(1)

Unit-IV

-: Non-Parametric Tests :-

Introduction:

So far we have studied the statistical tests based on the form of the frequency function of the parent population from which the samples have been drawn is assumed to be known and testing is concerned about the population parameters.

For example, almost all large and sample tests of significance are based on the fundamental assumption that the parent population is normal and they are concerned with testing or estimating the parameters (means, variances, etc).

→ The statistical tests which are concerned about the population parameters are known as parametric tests.

Also we have conducted some of the tests also donot depend on any form of frequency function such as chi-square tests of goodness of fit. Such type of tests provide a logical basis of conducting the statistical tests donot depend any form of frequency function.

Non-Parametric Tests: Defination :-

A Non-Parametric test is a test that doesnot depend on the particular form of basic frequency function from which samples are drawn and donot related with estimating the parameters.

A Non-parametric test (NP-test) doesnot make any assumption regarding the form of population.

Assumptions of Non-Parametric Tests:-

Assumptions associated with Non-Parametric tests are

i) The sample observations are independent.
ii) The variable under study is continuous.
iii) Probability generating function is continuous.
iv) Lower order moments exists.

Advantages of Non-Parametric tests:-

1. Non-Parametric methods are very simple and easy to apply
2. Non-Parametric methods dont require much complicated sampling theory.
3. Non assumption is made about the form of the frequency function of the parent population from which the samples have drawn
4. Generally, the socio-economic data are not normally distributed, Non-parametric methods have found applications in psychometry, sociology and educational statistics.
5. If the data is given in ranks or signs or grades, then Non-Parametric methods are more usefull and easy to apply. ~~But~~ ~~put~~
6. No parametric technique will apply to the data which are measured in nominal scale, while Non-Parametric methods exists to deal with such type of data.

Scanned with CamScanner

(2) **Disadvantages of Non-Parametric tests:-**

1. If the measurements are nomial or ordinal we can apply NP methods. Even in this case, If a parametric tests exist it is more powerfull than the NP-test.

2. NP-tests are designed to test statistical hypothesis only and not for estimating parameters.

3. So far, No NP-methods exist for testing interactions in "Anova" unless special assumptions about the additivity of the model are made.

Parametric tests Vs Non-Parametric tests:-

Parametric tests	Non-Parametric tests.
1. The sample observation are independent	1. The sample observations are independent.
2. The variable in the study need not be continous (may be discrete or continous)	2. The variable in the study continous.
3. It depends on form of frequency function from which the samples have drawn	3. It doesnot depends on any particular form of frequency function from which the samples are drawn.
4. Assumptions taken about a parameters	4. No assumptions about population parameters.
5. It is based on the assumption that the parent population is assumed to be known.	5. Non-Parameteric tests do not have any assumption regarding the parent population

Scanned with CamScanner

Parametric tests	Non-Parametric tests
6. In this method, we can estimate the population parameters which are not known	6. In this method we can not estimate the population parameter.
7. This method cannot be applied for the data is in ranks and grades	7. This method can be applied for the data of ranks and grades.
8. If the measurements are nominal or ordinal, we cannot apply parametric tests	8. If the measurements are nomial or ordinal we can apply non-parametric tests
9. The methods are more efficient or greatest power.	9. The methods are less efficient or less power.
10. These methods are strong	10. These methods are weak.
11. In this method we can draw more conclusions	11. In this method, we cannot draw more conclusions.

→ In Non-Parametric tests are different in one sample observations and two sample observations.

1. Sign test for one sample
2. Sign test for two samples
3. Wald-wolfowitz Runtest for one sample
4. wald-wolfowitz run test for two samples
5. wilcoxon signed rank test for two samples
6. Mann-whitney U-test for two samples
7. Median-test for two samples.

Scanned with CamScanner

(3) * **Sign Test:** Sign test is the simplest of all the Non-Parametric Tests. As the name suggests, it is based on signs (plus or minus) of the deviations rather than the exact magnitude of the variable values.

Types of Sign tests:

1. single Sample sign test
2. paired Sample sign test.

Sign test for one Sample :-

One Sample sign test is used to the test hypothesis. ~~Concering~~ Concerning the median for the population. Suppose we want to test the ~~median~~ ~~the~~ hypothesis that the median (Md) of a population has a specified value say M_0.

Let $x_1, x_2 \dots x_n$ be a random sample drawn from a population of size 'n' and let M be median of the given distribution or for the given observation, M_0 say.

Subtract the value M_0 from each of the sample observations ie; $x_1 - M_0$, $x_2 - M_0$, ... $x_n - M_0$.

If we observe the signs of the above expersion, we get some positive signs, some negative signs and some values may be zero. It is written as

i) The plus sign (+) : If the deviation is positive
ii) The minus sign (−) : If the deviation is negative
iii) zero (0) : If the deviation is zero.

Scanned with CamScanner

Null Hypothesis : H_0 : $M_d = M_0$

The Null hypothesis is written as

$\rightarrow H_0 : P(X > M_0) = P(X < M_0) = \frac{1}{2}$.

Alternative Hypothesis : $H_1 : M_d \neq M_0$ (T.T.T)

(or) $H_1 : M_d > M_0$ (R.T.T)

(or) $H_1 : M_d < M_0$ (L.T.T)

The Level of significance is fixed in advance . i.e $\alpha = 5\%$ (or) 1%.

