Non-Parametric Inferential Statistics

The entire topic is divided into 5 sub-divisions

1. **Objectives**
2. **Introduction**
3. **One Sample tests and their procedures**
4. **Two or more sample tests and their procedures**
5. **Summary**

1. **Objectives:** Dear learners, why to study Non-Parametric Inferential Statistics?
    i. **To test the hypothesis when we can not make any strict assumptions about the form of the distribution from which we are sampling.**
    ii. **To know which non-parametric test is more appropriate for different situations.**

2. **Introduction :** All of we know that, parametric tests are based on some strict assumptions, about the form of distribution, from which the sample was drawn.

The most common parametric assumptions are:

i. Data are approximately normally distributed and
ii. the key parameters(e.g., the mean or difference in means or the sd) of the distribution that are involved in estimation from the sample data.

But in real world, data may not be normally distributed and thus, to analyze such data non-parametric methods can be used as a counter part. That is, non-parametric procedures are one possible solution to handle non-normal data. It uses data that is often nominal or ordinal, that is, it does not rely on numbers, but rather rankings.

3. **One sample Tests and their procedures:**

    There are several one sample tests and some of them are listed below:

    i. Sign test for one sample
    ii. Wilcoxon signed rank test
    iii. Kolmogorov – Smirnov(KS) one sample test.

i. Kolmogorov-Smirnov(KS) One-sample test:

The Kolmogorov – Smirnov one sample test is a test of goodness of fit i.e., it is concerned with the degree of agreement between distribution of a set of sample values and some specified theoretical distribution. It determines whether the scores in the sample have come from a population having the theoretical distribution or not.

Procedure : Let $x_1, x_2,\ldots, x_n$ be a random sample from F(x). We define empirical cumulative distribution function of a random variable X as

$$F_n(x) = \frac{number\,of\,observations\,X_j \le x}{n}.$$

For fixed x, $F_n(x)$ is a statistic, since it depends on the sample. Under $H_0 : F(x)= F_0(x)$, for all x, the KS one sample test for goodness of fit statistic is defined as

$$D_n = Max| F_n(x) – F_0(x)|$$

Where $F_0(x)$ is the specified cumulative frequency distribution under H_0 against the two sided alternative

$$H_1: F(x) \neq F_0(x), \text{ for some x.}$$

For testing H_0 V/s $H_2 : F(x) \geq F_0(x)$, for all x, with strict inequality for some x, one sided KS statistic is

$$D^+_n=Max\{F_n(x) – F_0(x)\}$$

For testing H_0 V/s $H_3 : F(x) \leq F_0(x)$, for all x, with strict inequality for some x, one sided KS statistic is

$$D^-_n=Max\{ F_n(x) – F_0(x)\}$$

Critical region :

If $D_n \geq D_{n,\alpha}$ reject H_0. Where $D_{n,\alpha}$ is a critical value which can be obtained by referring to the Kolmogorov – Smirnov(KS) one sample test statistic table for different values of n.

It is expected that for every value of 'x', $F_n(x)$ should be fairly close to $F_0(x)$. i.e., under H_0, we would expect the differences between $F_n(x)$ and $F_0(x)$ to be small and with the limits of random errors.

Eg.1. (Kolmogorov-smirnov) : The following is a random sample of size 5. Test whether the sample can be considered as a sample from a N(0, 1) distribution: -1.152, -0.625, 0.682, -0.870, 1.405.

X	$F_0(x)$	$F_n(x)$	$\lvert F_0(x) - F_n(x) \rvert$
-1.152	0.1247	1/5=0.2	0.0753
-0.870	0.3078	2/5=0.4	0.0922
-0.625	0.2341	3/5=0.6	0.3659
0.682	0.2517	4/5=0.8	0.5483
1.405	0.4199	1	0.5801

Here, the null hypothesis $H_0 : F(x) = F_0(x)$ V/s $H_1 : F(x) \neq F_0(x)$.

Under H_0, the KS- statistic D_n is

$D_5 = \text{maximum}\lvert F_0(x) - F_n(x) \rvert = 0.5801 > 0.563 = D_{5, 0.05}$(from Table).

