12/02/2020

MODULE - 3

NON PARAMETRIC INFERENCE

Till now most of the statistical test procedures what we have disscussed so far had the following two features in common -

(i) The form of the frequency function of the parent population from which the samples have been drawn is assumed to be known.

(ii) They were concerned with testing statistical hypothesis about the parameters of this frequency function or estimating its parameters.

If all the exact (small) sample tests of significance are based on the fundamental assumption that the parent population is normal and are concerned with testing or estimating the means and variences of these populations. Such tests are known as parametric tests.

If a test doesn't depend on the particular form of the basic frequency function from which the samples are drawn then the test procedure is called NON-PARAMETRIC TESTS or METHODS.

In other words N-P test doesn't make any assumptions regarding the form of the population.

However certain assumptions associated with N-P test are

1) Sample observations are independent
2) Variable under study is continuous.
3) Population density function is continuous.
4) Lower order moments exist (general mean & variance)

Obviously these assumptions are fewer and much weaker than those associated with parametric tests.

ADVANTAGES and DRAWBACKS of N-P Methods over Parametric Methods :—

ADVANTAGES	DRAWBACKS
▶ N-P methods are readily comprehensible, very simple and easy to apply and don't require complicated sample theory	▶ N-P test can be used only if the measurements are normal or ordinal. Even in that case if a parametric test exist, it is more powerful

ADVANTAGES	DRAWBACKS
	than the N-P test. In other words if all the assumptions of a statistical model are satisfied by the data and if the measurements are required strength, then the N-P tests are wasteful of time and data
▶ No assumption is made about the form of the frequency function of the parent population from which sampling is done.	▶ So far no N-P methods exists for testing interactions in "analysis of varience" (ANOVA) model unless special assumptions about the addi-tivity of the model are made.
▶ No parametric technique will apply to the data which are measured in normal scale, while N-P methods exists to deal with such data.	▶ N-P tests are designed to test statistical hypothesis only and not for estimating the parameters.
▶ Since the socioeconomic data are not in general, normally distributed, N-P test have found	

ADVANTAGES	DRAWBACKS
applications in psychometry, sociology and educational statistics.	
▶ N-P tests are available to deal with the data which are given in ranks or numerical scores have the strength of the ranks. No parametric test can be applied if the scores are given in grades such as A^+, A, B, B^+ etc.	

The Sign test -

The sign test is the simplest of the non-parametric test.

Its name comes from the fact that it is based on the direction (or signs for '+'s and '-'s) of a pair of observations.

In any problem in which sign test is used, we count,

i) Number of +ve signs
ii) Number of -ve signs.
iii) Number of zeros (i.e., which cannot be included either +ve or -ve)

We take $H_0 : p = 0.5$ (Null Hypothesis).

The difference is chance effects the probability of a positive sign for any particular pair is $1/2$ as the probability of -ve sign.

If "s" is the number of times the less frequency sign occurs then "s" as the binomial distribution with $p = 1/2$.

The critical value for a two-sided alternative $\alpha = 0.05$ can be conveniently found by the expression,

$$k = \frac{(n-1)}{2} - (0.98)\sqrt{n}.$$

Note:—

1) H_0 is rejected if $S \leq k$

2) H_0 is accepted if $S > k$ for the sign test.

Types of Sign test :—

i) One sample sign test

ii) Paired sample sign test.

▶ One Sample Sign Test :—

In a one sample sign test, we test the Null Hypothesis $(H_0): \mu = \mu_0$, against an appropriate alternative on basis of a random sample of size n, we replace each sample value greater than μ_0 with a '+' sign and each sample value less than μ_0 with a '−' sign.

And discard sample value exactly equal to other sample value (put 0).

Then the Null Hypothesis (H_0) is testing for these '+' and '−' signs of a random variable following binomial distribution with $p = 1/2$ or not.

▶ Paired Sample Sign Test :—

This sign test is very important application in problems involving paired data such as data related

to the collection of an accounts receivable before and after a new collection policy.

In these problems each pair of sample value can be replaced with a '+' sign if the first value is greater than the second and with a '-' sign if the first value is smaller than the second, or discard if both are equal.

Then we proceed in the same manner as in One sample Sign test.

Q1) A Typing institute that in a six week intensive course, it can train students to type, on the average atleast 60 words per minute. A random sample of 15 graduates is given a typing test and the median number of words per minute typed by each of these students is given below.

Test the hypothesis that the median typing speed of graduates is atleast 60 words/min.

