis is: H_o: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

1st Solution: Using ANOVA

The following ANOVA table was obtained using the technique explained in Chapter 5.

ANOVA						
Source of Variation	***SS***	***df***	**Variance**	***F***	***F crit***	***p*-value**
Between Groups	2.728846	3	0.909615	**2.126**	**3.098**	**0.129**
Within Groups	8.55775	20	0.427888			
Total	11.286596	23				

Since $F_{calc.}$ = 2.126 < 3.098 ($F_{crit.}$) Then the hypothesis is accepted.

2nd Solution: Using the Kruskal – Wallis method

The following table shows the ranking of the 24 values (including averaging ranks of tied values)

Values	Rank	Values	Rank	Values	Rank	Values	Rank
1.5	1	1.97	7	2.56	13	3.05	19
1.55	2	2.05	8	2.58	14	3.25	20.5
1.65	3.5	2.33	9	2.85	15	3.25	20.5
1.65	3.5	2.54	10	2.87	17	3.45	22
1.85	5.5	2.55	11.5	2.87	17	3.85	23
1.85	5.5	2.55	11.5	2.87	17	3.9	24

From this table the following table assigning ranks for each individual group is obtained.

Values	Rank	Values	Rank	Values	Rank	Values	Rank
A		**B**		**C**		**D**	
2.55	11.5	3.45	22	1.55	2	2.05	8
1.97	7	2.87	17	2.58	14	1.85	5.5
2.87	17	2.33	9	3.25	20.5	2.54	10
1.5	1	3.85	23	1.85	5.5	3.25	20.5
2.56	13	3.05	19	1.65	3.5	3.9	24
1.65	3.5	2.55	11.5	2.87	17	2.85	15
SUM	**53**		**101.5**		**62.5**		**83**

Hence

$R_1^2 = 53^2 = 2809, R_2^2 = 101.5^2 = 10302.25, R_3^2 = 62.5^2 = 3906.25, R_4^2 = 83^2 = 6889$

The following table was computed to show the corresponding values of R^2 for each group.

	Values	R^2	Values	R^2	Values	R^2	Values	R^2
	A		**B**		**C**		**D**	
	2.55	132.25	3.45	484	1.55	4	2.05	64
	1.97	49	2.87	289	2.58	196	1.85	30.25
	2.87	289	2.33	81	3.25	420.25	2.54	100
	1.5	1	3.85	529	1.85	30.25	3.25	420.25
	2.56	169	3.05	361	1.65	12.25	3.9	576
	1.65	12.25	2.55	132.25	2.87	289	2.85	225
SUM		**652.5**		**1876.25**		**951.75**		**1415.5**

Hence the total sum of squares of ranks = R^2 = 652.5 + 1876.25 + 951.75 + 1415.5 = 4896.25

The variance of ranks is therefore:

$$s^2 = \frac{1}{N-1}.[\sum_{ranks} R_{ij}^2 - \frac{N.(N+1)^2}{4}] = \frac{1}{24-1}.[4896.25 - \frac{24\times(24+1)^2}{4}] = 49.84$$

The test statistic is: $T = \frac{1}{s^2}.\left[\sum_{i=1}^{k}\frac{(\sum_{j=1}^{n} R_{ij})^2}{n_i} - \frac{N.(N+1)^2}{4}\right] =$

$$\frac{1}{49.84}\times\left(\frac{2809+10302.25+3906.25+6889}{6} - \frac{24\times(24+1)^2}{4}\right) = 4.7$$

For a number of degrees of freedom = 4 – 1 = 3, $\chi^2_{0.95}$ = 0.35, $\chi^2_{0.05}$ = 7.81. Since 0.35 < 4.7 < 7.81, the **null hypothesis is accepted**.

Example 5.5

An industrial facility produces tiles which are graded as Grade 1, 2, 3 and 4, by decreasing order of quality. The following table shows the grades of a number of tile specimens collected from the 3 production lines.

