NONPARAMETRIC INFERENCE

BY

TANUJIT CHAKRABORTY

Indian Statistical Institute

Mail : tanujitisi@gmail.com

NON-PARAMETRIC INFERENCE

In many cases an experimentor doesn't know the form of the basic distn. and needs statistical techniques which are applicable regardless of the form of the density function. This technique is called non-parametric/ distribution-free method.

Estimation:- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with CDF F which is unknown. The family of distn. consists of absolutely continuous or discrete distribution.

Theorem:- 1. The order statistic $(X_{(1)}, X_{(2)}, X_{(3)}, \ldots, X_{(n)})$ is minimal sufficient for population distribution.

2. Any Unbiased estimator of $\psi(\theta)$ based on order statistic is unique and UMVUE of $\psi(\theta)$.

Example 1:- Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution function F (unknown). Find the UMVUE of $\mu(F)$ and $\sigma^2(F)$.

Solution:- $\bar{X}$ is an unbiased estimator of $\mu(F)$ for any F.

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}$$ is a function of $(X_{(1)}, X_{(2)}, \ldots, X_{(n)})$ complete sufficient statistic.

$\bar{X}$ is the UMVUE of $\mu(F)$.

$$S^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$ is an unbiased estimator of $\sigma^2(F)$ for any F.

$$\therefore\ S^2 = \frac{1}{2n(n-1)}\sum\sum_{i \neq j} \left(X_{(i)} - X_{(j)}\right)^2$$ is a function of $(X_{(1)}, \ldots, X_{(n)})$

So, S^2 is the UMVUE of $\sigma^2(F)$.

Example 2:- Let $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ be independent random variables from two absolutely continuous distribution functions. Find the UMVUE's of (i) $E(XY)$ (ii) $V(X+Y)$.

Solution:- $\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$ is the UMVUE of $E(X)$.

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ is the UMVUE of $E(Y)$.

$\Rightarrow$ $\bar{X}\bar{Y}$ is the UMVUE of $E(X)E(Y) = E(XY)$ due to independence.

$\Rightarrow$ $\frac{1}{mn}\left(\sum X_i\right)\left(\sum Y_i\right)$ is the UMVUE of $E(XY)$.

Now, $s_1^2 = \frac{1}{m-1}\sum_{1}^{m}(X_i - \bar{X})^2$ is the UMVUE of $V(X)$

$s_2^2 = \frac{1}{n-1}\sum_{1}^{n}(Y_i - \bar{Y})^2$ is the UMVUE of $V(Y)$

So, $s_1^2 + s_2^2$ is the UMVUE of $V(X)+V(Y) = V(X+Y)$.

■ <u>Testing of Hypothesis</u> :- A test of a hypothesis H_0, based on a statistic T, whose distribution, under H_0 does not depend on the specified distribution or any parameter of that distribution, is called a 'distribution-free' or 'Non-parametric' test.

(I) <u>Single Sample Problem</u> :-

<u>Problem of location</u> :- Let $X_1, \ldots, X_n$ be a sample of size n from some unknown CDF $F_X(x)$. We assume that $F_X(x)$ is absolutely continuous. Here an appropriate measure of location is median or the pth quantile. Let ξ_p be the pth quantile of $F_X(x)$. To test $H_0: \xi_p = \xi_0$, we consider sign test.

● <u>SIGN TEST</u> :- Let $(X_1, \ldots, X_n)$ be a random sample from a pdf $f_X(x)$.

To test $H_0: \xi_p = \xi_0$ Vs. $H_1: \xi_p > \xi_0$.

Let Z denotes the number of $(X_i - \xi_0)$ that are positive $(i = 1(1)n)$

Note that $P[X_i - \xi_0 = 0] = 0$.

Under H_0, the statistic Z can be thought of as the number of successes in n independent bernoullian trials.

$$\Rightarrow Z \sim Bin(n, 1-p), \text{ where } P_{H_0}[X_i - \xi_0 > 0]$$
$$= P_{H_0}[X > \xi_0] = 1 - p.$$

Under H_0, $E(Z) = n(1-p)$.

∴ One should expect Z to be near $n(1-p)$.