Students	words per min	words per min - 60
A	81	+
B	76	+
C	53	-

Students	Words per min	Words per min > 60
D	71	+
E	66	+
F	59	−
G	88	+
H	73	+
I	80	+
J	66	+
K	58	−
L	70	+
M	60	0
N	56	−
O	55	−

Sol:- No. of + signs = 9

No. of − signs = 5

No. of zeros = 1

1) Null Hypothesis : H_0 = Medium = 60

2) Alternative Hypothesis : H_1 : Medium > 60

3) Level of Significance : 5%.

4) Test statistics :

$$k = \frac{(n-1)}{2} - (0.98)\sqrt{n} = \frac{(14-1)}{2} - (0.98)\sqrt{14}$$

$= 6.5 - (0.98)(3.8)$

$= 2.83.$

5) Conclusion : Since $S > k$.

We accept H_0.

Note:-

For large samples, generally considered $n > 2.5$ for the sign test, the normal approximation to the binomial may be used, correcting for continuity.

Since $P = 0.5$, for this we have the mean value $= \frac{1}{2}n$. and standard deviation $= \frac{1}{2}\sqrt{n}$.

The actual value of z can be computed using formula,

$$z = \frac{X - nP_0}{\sqrt{nP_0(1-P_0)}}$$

where X is no. of '+' signs.

The value obtained can be compared to the critical value of z which is approximated for the direction of the test.

Q3) The following data related to the daily production of cement (in tons). A large plant for 30 days

11.5	10.0	11.2	10.0	12.3	11.1	10.2	9.6
8.7	9.3	9.3	10.7	11.3	10.4	11.4	12.3
11.4	10.2	11.6	9.5	10.8	11.9	12.4	9.6
10.5	11.6	8.3	9.3	10.4	11.5		

Use sign test to test the Null Hypothesis that if the plants average daily production of cement is 11.2 metric tons. against Alternative hypothesis it is less than 11.2 metric tons at 5% los.

Sol:- Taking '+' signs and '-' signs for the above data,

+	−	0	−	+	−	−	−	−	−
−	−	+	−	+	+	+	−	+	−
−	+	+	−	−	+	−	−	−	+

No. of '+' signs = 11

No. of '−' signs = 18

No. of zeros = 1

$\therefore \delta = 11 = X$

$n = 29$

$P = 1/2 = 0.5$

$$z = \frac{X - np}{\sqrt{npq}} = \frac{11 - 29(1/2)}{\sqrt{29(1/2)(1/2)}} = \frac{-3.5}{2.69}$$

$$= -1.3011$$

1) Null Hypothesis:

The plant average daily production is $\mu = 11.2$ metric tons.

2) Alternative Hypothesis:

Average daily production is $\mu < 11.2$ metric tons (L.T.T)

3) Level of Significance:

for z-table at 5% l.o.s for L.T.T table

value is -1.64

4) Test Statistic :

$$Z = -1.29 \simeq -1.30$$

5) Conclusion :

$\therefore Z_{cal} = -1.29.$

$Z_{tab} = -1.64$

$\therefore Z_{cal} > Z_{tab}$

Hence we reject H_0 Null Hypothesis i.e, plant average production is less than 11.2 metric tons.

Q4) Use the sign test to see if those is a difference between the no of days until the correction of an account receivable before and after a new collection policy. Use at 5% l.o.s.

Before :-	30	28	34	35	40	42	33	38	34	45
	28	27	25	41	36					
After :-	32	29	33	32	37	43	40	41	37	44
	27	33	30	38	36					

Sol:- Taking '+' signs and '-' signs for the above data,

$- \quad - \quad + \quad + \quad + \quad - \quad - \quad - \quad - \quad +$

$+ \quad - \quad - \quad + \quad 0$

No. of '+' signs = 6

No of '−' signs = 8

No of zeros = 1

$\therefore S = 6$

$\therefore k = \frac{(n-1)}{2} - (0.98)\sqrt{n}$ $= \frac{14-1}{2} - (0.98)\sqrt{14}$

$= \frac{13}{2} - (0.98)(4)$

$= 2.83.$

$\therefore$ Since $S > k$ then the null hypothesis is accepted i.e., there is no significant difference before and after the new correction policy in the accounts receivable.