Verify, at a 0.05 significance level the hypothesis that there are no radical differences between the qualities of tiles produced by each of the three lines.

A	B	C
1	4	1
2	1	3
1	2	4
3	3	4
2	2	2
1	3	1
		2

Solution:

Here the ANOVA method cannot be applied since the variable is ordinal.

So the Kruskal – Wallis method is applied.

The null hypothesis is: H_o: $M_1 = M_2 = M_3$ (Equal median values)

The following table summarizes the results obtained by applying the methods for ranking.

A	Rank	B	Rank	C	Rank
1	3.5	4	18	1	3.5
2	9.5	1	3.5	3	15.5
1	3.5	2	9.5	4	18
3	15.5	3	15.5	4	18
2	9.5	2	9.5	2	9.5
1	3.5	3	15.5	1	3.5
				2	9.5
SUM	**44**		**69.5**		**76.5**

$R_1^2 = 44^2 = 1936, R_2^2 = 69.5^2 = 4830.25, R_3^2 = 76.5^2 = 5852.25$

A	R^2	B	R^2	C	R^2
1	12.25	4	324	1	12.25
2	90.25	1	12.25	3	210.25
1	12.25	2	90.25	4	324
3	210.25	3	210.25	4	324
2	90.25	2	90.25	2	90.25
1	12.25	3	210.25	1	12.25
				2	90.25
SUM	**427.5**		**937.25**		**1063.25**

Hence the total sum of squares of ranks = R^2 = 427.5 + 937.25 + 1063.25 = 2428

The variance of ranks is therefore: $\frac{1}{19-1} \times [2428 - \frac{19 \times (19+1)^2}{4}]$ = 29.33

The test statistic is therefore:

$$\frac{1}{29.33} \times \left(\frac{1936}{6} + \frac{4830.25}{6} + \frac{5852.25}{7} - \frac{19 \times (19+1)^2}{4} \right) = 2.17$$

For a number of degrees of freedom = 3 – 1 = 2, $\chi^2_{0.95}$ = 0.105 and $\chi^2_{0.05}$ = 5.99.

Since 0.105 < 2.17 < 5.99, the **null hypothesis is accepted**

5.3 Tests for dependent samples

5.3.1 Degree of dependence between two samples: Spearman correlation coefficient

One important starting point in dealing with dependent samples is to assess whether the samples are independent or not. One classical way is to use the Spearman rank correlation coefficient (R_s). This correlation coefficient differs from that of Pearson in that it does not necessarily deal with interval or ratio variables. Actually it matches the ranks of corresponding values and gives no inference about the linearity of any relation linking these two variables. As with the Pearson coefficient its value lies between – 1 and + 1. The variable level has to be at least ordinal. The method of its computation is presented in what follows:

- Let *n* be the number of paired observations of the two samples
- Rank these observations, each pair separately, either in ascending or descending way. Use mid values for tied data.
- Get the absolute difference between the ranks of corresponding pairs $|D_i|$
- The Spearman correlation coefficient is then calculated from:

$$R_s = 1 - \frac{6.\sum_{i=1}^{n} D_i^2}{n.(n^2 - 1)} \qquad (5.5)$$

Confidence interval for R_s

The value of R_s as determined for a sample necessitates calculating a corresponding confidence interval for the correlation coefficient of population. The steps are follows:

- Calculate a value of z^* from:

$$z^* = \frac{1}{2} \ln \frac{1 + R_S}{1 - R_S} \qquad (5.6)$$

- Evaluate an error in the value of z from the equation:

$$E(z) = z_{1-\frac{\alpha}{2}} \sqrt{\frac{1+\frac{r^2}{2}}{n-3}} \qquad (5.7)$$

- Obtain lower and upper limits of z* from:

$$z_L = z^* - E(z) \text{ and } \qquad z_U = z^* + E(z) \qquad (5.8)$$

- The confidence interval of ρ_s, the population correlation coefficient is:

$$\tanh z_1 < \rho_s < \tanh z_2 \qquad (5.9)$$

Example 5.6

A research engineer needs to investigate whether there is any relation between the hardness of different alloys and their tensile strength. So he chooses 16 specimens containing different alloying elements and determines their Mohs hardness and their tensile strength (MPa). He obtains the following table.