Test statistic : Let $x_1, x_2 \ldots x_n$ be the observations on a random sample of size 'n' from the given population.

find the $x_1 - M_0$, $x_2 - M_0 \ldots x_n - M_0$ deviations.

If any deviation zero, it is eliminating from the observations.

n = number of +ve signs and -ve signs

Let $u = \min$(number of +ve signs, number of -ve signs)

If we consider getting a sign is success. The 'u' obviously follows binomial distribution with parameters $n, \frac{1}{2}$.

The probability mass function is $P(X=x) = {}^nC_x \, p^x \, q^{n-x}$; $x = 0, 1, 2 \ldots n$

$$\Rightarrow P(X=u) = {}^nC_u \, p^u \cdot q^{n-u}$$

$$\Rightarrow P(X=u) = {}^nC_u (\tfrac{1}{2})^u (\tfrac{1}{2})^{n-u}$$

$$\Rightarrow P(X=u) = {}^nC_u \cdot (\tfrac{1}{2})^n \quad ; u = 0, 1, 2, \ldots n$$

By using cumulative probability

$$P(X \leq u) = \sum_{r=0}^{u} ({}^nC_r)(\tfrac{1}{2})^n = p \text{ (say)}$$

Scanned with CamScanner

Conclusion: If $P' \geq \alpha$ (Level of significance), then we accept H_0. otherwise we reject H_0.

<u>If large samples</u>: If sample size 'n' is large (ie $n \geq 25$) then 'u' is follows to normal distribution with

$$\text{mean} = E(u) = n/2 \qquad \left[\because \begin{array}{l} E(u) = np = n \cdot \tfrac{1}{2} = n/2 \\ V(u) = npq = n \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = n/4 \end{array}\right]$$

$$\text{and variance} = V(u) = n/4$$

$$Z = \frac{u - E(u)}{S.E(u)} \sim N(0,1)$$

$$Z = \frac{u - n/2}{\sqrt{n/4}} \sim N(0,1)$$

If $Z_{cal} \leq Z_{tab}$ at given level of significance $\alpha\%$, then we accept H_0 otherwise we reject H_0.

* <u>Sign test for two sample (or) Paired Samples sign test</u>:-

Let (x_i, y_i); $i = 1, 2, \ldots n$ be a random sample of the 'n'-paired observations, the pair (x_i, y_i) corresponding to the value on the same unit. In paired t-test, where we assumed that the samples have been drawn from the normal population. Here we shall do the problem using sign test. without making assumptions regarding the parent population.

Assumptions:

1. measurement are such that the deviations $d_i = x_i - y_i$ $i = 1, 2, \ldots n$, can be express in terms of +ve (or) -ve signs
2. Variables have continous distribution
3. d_i's are independent.

Scanned with CamScanner

Here, we want to test the if the two populations have same median or not. (or) to test the difference between two populations

Null Hypothesis: H_0: $f_1(x) = f_2(y)$ i.e, the two populations are same (or) The two populations have same Median

Alternative Hypothesis: H_0: $f_1(x) \neq f_2(y)$; i.e, the two populations are not same (or) The 2 population doesnot have same Median.

The level of significance is fixed in advance i.e $\alpha = 5\%$ (or) 1%.

<u>Test statistic</u>: Let (x_i, y_i); $i = 1, 2, \dots, n$ be the paired observations in a sample size 'n'. Here 'x' values refers to the first sample and y-values refers second sample.

Now, find the deviations $d_i = x_i - y_i$; $i = 1, 2 \dots n$ and consider their signs $(+, -, 0)$. In case $x_i = y_i$ i.e; $d_i = 0$, remove those deviations and reduce the sample size by the number of such zero.

Now proceed as in the case of single sample test.

$$U = \min(\text{No. of +ve signs, No. of -ve signs})$$

'U' follows the Binomial distribution with parameters $n, \frac{1}{2}$

The p.m.f of Binomial distribution is $P(X=x) = n_{C_x} p^x q^{n-x}$

$$\rightarrow P(X=U) = (n_{C_U})\, p^U q^{n-U}$$

$$P(X=U) = (n_{C_U}) \left(\tfrac{1}{2}\right)^U \left(\tfrac{1}{2}\right)^{n-U}$$

$$P(X=U) = n_{C_U} \cdot \left(\tfrac{1}{2}\right)^U \quad ; U = 0, 1, 2 \dots n$$

Scanned with CamScanner

By using cumulative probability

$$P(X \leq r) = \sum_{r=0}^{U} \binom{n}{r} \left(\frac{1}{2}\right)^n = p' \text{ (say)}$$

If $p' \geq \alpha$ then we accept H_0 otherwise we reject H_0.

__For large samples : ($n \geq 25$):-__

In case the large samples, the test statistic 'U' is follows to normal distribution with mean $= np = n \cdot \frac{1}{2} = n/2$ and variance $= npq = n \cdot \frac{1}{2} \cdot \frac{1}{2} = n/4$

Then the test statistic is given by

$$Z = \frac{U - n/2}{\sqrt{n/4}} \sim N(0,1)$$

If $Z_{cal} \leq Z_{tab}$ at given level of significance then we accept H_0 otherwise we reject H_0.