MANN - WHITNEY U - TEST

(OR)

WILCOXON SIGN RANK SUM TEST

Rank sum test is a whole family of tests, here we shall discuss only one of type i.e, Mann - Whitney U-test. With this test we can test the Null Hypothesis $\mu_1 = \mu_2$ without assuming whether the population samples have roughly have the shape of normal distribution.

The test of the Null Hypothesis that these two samples come from identical populations may either be based on R_1 (sum of the sample 1 ranks) and R_2 (sum of the sample 2 ranks).

It may be noted that in practise it doesn't matter which sample we call sample 1 and which sample we call as sample 2.

The sample sizes are n_1, n_2 & the sum of R_1 and R_2 is simply the sum of n_1 and n_2 positive integers which is known do be $\frac{(n_1+n_2)(n_1+n_2+1)}{2}$. This formula enables us to find R_2 if we know R_1 & to find R_1 if we know R_2.

It is an alternative test of two samples t-test.

(10)

Decision was based on R_1 and R_2. Now the decision is usually based on either of the related statistics i.e.,

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$$

Note:-

1) When n_1 and n_2 are the size of the samples and R_1, R_2 are the Rank sums of the corresponding samples for small samples if both n_1 and n_2 are less than 10 (some statisticians can consider 8). Special tables must be used.

2) If U is smaller than the critical value, H_0 can be related to a standard normal curve by the statistic,

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

Q-1) Use a mann whitney U-test for the following 2 groups to determine whether there was a difference in their course of the 2 groups. Use the 5% l.o.s.

Suitability Scores

Group - A	Group - B
7	8
11	9
9	13
4	14
8	11
6	10
12	12
11	14
9	13
10	9
11	10
11	8

Sol:-

4	6	7	8	8	8	9	9	9	9	10	10	10
1	2	3	5	5	5	8.5	8.5	8.5	8.5	12	12	12
A	A	A	A	B	B	A	A	B	B	A	B	B

11	11	11	11	11	12	12	13	13	14	14	
16	16	16	16	16	19.5	19.5	21.5	21.5	23.5	23.5	
A	A	A	A	B	4	B	B	B	B	B	

R_1 → Rank sum of A = 123.5

R_2 → Rank sum of B = 176.5

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 98.5$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 = 45.5$$

$\therefore U$ = minimum of $\{U_1, U_2\}$ = min $\{98.5, 45.5\}$

$\therefore$ min value = 45.5

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

$$= \frac{45.5 - \frac{(12)^2}{2}}{\sqrt{\frac{(12)(12)(12+12-1)}{12}}}$$

$$= -1.59$$

At 5% l.o.s significance for T.T.T, Z - critical value is 1.96.

$\therefore$ Z_{cal} is -1.56 i.e., $|Z| = 1.56$.

$\therefore$ Z_{tab} value is 1.96, hence $Z_{cal} < Z_{tab}$ value

$\therefore$ We accept H_0.

i.e., There is no significant difference in the scores of two groups.

Q2) Test the difference between following two groups using Mann Whitney U-test.

Group - 1	Group - 2
24	28
18	42
45	63
57	57
12	90
30	68

Soli-

12	18	24	28	30	42	45	57	57
A	A	A	B	A	B	A	A	B
1	2	3	4	5	6	7	8.5	8.5

63	68	90
B	B	B
10	11	12

R_1 → Rank sum of A = 26.5

R_2 → Rank sum of B = 51.5

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 = 30.5$$

$$= 30.5$$

$$U_2 = n_1 n_2 + \frac{(n_2(n_2+1))}{2} - R_1 = 5.5$$

$\therefore$ minimum value = 5.5

If $n_1 = 6$, $n_2 = 6$ the table values for Wilcoxon rank sum test is 26.

The calculated value is 5.5

$\therefore$ Calculated value < Tab value.

$\therefore$ We accept H_0

ONE SAMPLE RUN TEST :-

The various test procedures discussed above a non-parametric test. In many situations it is difficult to decide this assumption is suitable for particular NP tests.

Eg:- If you want to predict the sales of a readymade garments store for a given we have no choice to sales data from previous years and take the stock of the economic condition in general.

For the above problem the no of methods developed for judging the randomness of the sample on the basis of the order on which the observations are taken.

It is possible to determine the sample that took suspiciously non-random may attributed to chance. The technique discussed below is based on the theory of "Runs".

A Run is a success of identical letters (or other kind of symbols) which is followed or proceeded by different letter or no letters at all.

Eg:- A A A B B B B

K K N N N

In the above example we have 6 runs of varying different lengths. The no of runs denoted by 'r', is a statistic with its own special sample distribution and its own test.