MHO nº	4	5	3	4	5	6	5	3	6	2
Strength	250	260	240	240	270	270	240	240	270	220

At 0.05 significance level, construct a confidence interval for the coefficient of population.

Solution:

$$R_s = 1 - \frac{6 \times 22}{10 \times (10^2 - 1)} = \mathbf{0.867}$$

From equation (5.6):

$$z^* = \frac{1}{2} \ln \frac{1 + 0.867}{1 - 0.867} = 1.32$$

From equation (5.7), with $z_{1-\frac{\alpha}{2}} = 1.96$

$$E(z) = 1.96 \sqrt{\frac{1+\frac{0.867^2}{2}}{10-3}} = 0.869$$

Moh Hardness	R Moh	Strength	R strength	D^2
4	5.ₒ	250	6	2.25
5	7	260	7	0
3	2.5	240	3.5	1
4	5.ₒ	240	3.5	1
5	7	270	9	4
6	9.5	270	9	0.25
5	7	240	3.5	12.25
3	2.5	240	3.5	1
6	9.5	270	9	0.25
2	1	220	1	0
	TOTAL			**22**

From equation (5.8): $z_L = 1.32 - 0.869 = 0.451$ $z_U = 1.32 + 0.869 = 2.189$

Finally, from equation (5.8):

$\tanh 0.451 < \rho_s < \tanh 2.189$ **$0.423 < \rho_s < 0.975$**

Example 5.7

It is known that the presence of quartz in a mineral ore reduces its grindability. To assess this effect, an engineer with the R&D department of a factory chooses 8 specimens of ore and subjects equal masses to grinding in a laboratory ball mill under the same conditions. He then determines the specific surface of product (cm^2/g) for the 8 specimens and obtains the following results:

%Quartz	12	21	16	18	28	30	10	15
Sp. Area	1260	1120	1250	1230	990	1020	1250	1300

Estimate:

(1) The Pearson correlation coefficient
(2) The Spearman correlation coefficient.

Solution:

The Pearson correlation coefficient can be directly obtained from EXCEL function: Correlation.

***R* = – 0.9265**

On the other hand, the Spearman rank coefficient can be easily calculated as practically no tied values are present.

% *Q*	R (*%Q*)	Area	*R*(area)	|*D*|	D^2
12	2	1260	7	5	25
21	6	1120	3	3	9
16	4	1250	5.5	1.5	2.25
18	5	1230	4	1	1
28	7	990	1	6	36
30	8	1020	2	6	36
10	1	1250	5.5	5.5	20.25
15	3	1300	8	5	25
				SUM	**155.5**

$$R_s = 1 - \frac{6 \times 154.5}{8 \times (8^2 - 1)} = \mathbf{-0.84}$$

The two coefficients are highly negative indicating a possible inverse correlation.

5.3.2 Cramer's test for independence of nominal variables

While Spearman correlation coefficient can be applied to test the degree of correlation between values of at least ordinal scale, it cannot be applied for nominal scale data. In that case the Cramer test is used that relies on constructing contingency tables and calculating the χ^2 value. The Cramer correlation coefficient is then calculated from:

$$\varphi = \sqrt{\frac{\chi^2_{calc}}{n.(k-1)}} \qquad (5.10)$$

Where: n = total sample size
k = smaller number of categories of the two considered variables

A value close to 1 will indicate a strong correlation while a small value will mean poor correlation.

Note that this coefficient is always positive and its interpretation is often subjective.