__Problems :__

1) A test conducted for 20 students in a school and marks are given below. 93, 88, 107, 115, 82, 97, 103, 86, 113, 107, 112, 90, 98, 93, 99, 103, 100, 101, 96, 104.

The statistical hypothesis of the median of the student in a school is 99 marks.

__Sol:__

H_0: $M = M_0 = 99$. ie, The median of students marks is 99.

H_1: $M \neq M_0 \neq 99$. ie, The median of students marks is not 99

obtains the deviations ie $x_i - M_0$; $i = 1, 2, \dots 20$, then we will get the signs

93, 88, 107, 115, 82, 97, 103, 86, 113, 107, 112, 90, 98; 93, 99, 103, 100, 101, 96, 104

−, −, +, +, −, −, +, −, +, +, +, −, −, −, 0, +, +, +, −, +

No. of +ve signs = 10

No. of −ve signs = 9

No. of zeros = 1

Now adjust out $n = 20 \Rightarrow n = 20 - 1$

$n = 19$

$U = \text{Min}(+\text{ve signs}, -\text{ve signs})$

$U = \text{Min}(10, 9)$

$U = 9$

from cumulative distribution is

$$p' = \sum_{r=0}^{9} \binom{n}{r} (1/2)^n = \sum_{r=0}^{9} ({}^{9}C_r)(1/2)^{19}$$

$p' = 0.05$

$\Rightarrow p' > 0.05$

$\Rightarrow$ we accept H_0

$\therefore$ The median of students marks of the school is 99 marks.

Scanned with CamScanner

Q. The data of weights of 100 students are collected in a university and they are divided into two groups ie, below 60 kg are 30 students and remaining students are above 60 kgs. According to this data can we conclude that students are in equal ratio in two groups?

Sol: given that $n=100$ (large sample)

Null Hypothesis: H_0: The population median is 60 kgs (or) students are in equal ratio in two groups

ie $H_0: M = 60$ kgs

Alternative hypothesis: H_1: The population median is not 60 kgs (or) students are not in equal ratio in two groups

ie $H_1: M \neq 60$ kgs

α is fixed in advance. ie $\alpha = 5\%$.

Test statistic: Here $n > 30$ if it follows to normal

From given data, $U = 30$

Then test statistic is

$$Z = \frac{U - E(U)}{SE(U)} \sim N(0,1)$$

$$E(U) = np = n/2 = \frac{100}{2} = 50$$

$$V(U) = npq = n \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \frac{100}{4} = 25$$

$$Z = \left| \frac{30 - 50}{\sqrt{25}} \right|$$

$$Z = 20/5$$

$$\therefore Z = 4$$

Scanned with CamScanner

Table value at given level of significance 5% is 1.96

$Z_{cal} = 4$ and $Z_{tab} = 1.96$

$Z_{cal} > Z_{tab}$, we reject H_0

∴ The population median is not 60 kgs.

3) 17 students are selected for a special training and a test is conducted to them. Another test is conducted after completion of the training. The marks are given below in grades in the following table. Test the difference in their talent of two tests.

No. of students:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Grades in test I:	2	3	3	3	3	3	3	3	2	3	2	2	5	2	5	3	1
Grades in test II:	4	4	5	5	3	2	5	3	1	5	5	5	4	5	5	5	5

Sol^n: $H_0: f_1(x) = f_2(y)$

$H_1: f_1(x) \neq f_2(y)$

$\alpha = 5\%$ l.o.s

To find $d_i = x_i - y_i$ = grades of test I − grades of test II

−, −, −, −, 0, +, −, 0, +, −, −, −, +, −, 0, −, −

No. of +ve signs = 3

No. of −ve signs = 11

No. of zeros = 3.

Scanned with CamScanner

(7)

Adjusted sample size $\Rightarrow n = 17-3$

$n = 14$

$U = \min(\text{No. of +ve signs}, \text{No. of -ve signs})$

$U = \min(3, 11)$

$U = 3$

$$p = \sum_{r=0}^{u} \left({}^{n}C_r\right)\left(\tfrac{1}{2}\right)^{n}$$

$$p = \sum_{r=0}^{3} \left({}^{14}C_r\right)\left(\tfrac{1}{2}\right)^{14}$$

$p = 0.0470 = p'$ (say)

$\Rightarrow p' < \alpha$

ie. $0.0470 < 0.05$ we reject H_0.

$\therefore$ There is a difference in their talents in the two tests.

* Wald – Wolfowitz Run Test (One Sample):–

Single Sample Run test (or) Run test for Randomness:–

The tests of significance is based on the fundamental assumption that the samples drawn from the population is random. Hence, it is desirable to verify if this assumption of randomness of the samples is true or not. The test of significance devised to the test of randomness of a sample is the "Run test", which is based on the theory of "Runs".

Defination of Run :–

A Run is sequence of letters of one kind surrounded by the sequence of letters of other kind. Number of elements in a run is usually ~~reffere~~ referred to as the length of the Run.

Example: Suppose the applications for the posts of mangement trainees in a business house arrive in the following order (M: Male, F: Female)

M FF MMM FF M FF MMM FFFF MM

There are in all nine runs in this sequence, as indicated above,

n_1 = Number of letters of first kind (Males) = 10

n_2 = # of letters of second kind (females) = 10

n = Total Sample Size $= n_1+n_2 = 20$

r = number of runs in the sequence = 9. $\left[\because r_1 = 5,\ r_2 = 4\right]$ $r = r_1 + r_2 = 9$

It is not necessary that the sample data are given in the form of symbols. Any set of sample observations in the form of numerical data can be converted into symbolic form and hence into runs.