To derive the mean of the sampling distribution of 'r' statistic the following test statistic is used,

$$M = \frac{2n_1n_2}{n_1+n_2} + 1$$

where M = mean of 'r' statistic

The standard error of 'r' statistis is calculated by the formula,

$$\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

It may be noted that sampling distribution of 'r' can be closely approximated by the normal distribution if either n_1 or n_2 is larger than 20

then the test statistic,

$$Z = \frac{r - \mu_r}{\sigma_r}$$

NOTE :-

If the calculated value is less than table value, then we accept H_0

Q.1) The following is an arrangement of 25 men (M) and 15 women (W) lined up to purchase tickets for a premier picture show.

M WW MMM W M,M W M W M WWW MMM W MM WWW MMMMMMM

Test the Randomness at 5% l.o.s.

Sol:- Here $n_1 = 25$

$n_2 = 15$

$r = 17$

$$\mu = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2(25)(15)}{25+15} + 1 = 19.75$$

$$\sigma = \sqrt{\frac{2n_1n_2\,(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2\,(n_1+n_2-1)}}$$

$$= \sqrt{\frac{2(25)(15)\left((2)(25)(15) - 25 - 15\right)}{(25+15)^2\,(25+15-1)}}$$

$$= \sqrt{\frac{750\,(750-40)}{(40)^2\,(40-1)}}$$

$$= 2.921$$

$$z = \frac{r-\mu}{\sigma} = \frac{17 - 19.75}{2.921} = -0.94$$

$$|z| = 0.94$$

Since the value is less than 1.96 (1.e. 1.0.s)

$\therefore z_{cal} < z_{tab}$

Hence we accept H_0, i.e., no real evidence to suggest that the arrangement is not random.

Q2) The following is the arrangement of defective (d) and no defective (n) peices produced in a given order by a certain machine.

nnnnn dddd nnnnnnnnnnn dd nn dddd.

Jest the randomness at 1% l.o.s.

Sol:- Here $n_1 = 17$

$n_2 = 10$ $\quad r = 6$

$$\mu = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2(17)(10)}{17+10} + 1$$

$$= \frac{340}{27} + 1 = 13.59$$

$$\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

$$= \sqrt{\frac{2(17)(10)(2(17)(10) - 17 - 10)}{(17+10)^2(17+10-1)}}$$

$$\sigma = 2.376$$

$$\therefore z = \frac{r - \mu}{\sigma} = \frac{6 - 13.59}{2.376}$$

$$= -3.2$$

$$|z| = 3.2$$

$$\therefore z_{cal} = 3.2$$

$$z_{tab} = 2.58$$

$$\therefore z_{cal} > z_{tab}$$

Hence we reject the H_0 Null Hypothesis.

Q3) An engineer is concerned about the possibility that too many changes are being made in the setting of an automatic machine. Given the following mean diameters (in inches) of 40

successive outputs termed on the machine.

0.261	0.258	0.249	0.251	0.247
0.256	0.250	0.247	0.255	0.243
0.252	0.250	0.253	0.247	0.251
0.243	0.258	0.251	0.245	0.250
0.248	0.252	0.254	0.250	0.247
0.253	0.251	0.246	0.249	0.252
0.247	0.250	0.253	0.247	0.249
0.253	0.246	0.251	0.249	0.253

Use 1% los to test the null hypothesis of randomness against the alternative that there is a frequently alternating pattern.

Sol:-

1) Null Hypothesis (H_0): Arrangement of sample value are in random

2) Alternative Hypothesis (H_1): Arrangement of sample values are not in random.

3) Level of Significance: At 1% los for T.T.T, Z_{tab} value is 2.56.

4) Test statistic: $Z = \frac{r - \mu}{\sigma}$

r = no. of runs

$$\mu = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

In the above data the median of the 40 arrangements is 0.250

Median = 0.250 = $\left(\frac{N}{2}\right)^{th}$ = 20th observation

∴ for $> 0.250 = a$

$< 0.250 = b$

a	a	b	a	b	a	b	b	a	b
a	b	a	b	a	b	a	a	b	b
b	a	a	b	b	a	a	b	b	a
b	b	a	b	b	a	b	a	b	a

Here $n_1 = 19$

$n_2 = 21$

$r = 29$

$$\mu = \frac{2n_1n_2}{n_1+n_2} + 1 = \frac{2(19)(21)}{19+21} + 1 = \frac{798}{40} + 1$$