We reject H_0 at 5% level of significance and conclude that $H_1 : F(x) \neq F_0(x)$.

4. Two or more sample tests and their procedures:

There are several two ore more sample tests and some of them are listed below:

i. Wilcoxon-Mann-Whitney U-test
ii. Wald – Wolfowitz run test
iii. The Median test
iv. The Kruskal - Wallis – one way Analysis of Variance test
v. Spearman's Rank Correlation Test

i. Wilcoxon-Mann-Whitney U-test.

Wilcoxon-Mann-Whitney -U-test is one of the most powerful non-parametric tests. It is used as an alternative to a parametric two sample t-test.

Without assuming that the two samples have come from normal population when it is ordinal measurement has been achieved, the Mann-Whitney - U-test may be used to test whether two independent groups have been drawn from the same population or not.

Procedure:-

Let $x_1, x_2, \ldots, x_{n1}$ be a random sample from the first population and let $y_1, y_2, \ldots, y_{n2}$ be a random sample from the second population with their pdf f(x) and f(y) respectively.

Let $z_1, z_2, \ldots, z_{n1+n2}$ be the combined ordered sample. Assign rank '1' to the lowest score, rank '2' to the next lowest score and so on. If there is any significance difference between two populations. [i.e., if the two distribution are identical], most of the lower ranks are likely to go to values of the first sample and most of the higher ranks are likely to go to the values of the second sample and vice versa.

Let R_1 - be the sum of the ranks of the values of the first sample and R_2 - be the sum of the ranks of the values of the second sample . The problem is to test null hypothesis H_0 : f(x)= f(y) against the alternative H_1 : $f(x) \neq f(y)$.

Let $U_1 = R_1 - \frac{n_1(n_1+1)}{2}$ and $U_2 = R_2 - \frac{n_2(n_2+1)}{2}$

then we have

$U = \text{minimum}(U_1, U_2)$.

It is derived from that

$$E(U) = \frac{n_1 n_2}{2} \quad \text{and} \quad Var(U) = \frac{n_1 n_2 (n_1 + n_2 + 1)}{12}.$$

For large n, under H_0 the test statistic Z is defined by

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}} \sim N(0, 1), \text{ asymptotically.}$$

Reject H_0 if $|Z| \geq Z_{\alpha/2}$ otherwise accept H_0. One may draw conclusion by referring to the standard normal table for the fixed level of significance α.

Eg. 2.(**Wilcoxon-Mann-Whitney U-test.**):

The following data refers to the percentage of recovery of dextroamphetramine extracted after seven hours bay a sample of children having organically related disorder and a sample of children having non-organic disorders.

Organic (X) : 17.53 20.60 17.62 20.93 27.10

Non-organic(Y): 15.59 14.76 13.32 12.45 12.79

Use Mann-Whitney-Wilcoxon test to test at $\alpha = 0.05$ the hypothesis that the two distributions are the same against the alternative that $\mu_1 > \mu_2$.

Solution : Here we Combine the two samples and the combined ordered Sample :

12.45, 12.79, 13.32, 14.76, 15.59, 17.53, 17.62, 20.60, 20.93, 27.10

Rank : 1 2 3 4 5 6 7 8 9 10

R_1= sum of ranks(X) = 6+7+8+9+10 = 40

R_2= sum of ranks(Y) = 1+2+3+4+5 = 15

$H_0 : \mu_1 = \mu_2$ V/s $H_1 : \mu_1 > \mu_2$

$U_1 = 40-5(6)/2 = 25$ and $U_2 = 15-5(6)2= 0$

Therefore, $U = \min(U_1, U_2) = \min(25, 0) = 0$.

E(U) =12.5 and Var(U) = 22.9167

We use large sample approximation so that under H_0 the test statistic Z is defined by

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}} \sim N(0, 1), \text{ asymptotically.}$$

Z = -2.61

We reject H_0 since

$Z \geq Z_{\alpha/2} = 1.96$ at 5% level of significance and conclude that $H_1: \mu_1 > \mu_2$.

ii. Wald- Wolfowitz Run Test :

"A run is defined as a sequence of letters of one kind surrounded by a sequence of letters of the other kind and the number of elements in a run is usually referred to as the length(L) of the run".