Scanned with CamScanner

(8) **Example:** Suppose $x_1, x_2 \ldots x_n$ is the set of 'n' sample observations (numerical values) where, x_i = i^{th} sample observation in the occurrence of the experiment ($i = 1, 2, \ldots n$).

we obtain the ~~Mean~~ Median value (M) of the set of sample observations and see if each of the given observations is $\geq M$ or $< M$

→ If the observation is $\geq M$, replace it by letter 'A'.

→ If the observation in $< M$, replace it by letter 'B'.

we shall obtain a sequence of the symbols A and B of the type, say.

B A A B A B B B A A A A B B A.

n_1 = number of letters of A is = 8

n_2 = number of letter of B is 7

$n = n_1 + n_2 = 15$

r_1 = number of runs of letter 'A' is 4

r_2 = number of runs of letters B is 4

$r = r_1 + r_2 = 8$ (length of run)

Testing of Hypothesis of randomness :-

The sequence of sample observations cannot be considered to be random,

→ If similar items call cluster together resulting in too few runs, consider the following results obtained in a toss of a coin 10 times

T T T T T H H H H H.

Here, the proportion of heads and tails is same, the sequence cannot be considered as random because there are only two runs, a run of five tails followed by a run of five heads.

Scanned with CamScanner

→ If similar items would alternating mix, resulting in too many runs consider the following sequence of extreme outcomes in toss of a coin 10 times.

THTHTHTHTH

Here also, the proportion of heads and tails is same. the sequence cannot be considered as random because there are only too many (10) runs, 5 runs of one tail each and 5 runs of one head each.

Null Hypothesis: H_0: The process that generates the sample data is random

Alternative hypothesis: H_1: The process that generates the sample data is not random (T.T.T)

Level of significance fixed in advance. ie $\alpha = 5\%$ or 1%.

Setup the level of significance $\alpha\%$.

→ The basic sample principle in the single sample run test is to reject the null hypothesis of randomness if the given sequence of observations results in too few or too many runs.

→ Since too few or too many runs indicate lack of randomness of the sample, a two-tailed alternative hypothesis is appropriate.

→ To determine the values of r which are too few or too large, we need the sampling distribution of runs (r) under the null hypothesis (H_0) of randomness.

Scanned with CamScanner

(9)

Test statistic:

If $n_1 \le 20$ and $n_2 \le 20$, the test statistic is r = the observed number of runs in the sample.

Table values for $n_1 \le 20$ and $n_2 \le 20$ and at α% level of significance, the critical values of 'r' for a two tailed test for different combinations of n_1 and n_2 are given in the Run test table.

Conclusion:

If 'r' lies between lower critical value (r_1) and upper critical value (r_2), then we accept H_0 otherwise we reject H_0.

Reject null hypothesis of the randomness of the sample at 5% LOS if for given combinations of n_1 and n_2.

$r \le$ the lower critical value (r_1) given in Run test table (or)

$r \ge$ the upper critical value (r_2) given in Run test table.

Large sample: If anyone of n_1 and n_2 is greater than 20, the sampling distribution of the number of runs 'r' is approximately normally distributed

$$\text{with mean} = E(r) = \frac{2n_1n_2}{n_1+n_2} + 1$$

$$\text{and variance} = V(r) = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}$$

n_1 = number of symbols of on kind

n_2 = number of symbols of other kind

$$\text{The test statistic is } Z = \frac{r - E(r)}{S.E(r)} \sim N(0,1)$$

Scanned with CamScanner

$$Z = \frac{r - \left(\frac{2n_1 n_2}{n_1+n_2} + 1\right)}{\sqrt{\frac{2n_1 n_2 (2n_1 n_2 - n)}{n^2(n-1)}}} \sim N(0,1) \text{ ; where } n = n_1 + n_2.$$

Conclusion: Calculat Z_{cal} value and get the table value at given α% L.O.S, then compare Z_{cal} with Z_{tab} value. If calculated value is less than or equals to table value we accept H_0, otherwise we reject H_0.

→ If $Z_{cal} \le Z_{tab}$, we accept H_0, o/w we reject H_0.

---

* Wald – Wolfowitz Run test (Two samples):-

Let $X_1, X_2 \dots X_{n_1}$ and $Y_1, Y_2 \dots Y_{n_2}$ be two independent random samples of sizes n_1 and n_2 respectively from two populations with probability density functions $f_1(x)$ and $f_2(y)$ respectively. We want to test the hypothesis that the samples have been drawn from the same population or from two identical populations with same probability density functions.

Null Hypothesis: H_0: $f_1(x) = f_2(y)$.

ie; two samples are drawn randomly from two different populations with same distribution (or) two samples coming from same distribution

Alternative hypothesis: H_1: $f_1(x) \neq f_2(y)$

ie; two samples coming from different distribution.

(10) Level of significance is fixed in advance. i.e; $\alpha = 5\%$ or 1%.