$$= 20.95$$

$$\sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

$$= \sqrt{\frac{2(19)(21)\,(\,2(19)(21) - 19 - 21)}{(19+21)^2\,(19+21-1)}}$$

$$= \sqrt{\frac{798\,(798-40)}{1600\,(40-1)}}$$

$$= 3.11$$

$$\therefore Z = \frac{r - \mu}{\sigma} = \frac{29 - 20.95}{\left(\frac{3.11}{2}\right)}$$

$$= 2.588$$

$$\therefore Z_{cal} = 2.588$$

$$Z_{tab} = 2.58$$

$$\therefore Z_{cal} > Z_{tab}$$

∴ We reject H_0.

NOTE :-

1) For dividing different sample sizes we can consider the three mathematical averages like mean, median and mode, preferably median.

2) If any value is equivalent to the above sited mathematical averages then remove the corresponding data.

The Kruskal - Wallis Test :-

If several independent samples are involved, ANOVA is the usual procedure. If ANOVA is failure to meet the assumptions needed for analysis then the alternative technique is to be used called as "Kruskal - Wallis" test or "H-test".

This test helps in testing the null Hypothesis that K independent random samples come from identical populations against the alternative hypo-thesis that the means of these samples are not equal.

The H - statistic is calculated from the formula,

$$H = \frac{12}{N(N+1)} \left[\frac{R_1^2}{n_1} + \frac{R_2^2}{n_2} + \dots + \frac{R_k^2}{n_k} \right] - 3(N+1)$$

where $n_1, n_2, \dots n_k$ are the number in each of K samples

$$N = n_1 + n_2 + \dots + n_k$$

$R_1, R_2, \dots R_k$ are the rank sums of each sample

If there is any tie occurs the usual procedure is followed.

NOTE :-

1) For small samples H is approximately distributed as χ^2 with $(k-1)$ d.f.

Q1) A companies trainees are randomly assigned to groups which are taught a certain industrial inspection procedure by 3 different methods. At the end of the inspection period they are tested for inspection performance quality. The following are the scores.

Method - A :	80	83	79	85	90	68	
Method - B :	82	84	60	72	86	67	91
Method - C :	93	65	77	78	88		

Use the H-test to determine a 5% l.o.s whether the 3 methods are equally effective.

Sol:- Here, $n_1 = 6$

$n_2 = 7$

$n_3 = 5$

Values	Ranks	Ranks of Method A	Ranks of Method B	Ranks of Method C
60	1		1	
65	2			2
64	3		3	
68	4	4		
72	5		5	
77	6			6
78	7			
79	8	8		
80	9			
82	10		10	
83	11			
84	12		12	
85	13	13		
86	14		14	
88	15			15
90	16	16		
91	17		17	
93	18			18

R_1 = Sum of Ranks of sample A = 61

R_2 = Sum of Ranks of sample B = 62

R_3 = Sum of Ranks of sample C = 48

1) Null Hypothesis (H_0): Three methods are equally effective

2) Alternative Hypothesis (H_1): Three methods are not equally effective.

3) Level of significance: At 5% los, for χ^2 table value with $(k-1)$ d.f $= (3-1)$ df
$= 2$ df i.e, 5.991

4) Test statistic:

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

$$= \frac{12}{18(18+1)}\left[\frac{61^2}{6} + \frac{62^2}{7} + \frac{48^2}{5}\right] - 3(18+1)$$

$$= \frac{12}{342}\left[\frac{3721}{6} + \frac{3844}{7} + \frac{2304}{5}\right] - 57$$

$$= 0.035\,[620.16 + 549.14 + 460.8] - 57$$

$$= 0.196$$

5) Conclusion: $\chi^2_{tab} = 5.991$

$\chi^2_{cal} = 0.196$

$\therefore \chi^2_{cal} < \chi^2_{tab}$

$\therefore$ Hence we accept H_0.

Q2) An experiment designed to compare 3 methods for preventing the depth of the pits (in thousands of an inches), the following maximum depth of pits in pieces of wire subjected w.r.t treatments

Method - A	–	77	54	67	74	71	66
Method - B	–	60	41	59	65	62	64	52
Method - C	–	49	52	69	47	56		

Use 5% los to test the null Hypothesis that the 3 samples come from identical populations.