Procedure : Consider the two ordered samples ($x_1, x_2, \ldots, x_{n1}$) and ($y_1, y_2, \ldots, y_{n2}$), which are drawn from two populations with density $f_1(x)$ and $f_2(y)$ respectively. The problem is to test if the samples have been drawn from the same population or from populations with the same density functions i.e., if $f_1(x) = f_2(y)$.

Let us combine the two samples and arrange the observations in order of magnitude to give the combined ordered sample as,(say)

$$x_1, x_2, y_1, y_2, y_3, y_4, x_3, x_4 \ldots\ldots \qquad (1)$$

Thus in (1) we have in order, a run of x = 2 (i.e. L = 2), a run of y = 4,(i.e. L=4), and so on. Here the problem is to test the null hypothesis H_0: $f_1(x) = f_2(y)$. That is the samples have come from the same population. Let U be the number of runs in the combined ordered sample. Then the we shall find the distribution of U under H_0. Under H_0, all permutations of the n_1 observations of X and n_2 observations of Y have equal probabilities. We can select the n_1 positions for the n_1 values in $\binom{n_1+n_2}{n_1}$ ways with probability $\dfrac{1}{\binom{n_1+n_2}{n_1}}$

Null hypothesis is rejected if $U < u_0$, where the value of u_0 for given level of significance is determined from considering the distribution of u under H_0.

To find P[U = u], we must determine the number of permutations that yield u runs. First we consider u = 2k,

(k > 0 is an integer) the even number of runs. In this case we have k-runs of X's and k-runs of Y's. n_1 X's will give k runs if they are separated by (n_1-1) dividers in distinct places between the X's with no more than one divider per space. This can be done in $\binom{n_1-1}{k-1}$ ways. Similarly k runs of the n_2 values of Y's can be done in $\binom{n_2-1}{k-1}$ ways. These two sets of runs can be placed together to form u = 2k runs of which $\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}$

begin with runs of X's and $\binom{n_2 - 1}{k - 1}$ $\binom{n_1 - 1}{k - 1}$ begin with runs of Y's. Thus the sequence of runs of the

above type may start with either X's or Y's and we have

$$P[U = 2k] = \frac{2\binom{n_1 - 1}{k - 1}\binom{n_2 - 1}{k - 1}}{\binom{n_1 + n_2}{n_2}}$$

If the number of runs in (1) is odd, i.e., u = 2k+1, it is possible to have either

(i) k+1 runs of the ordered values of X and k runs of the ordered values of Y or

(ii) k runs of X's and k+1 runs of Y's. Hence,

$$P[U = 2k+1] = P[i] + P[ii] = \frac{\binom{n_1 - 1}{k}\binom{n_2 - 1}{k - 1} + \binom{n_1 - 1}{k - 1}\binom{n_2 - 1}{k}}{\binom{n_1 + n_2}{n_1}}$$

Therefore, the distribution of U under H_0 is

$$P[U = 2k] = \frac{2\binom{n_1 - 1}{k - 1}\binom{n_2 - 1}{k - 1}}{\binom{n_1 + n_2}{n_2}}$$, when there are even number of runs

and

$$P[U = 2k+1] = \frac{\binom{n_1 - 1}{k}\binom{n_2 - 1}{k - 1} + \binom{n_1 - 1}{k - 1}\binom{n_2 - 1}{k}}{\binom{n_1 + n_2}{n_1}}$$, when there are odd number of runs

The null hypothesis is rejected if, the observed number of runs U is too small. That is critical region is of the form

$U \le c$, where c is the constant which can be determined by using the pdf of U. If n_1 and n_2 are large, then under H_0, U is distributed as asymptotic normal with

$$E(U) = \frac{2n_1n_2}{n_1+n_2} + 1 \qquad \text{and}$$

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

and we can use the normal test

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}} \sim N(0, 1) \text{ asymptotically.}$$

Reject H_0 if $|Z| \ge Z_{\alpha/2}$ otherwise accept H_0. By referring to the standard normal N(0,1) table for the level of significance α one may draw conclusion.