Now, Combine the two samples to obtain a single sample of size n_1+n_2 and arrange the observations in the Combined Sample in ascending order of magnitude. And denote the observation with 'A' if it comes from the first sample and denote with 'B' if it comes from second sample, then we are getting sequence and count the n_1, n_2, r_1, r_2 and r.

If $n_1 \leq 20$ and $n_2 \leq 20$, the test statistic is r.

r = the observed number of runs in the sample.

Table values for $n_1 \leq 20$ and $n_2 \leq 20$ and $\alpha\%$ L.O.S, the critical values of 'r' for a two tailed test, for different combinations of n_1 and n_2, are given in the Run test table

Conclusion:

If 'r' is lies between lower critical value (r_1) and upper critical value (r_2), then we accept Ho, otherwise we reject Ho.

Reject Null hypothesis of the randomness of the sample at 5% Level of significance, if for given combinations of n_1 & n_2

$r \leq$ The lower critical value (r_1) given in Run test table (or)

$r \geq$ The upper critical value (r_2) given in Run test table.

Scanned with CamScanner

Large Sample:

when $n_1 > 20$ or $n_2 > 20$, the sampling distribution of the number of runs 'r' is approximately normally distributed

with mean $= E(r) = \dfrac{2n_1 n_2}{n_1 + n_2} + 1$

and variance $= V(r) = \dfrac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}$, where $n = n_1 + n_2$

n_1 = number of symbols of one kind

n_2 = number of symbols of other kind

Test statistic is given by $Z = \dfrac{r - E(r)}{SE(r)} \sim N(0,1)$

$$Z = \frac{r - \dfrac{2n_1 n_2}{n_1 + n_2} + 1}{\sqrt{\dfrac{2n_1 n_2 (2n_1 n_2 - n)}{n^2 (n-1)}}} \sim N(0,1)$$

Conclusion: If $Z_{cal} \leq Z_{tab}$ at given $\alpha\%$ L.O.S, then we accept H_0, otherwise reject H_0.

Scanned with CamScanner

(11) Problems:

1) Test the randomness to the following data.

109, 124, 173, 167, 148, 132, 168, 165, 118, 112, 114, 164, 180, 123, 180, 152.

Sol: Null Hypothesis: H_0: The given sample has been drawn randomly from a population

first calculate median, arrange the sample in ascending order as

109, 112, 114, 118, 123, 124, 132, 148, 152, 164, 165, 167, 168, 173, 180, 180.

$$\text{median} = \frac{148+152}{2}$$

$$\text{median} = 150$$

Now denote A if $x_i \geq M$ and denote B if $x_i < M$ to the given data.

$\underline{BB}\ \underline{AA}\ \underline{BB}\ \underline{AA}\ \underline{BBB}\ \underline{AA}\ \underline{B}\ \underline{AA}$

Total number of run = 8 $[\because r_1 = 4, r_2 = 4]$

$n_1 = 8$

$n_2 = 8$

Corresponding to the values of $n_1=8$, $n_2=8$ and 5% L.O.S the critical values are $r_1=4$, $r_2=14$.

$\therefore r_1 < r < r_2$ (ie, $4 < r = 8 < 14$)

$\therefore$ we accept H_0

Hence the given sample has been drawn randomly from a population.

Scanned with CamScanner

2) 25 heads are observed of 37 throws of a coin. Test whether the coin is unbaised if the total runs are 13.

Soln: H_0: Coin is unbiased.

given that $n = 37$

n_1 = number of heads attained $= 25$

$\Rightarrow$ n_2 = number of tails attained $= 37 - 25 = 12$.

and $r = 13$

$n_1 > 20$ i.e. we have to use normal test as

$$Z = \frac{r - \left(\frac{2n_1n_2}{n_1+n_2} + 1\right)}{\sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}} \sim N(0,1)$$

$$Z = \frac{13 - \left(\frac{(2)(25)(12)}{25+12} + 1\right)}{\sqrt{\frac{(2)(25)(12)((2)(25)(12) - 25 - 12)}{(25+12)^2(25+12-1)}}}$$

$Z = 1.61$, ie, $Z_{cal} = 1.61$

at 5% level of significance, the table value is 1.96

$\therefore$ $Z_{tab} = 1.96$

$\therefore$ $Z_{cal} < Z_{tab}$, we accept H_0.

Hence the coin is unbiased.

Scanned with CamScanner

3) The data is given below of scores of 6 Secretariat clerks and 7 clerks of Directorate office. Can we assume that these scores of clerks follows same distribution.

Score of Secretariate	40	35	52	60	46	55	
Score of Directorate office	47	56	42	57	50	57	62

Sol: Ho: $f_1(x) = f_2(y)$; i.e, Scores of clerks in two offices are following same distribution.

→ Arrange in ascending order of two combined sample

35	40	42	46	47	50	52	55	56	57	57	60	62
A	A	B	A	B	B	A	A	B	B	B	A	B

The order is, <u>A A</u> B <u>A</u> <u>BB</u> <u>AA</u> <u>BBB</u> <u>A</u> <u>B</u>

Total number of runs = 8 $[\because r_1 = 4, r_2 = 4]$

$n_1 = 6$ and $n_2 = 7$

Corresponding to the values of $n_1 = 6$ and $n_2 = 7$ at 5% L.O.S, the Critical values from the run tables are $r_1 = 3$ and $r_2 = 12$

$\therefore r = 8$ lies in between $r_1 = 3$ and $r_2 = 12$

i.e $r_1 < r < r_2$

∴ we accept Ho.