Sol:-

1) Null Hypothesis : Populations are identical

2) Alternative hypothesis : populations are not identical

3) Level of significance : at 5% los kruskal wallis test approximately χ^2 distributed with

$(k-1)$ df $= (3-1)$ df

$= 2$df

χ^2 tab value is 5.99

4) Test statistic :

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

$n_1 = 6$; $n_2 = 7$; $n_3 = 5$ $\therefore N = 18$

Values	Ranks	Rank of Sample A	Rank of Sample B	Rank of Sample C
41	1		1	
47	2			2
49	3			3
52	4.5		4.5	
52	4.5			4.5
54	6	6		
56	7			7
59	8		8	
60	9		9	
62	10		10	
64	11		11	
65	12		12	
66	13	13		
67	14	14		
69	15			15
71	16	16		
74	17	17		
77	18	18		

$R_1 = 84$

$R_2 = 55.5$

$R_3 = 31.5$

$$\therefore H = \frac{12}{18(18+1)}\left[\frac{84^2}{6} + \frac{55.5^2}{7} + \frac{31.5^2}{5}\right] - 3(18+1)$$

$$= \frac{12}{342}\left[1176 + 4140.03 + 198.45\right] - 57$$

$$= 0.035\ [1814.48] - 57$$

$$H = 6.507$$

5) Conclusion:- $H_{cal} = 6.507$

$H_{tab} = 5.99$

$\therefore\ H_{cal} > H_{tab}$

Hence we reject the H_0

Spearman's Rank Correlation :-

Suppose $(x_1, x_2, \ldots x_n)$ and $(y_1, y_2, \ldots y_n)$ are ranks of two variables x and y in the order of merit with respect to some property.

The correlation between n-pairs of ranks is called the rank correlation.

Then

$$\rho = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

where $n \rightarrow$ no. of samples

$d_i \rightarrow$ deviation $= x_i - y_i$

For Repeated Ranks,

If two or more individuals are repeated, we have to add the correction factor $\frac{M(M^2-1)}{12}$ to Σd_i^2 where M is the no. of times an item is repeated value with both the x-series as well as y-series

NOTE :-

If the same value is there for m values, average rank should be alloted or while alloting the rank.

Eg:-

If two persons get same rank, the rank should be 8 and 9 but we consider the average of 8 and 9 ranks i.e., $\frac{8+9}{2} = 8.5$, should be allotted to each.

If three values are same, the respective ranks can be 7, 8, 9, then the average is 8, should be allotted to each.

Q1) Find the Rank correlation co-efficient to the following data.

x	y	Rank of x	Rank of y	$d_i = x_i - y_i$	d_i^2
10	30	10	10	0	0
15	42	6	3	3	9
12	45	9	2	7	49
17	46	4	1	3	9
13	33	8	9	-1	1
16	34	5	8	-3	9
24	40	1	4	-3	9
14	35	7	7	0	0
22	39	2	5	-3	9
20	38	3	6	-3	9

$\Sigma d_i^2 = 104$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6[104]}{10(10^2-1)}$$

$$= 1 - 0.63$$

$$= 0.36$$

1) Null Hypothesis:- There is no significant between x and y

2) Alternative Hypothesis:- There is significant between x and y

3) Level of significance:-

If $n = 10$ or $n < 10$, it follows Spearman's rank correlation table with n d.f = 10 d.f i.e., 0.648.

4) Test statistic:-

For $n \leq 10$ the test statistic will become

$$\rho = r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$= 0.36$$

5) Conclusion:-

$r_{s\,cal} = 0.36$

$r_{s\,tab} = 0.648$

$\therefore r_{s\,tab} > r_{s\,cal}$

$\therefore$ Hence we accept H_0

Q2) Obtain the rank correlation co-efficient for the following data.

x	y	Rank of x	Rank of y	d_i	d_i^2
68	62	4	5	-1	1
64	58	6	7	-1	1
75	68	2.5	3.5	-1	1
50	45	9	10	-1	1
64	81	6	1	5	25
80	60	1	6	-5	25
75	68	2.5	3.5	-1	1
40	48	10	9	1	1
55	50	8	8	0	0
64	70	6	2	4	16