Normal approximation is fairly good iff n_1 and n_2 are large say > 10 each. The run test is sensitive to both differences in location and differences in spread of the two distributions.

Eg.3. (Wald-Wolfowitz Run test) : Let the lengths of the male and female trident lynx spiders be denoted by X and Y respectively, with corresponding distributions functions F(x) and G(y). Measurements of the lengths in millimeters, of five male and five female spiders yielded the following observations as:

X : 5.40, 5.55, 6.00, 5.00, 5.70.

Y : 6.20, 6.25, 5.75, 5.85, 6.55.

Use run test to test the null hypothesis H_0 : F(x) = G(y) v/s H_1 : F(x) > G(y), take α=0.10.

Solution : Here we combine the two samples so that the combined ordered sample is

5.00, 5.40, 5.55, 5.70, 5.75, 5.85, 6.00, 6.20, 6.25, 6.55.

Runs of X : L_1=4, L_2=1

Runs of Y : L_1=2, L_2=3.

n_1=5, n_2=5,

U = No. of runs = 4(even); and k = 2, the number of runs of X's and Y's

H_0 : F(x)=G(y) v/s

H_1 : F(x) > G(y)

Under H_0 :

P[U = 2k] = 0.1269.

We use normal approximation to the above data so that

E(U) = 6 and Var(U) = 2.2222.

Under H_0 the test statistic Z is given by

$$Z = \frac{U - E(U)}{\sqrt{Var(U)}} = -1.3416$$

Since $Z < Z_\alpha = 1.282$, so we can not reject H_0 and conclude that $H_0 : F(x)=G(y)$.

iii. The Median test :

It is a statistical procedure for testing whether, two independent ordered samples differ in their central tendency or not. That is it gives the information of two independent samples are likely to be drawn from the population with the same median.

Procedure : Let $x_1, x_2, \ldots, x_{n1}$ and $y_1, y_2, \ldots, y_{n2}$ be the two independent ordered samples from the population with pdf f(x) and f(y) respectively. The measurement must atleast ordinal. Let $z_1, z_2, \ldots, z_{n1+n2}$ be the combined ordered sample. Let n_1 be the number of x's and n_2 be he umber of y's exceeding the median value M of the combined sample. Then under H_0 the samples come from the same population or from different populations with the same median.

That is under H_0: f(x) = f(y), the joint distribution of m_1 and m_2 is the hypergeometric distribution with probability function

$$P(m_1, m_2) = \frac{{}^{n_1}C_{m_1}\ {}^{n_2}C_{m_2}}{{}^{n_1+n_2}C_{m_1+m_2}} \qquad (1)$$

If $m_1 \le n_1/2$ then the critical region corresponding to the size of the type-I error α is given by $m_1 \le m'_1$ where m'_1 is computed from the equation $\sum_{m_1=1}^{m'_1} p(m_1, m_2) = \alpha$. It is proved that under H_0 the distribution of m1 is also Hypergeometric with

$$E(m_1) = \begin{cases} \frac{N}{2}, & if\ N = (n_1 + n_2)\ is\ even \\ \frac{n_1}{2}\left(\frac{N-1}{N}\right) & if\ N\ is\ Odd \end{cases}$$

and

$$\text{Var}(m_1) = \begin{cases} \dfrac{n_1 n_2}{4(N-1)}, \; if \; N \; is \; even \\ \dfrac{n_1 n_2 (N+1)}{4N^2}, if \; N \; is \; odd \end{cases}$$

Most of the time it is quite inconvenient use this distribution, however for large samples we may regard the distribution of m_1 to be asymptotically normal and one may use the 'Z' test. That is

$$Z = \frac{m_1 - E(m_1)}{\sqrt{\text{var}(m_1)}} \sim N(0, 1), \text{ asymptotically.}$$

Reject H_0 if $|Z| \geq Z_{\alpha/2}$ otherwise accept H_0. One may draw conclusion by referring to the standard normal table for the level of significance α.