Hence we Conclude, the two Scores of clerks in the two offices follows same distribution

H/W 4) The gains and losses are arranged in an order of a football team in the year 2000 and is given below. Test the randomness.

WLWWL W LLL WLL WW L WWL WL WLWLLWLWL

Scanned with CamScanner

* Wilcoxon - Signed rank test. (Single samples) :-

The sign test uses only the information of whether the value is above or below median. Consequently, it may fail in rejecting a false null hypothesis. In this section we introduce another test called the wilcoxon signed rank test that, in addition to the signs, also use the amount of deviation of data values from the median value and hence is considered to be better than the sign test.

procedure :

Let $x_1, x_2 \ldots x_n$ be a random sample drawn from a population of size n and ~~let~~ let M be the median of the given distribution (or) for the given observations M_0 say.

Null hypothesis: $H_0: M = M_0$

Alternative hypothesis: $H_1: M > M_0$ (R.T.T)

$M < M_0$ (L.T.T)

$M \neq M_0$ (T.T.T)

Level of significance fixed in advance. i.e $\alpha = 5\%$ or 1%.

Test statistic: In testing the null hypothesis that $M = M_0$. we subtract M_0 from the each sample observation. If the difference is zero for any observation and reduce sample size 'n' accordingly. The remaing differences are arranged in increasing order of their absolute value.

The ordering determines the rank (position) of each data value, and to each rank we attach sign (plus or minus) of the corresponding difference. These ranks are called the signed - ranks

Scanned with CamScanner

And then Calculate Sum of ranks of +ve signs and −ve signs respectively. obtain 'T' as minimum of these sums.

$$T^+ = \text{sum of +ve ranks}$$

$$T^- = \text{sum of −ve ranks.}$$

The test statistic is $T = \min(T^+, T^-) = T_0$ (say).

we will get the critical value of 'T' for the corresponding value of 'n' and at specified level of significance.

Compare Calculate value and table value. If $T > T_0$, we accept H_0. otherwise we reject H_0.

* <u>Single Sample wilcoxon Signed rank test (large samples)</u>

If the Sample size is large ($n \geq 25$), then we use normal test.

Hence 'T' follows normal distribution with mean and variance as

$$\text{mean} = E(T) = \frac{n(n+1)}{4} \quad \text{and} \quad \text{variance} = V(T) = \frac{n(n+1)(2n+1)}{24}$$

∴ Then the test-statistic $z = \dfrac{T - E(T)}{S.E(T)} \sim N(0,1)$

$$S.E(T) = \sqrt{V(T)}.$$

$$\Rightarrow \quad Z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \sim N(0,1).$$

Scanned with CamScanner

(14) Calculate $|Z|$ and compare it with the tabulated value of Z at 'α' level of significance.
If calculated value is greater than the tabulated value we reject H_0, otherwise we accept H_0.

* Wilcoxon signed rank test for two samples (paired sample)

The wilcoxon signed rank test can also be used to compare two non-normal populations.
~~Let $x_1, x_2, \ldots x_{n_0}$ and~~ Let $x_1, x_2, \ldots x_n$ and $y_1, y_2 \ldots y_n$ be two random samples of sizes 'n' from two different populations with probability density functions $f_1(x)$ and $f_2(y)$ respectively.

Null hypothesis : H_0: $f_1(x) = f_2(y)$
i.e: the two random samples are coming from the same distⁿ.
Alternative hypothesis: H_1: $f_1(x) \neq f_2(y)$
ie; the two random samples are not coming from the same distribution.

~~[illegible]~~ Level of significance fixed in advance ie $\alpha = 5\%$ (or) 1%.

Test statistic:

first we calculate $d_i = x_i - y_i$; $i = 1, 2, \ldots n$ for each pair of the observations. If the difference is zero for any ~~other~~ pair, and reduce the sample size 'n' accordingly. the remaing differences are arranged in increasing order of their absolute value.

Scanned with CamScanner

The ordering determines the rank (position) of each data value, and to each rank we attach the sign (+ve or −ve) of the corresponding difference. These ranks are called signed-rank. And then calculate sum of ranks of +ve signs and −ve signs separately. Obtain T as minimum of these sums.

i.e $T = [T^+, T^-]$

where T^+ = Sum of +ve ranks

T^- = Sum of −ve ranks.

we will get the critical value of T for the corresponding value of 'n' and at specified level of significance.

Compare calculated value with the tabulated value. If $T_{cal} > T_{tab}$, we accept H_0 otherwise we reject H_0.

<u>Wilcoxon signed-rank test for two samples:-</u>

If $(n > 25)$ the sample size is large, then we use normal test.

For large values of n, the statistic 'T' is approximately normal with mean and variance as

$$\text{mean} = E(T) = \frac{n(n+1)}{4} \text{ and variance} = V(T) = \frac{n(n+1)(2n+1)}{24}.$$

Then the test statistic is $Z = \frac{T - E(T)}{S.E(T)}$.

Calculate $|Z|$ and compare it with the tabulated value of 'Z' at 'α' level of significance. If calculated value is greater than the tabulated value, we reject H_0. otherwise we accept H_0.