$\Sigma d_i^2 = 72$

$$\rho = 1 - \frac{6\left[\Sigma d_i^2 + \frac{M(M^2-1)}{12} + \cdots + \frac{M(M^2-1)}{12}\right]}{n(n^2-1)}$$

$$C.F\ (75) = \frac{2(2^2-1)}{12} = 0.5$$

$$C.F\ (64) = \frac{3(3^2-1)}{12} = 2$$

$$C.F\ (68) = \frac{2(2^2-1)}{12} = 0.5$$

$$\rho = 1 - \frac{6[72 + 0.5 + 2 + 0.5]}{10(10^2-1)} = 1 - \frac{6(75)}{990}$$

$= 1 - 0.454$

$= 0.546$

1) Null Hypothesis : H_0 : There is no significant difference between x and y

2) Alternative Hypothesis = H_1 : There is significant difference between x and y

3) Level of significance =

If $n = 10$, or $n < 10$, it follows Spearman's rank correlation table with n.d.f = 10 d.f i.e., 0.648

4) Test statistic =

For $n \leq 10$, the test statistic will become,

$$\rho = r_s = 1 - \frac{6 \Sigma d_i^2}{n(n^2-1)}$$

$$= 0.546$$

5) Conclusion -

$$r_{s\,cal} = 0.546$$

$$r_{s\,tab} = 0.648$$

$\therefore r_{s\,cal} < r_{s\,tab}$

$\therefore$ Hence we accept H_0

Q3) A sample of fathers and their 12 daughters gave the following data about their heights in inches.

father height (X)	daughter height (Y)	Rank of x	Rank of y	d_i	d_i^2
65	68	9	5.5	3.5	12.25
63	66	11	9.5	1.5	2.25
67	68	6.5	5.5	1	1
64	65	10	11.5	-1.5	2.25
68	69	4.5	3	1.5	2.25
62	66	12	9.5	2.5	6.25
70	68	2	5.5	-3.5	12.25
66	65	8	11.5	-3.5	12.25
68	71	4.5	1	3.5	12.25
67	67	6.5	8	-1.5	2.25
69	68	3	5.5	-2.5	6.25
71	70	1	2	-1	1

$\Sigma d_i^2 = 72.5$

In the above table $n = 12$ ($\because n > 10$) we can follow z-test statistic about the spearman's rank correlation co-efficient i.e.,

$$z = \frac{6R - n(n^2-1)}{n(n+1)\sqrt{n-1}}$$

where $R = \Sigma d_i^2$

n = no. of samples

1) Null Hypothesis :- H_0 - There is no significant difference between father height and daughter height.

2) Alternative Hypothesis :- H_1 : There is significant difference between father height and daughter height.

3) Level of significance :- at 5% los for T.T.T, Z critical value is 1.96.

4) Test statistic -

$$Z = \frac{6R - n(n^2-1)}{n(n+1)\sqrt{n-1}} = \frac{6(72.5) - 12(143)}{12(12+1)\sqrt{12-1}}$$

$$= -2.47$$

$$|Z| = 2.47$$

5) Conclusion - $Z_{cal} = 2.47$

$Z_{tab} = 1.96$

$Z_{cal} > Z_{tab}$,

Hence we reject H_0.

There is a significant difference between fathers heights and daughters height.

TOLERANCE LIMITS -

Let L_1, L_2, be two functions of sample values with $L_1 < L_2$, such that the random interval (L_1, L_2) has a probability of β of containing 100% (γ) of the population then,

$$P\left[\int_{L_1}^{L_2} f(x)\,dx \geq \gamma\right] \cdot P\left[F(L_2) - F(L_1) \geq \gamma\right] = \beta$$

This interval (L_1, L_2) is called a 100% (γ) tolerance interval with probability β. The function L_1 and L_2 are called the lower and upper tolerance limits.

For general statistics L_1 and L_2, the probability β depends on the form of the distribution $F(x)$ hence, generally tolerance limits are not distribution free.

The tolerance limits can be used to analyse mass production of articles, the quality of articles as well as to check the specifications on the respective labels in a quality control department.

TOLERANCE LIMITS -

Let L_1, L_2, be two functions of sample values with $L_1 < L_2$, such that the random interval (L_1, L_2) has a probability of β of containing 100% (γ) of the population then,

$$P\left[\int_{L_1}^{L_2} f(x)\,dx \geq \gamma\right] = P\left[F(L_2) - F(L_1) \geq \gamma\right] = \beta$$

This interval (L_1, L_2) is called a 100% (γ) tolerance interval with probability β. The function L_1 and L_2 are called the lower and upper tolerance limits.

For general statistics L_1 and L_2, the probability β depends on the form of the distribution $F(x)$ hence, generally tolerance limits are not distribution free.

The tolerance limits can be used to analyse mass production of articles, the quality of articles as well as to check the specifications on the respective labels in a quality control department.