χ^2 - (Chi-Square) approximation for Median test :

In the median test the observations m_1 and m_2 can be classified into (2x2) contingency table and is given by

Number of observations	Sample -1	Sample -2	Total
> Median	m_1=a	m_2=b	m_1+m_2= (a+b)
< Median	n_1-m_1=c	n_2-m_2=d	$n_1 + n_2 - m_1 - m_2$=(c+d)
Total	n_1=(a+c)	n_2=(b+d)	n_1+n_2=N=(a+b+c+d)

If the frequencies are small we can compute the exact probabilities from (1). However, if the frequencies are large we may use χ^2 **(Chi-square) -** test with 1 degree of freedom for testing H_0, and the test statistic is given by

$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \sim \chi^2 \text{ with 1 degree of freedom.}$$

However, the normal approximation is fairly good if both n_1 and n_2 exceed 10.

iv. **The Kruskal - Wallis – one way Analysis of Variance Test**:-

The Kruskal Wallis one-way analysis of variance by ranks is an extremely useful test whether to decide the 'k' independent samples are from different populations or not. The Kruskal Walli's technique tests, the null hypothesis H_0 that the 'k' samples have come from the same population or from identical population with respect to averages. i.e.

The test assumes that, the variable under study has an underlined continuous distribution. It requires at least ordinal scale of measurement.

In the computation of Kruskal Walli's test each of the 'n' observations are replaced by ranks i.e., all the scores from all of the 'k' samples combined are ranked in a single series.

The smallest score is replaced by rank 1, the next smallest by rank 2 and the largest by N where 'N' is the total no. of independent observations in the 'k' samples.

The sum of the ranks in each sample is found. It can be shown that if the 'k' samples actually are from the same population or from identical population or if H_0 is true then 'H' statistic used in the Kruskal Wallis test is approximately distributed as χ^2 with

(k-1) degrees of freedom, provided the sizes of various 'k' samples are not too small(i.e., sample sizes to be at least 5). The Kruskal - Wallis test statistic is given by

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \; - \; 3(N+1)$$

Where k is the number of samples, n = no. of observations in i^{th} sample, N = $\sum_{i=1}^{k} n_i$, total no. of observations.

R_i = sum of the ranks in the i^{th} sample.

If the observed value of H is equal to or larger than χ^2 table value at the fixed level of significance for (k-1) degrees of freedom, then H_0 will be rejected at that level of significance.

v. **Spearman's Rank Correlation Test :**

Let (X_1, Y_1), (X_2, Y_2),…, (X_n, Y_n) be a sample from a bivariate population. The sample correlation coefficient(R) between X's and Y's is defined by

$$R = \frac{\sum_{i=1}^{n}(X-\overline{X})(Y-\overline{Y})}{\sqrt{\sum_{i=1}^{n}(X-\overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y-\overline{Y})^2}} \qquad (1)$$

If the sample values $X_1, X_2, ,\ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ are each ranked from 1,2, …, n in increasing order of magnitude separately, and if the X's and Y's have continuous df's we get a unique set of rankings. Thus

$$\sum_{i=1}^{n} X_i = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

and $\overline{X} = \overline{Y} = \frac{(n+1)}{2}$.

Also, we have

$$\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{n(n^2-1)}{12} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 .$$

$$\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{2}\left[\frac{n(n^2-1)}{6} - \sum_{i=1}^{n} d_i^2\right],$$

where $d_i = \text{rank}(X_i) - \text{rank}(Y_i) = R_x - R_y$

Thus under H_0, the random variables X's and Y's are independent, so that their ranks are also independent and then the statistic R is defined as

$$R = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

where R is called as the Spearman's rank correlation coefficient between the ranks of X's and Y's. For large samples it is possible to use normal test so that under H_0, the statistic

$Z = R\sqrt{n-1}$ has approximately a standard normal distribution and we should reject H_0

if $|Z| > Z_\alpha$ where Z_α is a critical value of Z at prefixed α level of significance. Normal approximation is fairly good iff n is large, say ≥ 10.

5. Summary :

We have learnt several non-parametric test procedures and their applications. These tests could be applied when strict assumptions about the form of the distribution from which we are sampling do not hold, i.e. especially when data are non normal and the data are of often nominal or ordinal. In other words when parametric assumptions do not hold we could use non-parametric tests as their alternative.