Scanned with CamScanner

(15) * **problem:**

1) A Sample of 12 pairs of twins are taken at random and their intelligence scores are given below. According to this data test the intelligence level of the first person of the twins is more than the second person.

Twin pair	1	2	3	4	5	6	7	8	9	10	11	12
Score of I person of Twin	86	79	77	68	91	72	77	91	70	71	88	87
Score of II person of Twin	88	77	76	64	96	72	65	90	65	80	81	82

Sol^n: $H_0: f_1(x) = f_2(y)$

∴ ie; Intelligence skills are same for twins.

$H_1: f_1(x) \neq f_2(y)$

ie; Intelligence skills are not same for twins.

Score of 1st person	Score of 2nd person	$d_i = x_i - y_i$	$\lvert d_i \rvert$	Ranks for d_i
86	88	-2	2	3.5
79	77	$+2$	2	3.5
77	76	-1	1	1.5
68	64	4	4	5
91	96	-5	5	6.5
72	72	0	0	—
77	65	12	12	10
91	90	1	1	1.5
70	65	5	5	6.5
71	80	-9	9	9
88	81	7	7	8
87	82	5	5	11

Scanned with CamScanner

Sum of +ve sign ranks $= 47 = T^+$

Sum of −ve sign ranks $= 19 = T^-$

$\therefore T = \min(T^+, T^-)$

$\Rightarrow T = \min(47, 19)$

$T = 19$.

Table value of T at 5% level of significance and for $n = 11$ (Since one ~~d~~. $d_i = 0$, $\therefore n = 12 - 1 = 11$) from wilcoxon ~~$\therefore n = 12 - 1 = 11$~~ signed rank test ~~statistics~~ tables

$T_{tab} = 11$.

$\because T_{cal} > T_{tab}$, so we accept H_0.

$\therefore$ Intelligence levels of both pair of twins are same.

(18) * **Mann-Whitney U-test (Two Samples):-**

The wilcoxon signed rank test can be used to compare two populations when the two sets of sample data are paired. Suppose there is no natural way to pair up the data values, that is the samples are independent. In this case we will introduce a test called Mann-whiteney test, that enables us to compare two independent samples by testing a hypothesis on the two population medians.

This non-parametric test for two independent samples was described by wilcoxon and studied by Mann and whitney. Mann-whiteney is the most widely used test as an alternative to the t-test for two independent samples when we do not make the t-test assumptions about the parent population. The only assumption is that the random variables are continous.

Assumptions:-

1) The two samples are random
2) The two samples are independent
3) The sizes of both the samples are more than or equal to 10. ie, $n_1 \geq 10$ and $n_2 \geq 10$.
4) No assumption is made about the form of the distributions for which samples are drawn.

Let $x_1, x_2, \ldots x_{n_1}$ and $y_1, y_2, \ldots y_{n_2}$ be independent random samples of size n_1 and n_2 from the populations with probability density functions $f_1(x)$ and $f_2(y)$ respectively.

Scanned with CamScanner

we want to test if the samples have been drawn from the same population or not. i.e. $f_1(x) = f_2(y)$.

The Mann-Whitney test uses the test statistic 'U' to compare the population medians M_x and M_y.

Null hypothesis: $H_0: M_x = M_y$ (or) $f_1(x) = f_2(y)$

i.e: The two samples are drawn from same distribution.

Alternative hypothesis: $H_1: M_x > M_y$ (or) $f_1(x) > f_2(y)$ (R.T.T)

(or)

$H_1: M_x < M_y$ (or) $f_1(x) < f_2(y)$ (L.T.T)

(or)

$H_1: M_x \neq M_y$ (or) $f_1(x) \neq f_2(y)$ (T.T.T)

Level of significance fixed in advance i.e $\alpha = 5\%$ or 1%.

The Mann-Whitney test, uses the sum of the ranks (or rank sum) of each sample. To assign rank to each observation of the two samples, we combine two samples into a single sample and arrange the observations of the combined sample in ascending order of their magnitude. Then ranks are assigned to the combined sample data.

Let T_x = Sum of the ranks of the observations of the first sample (x-values) in the combined sample.

T_y = Sum of the ranks of the observations of the 2nd sample (y-values) in the combined sample.

Scanned with CamScanner

(17) Mann-whitney U-statistics are defined as follows

$$U_x = T_x - \frac{n_1(n_1+1)}{2} \quad \text{and} \quad U_y = T_y - \frac{n_2(n_2+1)}{2}$$

where n_1 and n_2 are the number of observations in the first and second sample respectively.

Test Statistic is $U = \min(U_x, U_y)$.

The statistics U_x and U_y are also defined as

$$U_x = n_1 n_2 + \frac{n_2(n_2+1)}{2} - T_y$$

$$U_y = n_1 n_2 + \frac{n_1(n_1+1)}{2} - T_x .$$

It may be noted that the sum of the two samples sizes, that is $U_x + U_y = n_1 n_2$.

$$T_x + T_y = \frac{(n_1+n_2)(n_1+n_2+1)}{2}$$

'U' calculated value compare with the 'U' critical value at α% level of significance.

If $U_{cal} \geq U_{tab}$, we accept H_0, otherwise we reject H_0.

Scanned with CamScanner

* Mann-whitney U-test for large samples :-

Use the normal test in the usual manner. The normal approximation 'U' is fairly good if both n_1 and $n_2 > 10$.