Example 5.8
As survey was conducted to test whether there was any relation between the number of working hours per day and academic level. The results obtained were as follows for the number of students:

Working h.	Soph.	Junior	Senior 1	Senior 2	TOTAL
X: 0 – 2	25	12	8	2	**45**
Y: 2 – 4	12	19	12	7	**50**
Z: > 4	2	9	23	11	**45**
TOTAL	**39**	**38**	**43**	**20**	**140**

Calculate the Cramer correlation coefficient. What can you conclude?

Solution:

The contingency table is shown below:

	Soph.	**Soph.**	**Junior**	**Junior**	**Senior 1**	**Senior 1**	**Senior 2**	**Senior 2**
	X_{actual}	X_{hyp}	X_{actual}	X_{hyp}	X_{actual}	X_{hyp}	X_{actual}	X_{hyp}
X	25	12.54	10	12.21	8	13.82	2	6.43
Y	12	13.93	19	13.57	12	15.36	7	7.14
Z	2	12.54	9	12.21	23	13.82	11	6.43

The calculated χ^2 value = 40.52

Substituting in equation (5.10) with n = 140 and k = 3, we get: $\varphi = \sqrt{\frac{40.52}{140\times(3-1)}} = \mathbf{0.38}$

The moderately low value of the correlation coefficient infers that there is, to some extent, some relation between the number of studying hours and the academic level.

5.4 Exercise problems

(1) A cement plant can get his clay raw material from two different sources (A) and (B). The chlorine content of 10 specimens of produced cement was determined. The following table shows the results obtained. Use the Mann – Whitney test (at 0.05 significance level) to decide whether there are any differences in the chlorine content in both types of clays.

Zone	A	A	B	A	B	B	B	A	B	A
% Cl	0.08	0.1	0.11	0.12	0.09	0.08	0.12	0.07	0.1	0.09

(2) The following table shows the grades obtained by a sample of 15 students in two subjects X and Y. Calculate the correlation coefficient between the two sets of data and deduce (at 0.05 significance level) the inference on the correlation between these grades at population level.

X	A	C	D	C	C	D	B	B	A	B	B	C	F	D	F
Y	A	B	C	C	B	F	B	A	B	B	C	C	F	F	D

(3) The following table shows the relation between the sulfur content of crudes produced by 8 different wells and their kinematic viscosity (cSt) as determined for specimens collected from each well.

% S	0.3	0.25	0.5	0.22	0.21	0.15	0.28	0.32
υ	215	220	450	180	190	205	350	240

Estimate the Spearman coefficient from these data and deduce the probable limits of the coefficient of population at α = 0.05

(4) The percent solvent in the distillate produced by 4 different distillation units was determined for a sample of size 28 collected from the different units. These were compared as shown in the following table. At a 0.05 significance level would you accept the hypothesis that there is no appreciable difference between the mean percent solvent produced by the different units? Use two different methods of analysis.

A	B	C	D
35	42	48	40
39	50	42	48
37	40	50	44
42	48	39	34
38	39	38	48
32	35	48	45
35	40	42	38

(5) Eight students were selected at random from senior students from 5 departments X, Y, Z, U, V. Their grades were recorded. Estimate using a suitable test whether there are any differences of grades between the different departments (At 0.05 significance level).

X	Y	Z	U	V
B	A	B	D	B
B	A	B	D	B
C	C	A	F	A
A	D	B	B	A
A	F	B	C	D
D	C	C	C	C
C	C	D	A	A
D	A	B	B	C

(6) A survey is carried out on students in a certain academic program to investigate the presence of a possible relation between the grade of a student and his preferred hobby. The results were as follows:

Grade	Sports	Music	Reading	TV series	Other
A	4	5	5	2	1
B	8	8	6	8	7
C	11	9	5	7	5
D	4	3	2	3	6
F	3	2	1	2	3

Use the Cramer method to test whether there is any possible correlation between the two parameters.