Kolmogorov Smirnov Test

It is a simple non-parametric method for testing whether there is a significant difference between an observed frequency distribution and a theoretical frequency distribution.

In KS-one sample test is more powerful than the χ^2 test. since it can be used for small samples unlike χ^2 test.

The null hypothesis assumes no difference between the observed and theoretical distribution and the value of test statistic 'D' is calculated.

Procedure

1) Arrange the given number in ascending order.

2) Compute $D^+ = \max_{1 \le i \le N} \left\{ \frac{i}{N} - R_i \right\}$

$D^- = \max_{1 \le i \le N} \left\{ R_i - \left(\frac{i-1}{N} \right) \right\}$

where N = sample size or degrees of freedom (or) total no. of random numbers.

R_i = The observed frequency probabilities as in ascending order

3) Compute $D = \text{Max}[D^+, D^-]$

4) Compute the theoretical value D_α for a given L.O.S α from standard Kolmogoron - Smirnov's table.

5) If $D < D_\alpha$ then accept the hypothesis otherwise reject the hypothesis.

Q1) Using KS test, check for the property of uniformity for the input set of random numbers i.e.,

0.54, 0.73, 0.98, 0.11, 0.68.

Sol:- Let

i	1	2	3	4	5
R(i)	0.11	0.54	0.68	0.73	0.98

$$D^+ = \max_{1 \le i \le N} \left\{ \frac{i}{N} - R_i \right\}$$

$$= \text{Max}\left[\frac{1}{5} - 0.11, \frac{2}{5} - 0.54, \frac{3}{5} - 0.68, \frac{4}{5} - 0.73, \frac{5}{5} - 0.98 \right]$$

$$= \text{Max}[\boxed{0.09}, -, -, 0.07, 0.02]$$

$$D^- = \max_{1 \le i \le N} \left\{ R_i - \left(\frac{i-1}{N} \right) \right\}$$

$= \text{Max}\{0.11-0,\ 0.54-0.2,\ 0.68-0.4,$
$0.73-0.6,\ 0.98-0.8\}$

$= \text{Max}\{0.11,\ \boxed{0.34},\ 0.28,\ 0.13,\ 0.18\}$

$D = \text{Max}\{D^+, D^-\}$

$= \text{Max}[0.09, 0.34]$

$\boxed{D = 0.34}$

Now we have to verify the critical value of D_α at 5% in KS standard table where $N = 5$.

By using KS table at 5% los with $Ndf = 5\ df$, table value is 0.56328 i.e., D_{tab} value is 0.56328

D_{cal} value is 0.34

Conclusion :-

$\therefore D_{cal} < D_{tab}$ i.e., $0.34 < 0.56328$

$\therefore$ Hence we accept H_0.

Q2) The sequence of the random numbers, 0.63, 0.49, 0.24, 0.57, 0.71 and 0.89 has been generated. the KS test with $\alpha = 0.05$

to determine if the hypothesis that the numbers are uniformly distributed on the interval (0,1) can be rejected.

Sol:- Let

i	1	2	3	4	5	6
R(i)	0.24	0.49	0.57	0.63	0.71	0.89

$$D^+ = \max_{1 \le i \le N} \left\{ \frac{i}{N} - R_i \right\}$$

$$= \max \left\{ \frac{1}{6} - 0.24, \frac{2}{6} - 0.49, \frac{3}{6} - 0.57, \frac{4}{6} - 0.63, \frac{5}{6} - 0.71, \frac{6}{6} - 0.89 \right\}$$

$$= \max [0.16 - 0.24, 0.33 - 0.49, 0.5 - 0.57, 0.66 - 0.63, 0.83 - 0.71, 1 - 0.89]$$

$$= \max \{ -, -, -, 0.03, \boxed{0.12}, 0.11 \}$$

$$D^- = \max_{1 \le i \le N} \left\{ R_i - \left(\frac{i-1}{N} \right) \right\}$$

$$= \max [0.24, \boxed{0.32}, 0.236, 0.13, 0.043, 0.056]$$

$$D = \max [D^+, D^-]$$

$$= \max [0.12, 0.32]$$

$$D = 0.32.$$

By using Ks table at 5% l.o.s with n df = 6 d

① table value is 0.5

$\therefore D_{cal} = 0.32$

Conclusion:-

$\therefore D_{cal} < D_{tab}$

i.e., $0.32 < 0.5$

$\therefore$ Hence we accept H_0.