If the sample size is large, then we may use the normal test for testing H_0. 'U' is approximately normal with mean and variance are, given by

$$\text{mean} = E(U) = \frac{n_1 n_2}{2} \quad \text{and Variance} = V(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}$$

$\therefore$ The test statistic is $Z = \dfrac{U - E(U)}{S.E(U)} \sim N(0,1)$

where $SE(U) = \sqrt{V(U)}$.

Compare $|Z|$ value with the ~~tablo~~ tabulated value of Z at given $\alpha\%$ level of significance.

If $|Z| \geq Z_{tab}$, we reject H_0, otherwise we accept H_0.

—x—

Scanned with CamScanner

(8) * Median test :-

Median test is a statistical procedure to test if two independent samples have been drawn from two population with the same Median.

Let $x_1, x_2 \dots x_{n_1}$ and $y_1, y_2 \dots y_{n_2}$ be two independent random samples of sizes n_1 and n_2 respectively from two populations with probability density functions $f_1(x)$ and $f_2(x)$ respectively. We want to test if two independent samples have been drawn from two populations with the same Median.

Null hypothesis: H_0: Samples have been drawn from the populations with the same Median. i.e., $f_1(x) = f_2(y)$

Alternative hypothesis: H_1: population medians are different. i.e. $f_1(x) \neq f_2(y)$

Level of significance fixed in advance. i.e $\alpha = 5\%$ (or) 1%.

Test statistic: Let us take two independent random sample of sizes n_1 and n_2 from the given populations. Combine the two sample observations and arrange them in ascending order of magnitude. Obtain the Median say M. of the ~~both~~ combained sample. Find the number of observations in each sample which are above M and which are not above M.

Let m_1 = number of observations of the first sample which are above M. i.e, $> M$

m_2 = number of observations of the second sample which are above M. i.e, $> M$.

Scanned with CamScanner

$n_1 - m_1$ = number of observations of the first sample which are not above M. i.e, $\leq M$

$n_2 - m_2$ = number of observations of second sample which are not above M. i.e, $\leq M$

The data so obtained can be arranged in the 2x2 contigency table as given below.

	Sample-I	Sample-II	Total
No. of observations $> M$	m_1	m_2	$m_1 + m_2$
No. of observations $\leq M$	$n_1 - m_1$	$n_2 - m_2$	$n_1 + n_2 - (m_1 + m_2)$
Total	n_1	n_2	$n_1 + n_2$

If either n_1 or n_2, or both is less than 10, then we apply hyper geometric distribution.

$$P(m_1) = \frac{{}^{n_1}C_{m_1} \cdot {}^{n_2}C_{m_2}}{{}^{n_1+n_2}C_{m_1+m_2}}.$$

we calculate P, If $P > \alpha$ then we accept H_0., otherwise we reject H_0.

If both n_1 and n_2 greater than 10, we apply chi-square distribution. we calculate χ^2 using 2x2 contingency table.

Scanned with CamScanner

$$\chi^2 = \frac{N\,(ad-bc)^2}{(a+c)(b+d)(a+b)(c+d)}$$

where $N = a+b+c+d$

Compare this value of 'χ^2' with the significant value of Chi-square at 1 degrees of freedom. If $\chi^2 > \chi^2_{(\alpha,1)}$ at specified level of significance 'α', then we reject Ho, otherwise we accept Ho.

In this case, If any one of the frequency is less than 5, then we apply Yate's Correction for chi-square statistic.

$$\chi^2 = \frac{N\left[|ad-bc| - \frac{N}{2}\right]^2}{(a+c)(b+d)(a+b)(c+d)}$$

=o=

<u>problems</u>:

1) The number of defective items produced from two machines A and B are given below. Test whether these samples are drawn from the ~~Sampl~~ same population using median test.

No. of defectives from Machine A	26	27	31	26	19	21	20	25	30
No. of defectives from Machine B	23	28	26	24	22	19			

Sol^n: Null hypothesis: H_0: $f_1(x) = f_2(y)$

i.e, two samples have drawn from the same popl^n.

Combining the two samples in ascending order.

19 19 20 21 22 23 24 25 26 26 26 27 28 30 31

given that, $n_1 = 9$, $n_2 = 6$

$n_1 + n_2 = 9 + 6$

$n_1 + n_2 = 15$

Median $= 25$

$m_1 = 6$

$M_2 = 2$

$\therefore$ 2×2 Contingency table.

	Sample I	Sample II	Total
$>M$	6	2	8
$\leq M$	3	4	7
Total	9	6	15

Since n_1 and $n_2 < 10$

$$\therefore P = \frac{{}^9C_6 \; {}^6C_2}{{}^{15}C_8} \Rightarrow P = 0.1958$$

$P = 0.1958 > \alpha = 0.05$ at 5% Los, therefore we accept H_0, we conclude two samples have drawn from the ~~sample~~ same population.

Scanned with CamScanner

(19) H/w:

2) The sample observations of two sample given below :

Sample I : 235, 256, 258, 220, 250, 225, 224, 247
207, 248, 254, 206, 230, 251, 268, 245
315, 225

Sample II : 258 228 250 225 243 249 243 229
237 239 222 206 227 211 206 220
226 225 214 232 236 205 255

Test whether two random samples have drawn from the ~~sample~~ same population.

Scanned with CamScanner