Hence, an intuitively appealing test is reject H_0 iff

$$Z - n(1-p) \geq K$$

$\Leftrightarrow Z \geq c$, where c is such that $P_{H_0}(Z \geq c) = \alpha$.

The level α critical region is given by;

$W = \{Z : Z \geq c\}$, where $P_{H_0}\{Z \geq c\} = \alpha$.

Alternatively, we can compute the p-value,

$p = P_{H_0}(Z \geq Z_0)$, where Z_0 is the observed value of Z.

If $p < \alpha$, we reject H_0 in favour of H_1.

Remark:-

1. The sign test of $H_0: \xi_p = \xi_0$ Vs. $H_1: \xi_p \neq \xi_0$ is given by $Z \le c_1$ or $Z \ge c_2$, where $P_{H_0}\{Z \le c_1 \text{ or } Z \ge c_2\} = \alpha$.
Otherwise we can compute the p-value.

2. It can be shown that a level α UMP test of $H_0: \xi_p = \xi_0$ Vs. $H_1: \xi_p > \xi_0$ based on Z is given by

$$\phi(z) = \begin{cases} 1 & \text{if } z > c \\ \gamma & \text{if } z = c \\ 0 & \text{if } z < c \end{cases} \quad \text{where } E_{H_0}\{\phi(Z)\} = \alpha.$$

Hence the sign test given by $W = \{z : z \ge c\}$ is a UMP test of $H_0: \xi_p = \xi_0$ Vs. $H_1: \xi_p > \xi_0$ of its size.

3. <u>Asymptotic Sign Test</u> :- For large n,

$$\frac{Z - n(1-p)}{\sqrt{np(1-p)}} \overset{a}{\sim} N(0,1), \text{ under } H_0.$$

Hence,
$$\alpha = P_{H_0}\{Z \ge c\} = P_{H_0}\left\{\frac{Z - n(1-p)}{\sqrt{np(1-p)}} \ge \frac{c - n(1-p)}{\sqrt{np(1-p)}}\right\}$$
$$= P\left\{\tau \ge \frac{c - n(1-p)}{\sqrt{np(1-p)}}\right\}; \ \tau \sim N(0,1)$$

$$\Rightarrow \frac{c - n(1-p)}{\sqrt{np(1-p)}} \simeq \tau_\alpha .$$

Hence an asymptotic sign test of $H_0: \xi_p = \xi_0$ Vs. $H_1: \xi_p > \xi_0$ is to reject H_0 iff $Z \ge n(1-p) + \tau_\alpha \cdot \sqrt{np(1-p)}$.

4. Sign test for a sample from bivariate population (paired sample): Let (X_i, Y_i), $i = 1(1)n$ be a paired sample. Let $D_i = X_i - Y_i$ and assume that D_i has an absolutely continuous distribution. We are interested in the location of the distribution of D_i's.

To test $H_0: \xi_p(D) = \xi_0$

H_0 can be tested using sign test based on $D_1, D_2, \ldots, D_n$.

Note that, $\xi_p(D) \neq \xi_p(X) - \xi_p(Y)$

$$\xi_{1/2}(D) \not\Rightarrow \xi_{1/2}(X) = \xi_{1/2}(Y).$$

- **Wilcoxan Signed Rank Test:-** The sign test for ξ_p loses information as it ignores the magnitude of the deviation $(X_i - \xi_0)$'s and considers only the signs. Hence, a test can be provided that also takes into account, the magnitude of these deviations and this improvement is provided in Wilcoxon's signed rank test.

Let $X_1, X_2, \ldots, X_n$ be a random sample from a pdf $f(x)$ which is unknown.

To test $H_0: \xi_{1/2} = \xi_0$.

In all such cases, WLG, take $\xi_0 = 0$.

Hence, our condition on $F(x)$ becomes $F(-x) + F(x) = 1$.

The testing problem reduces to $H_0: \xi_{1/2} = 0$.

We proceed by first ranking $|x_1|, \ldots, |x_n|$ and keeping track of the original sign of x_i. Let R_i be the rank of $|X_i|$ $\forall\, i = 1(1)n$ and

$$Z_i = \begin{cases} -1 & \text{if } x_i < 0 \\ 1 & \text{if } x_i > 0 \end{cases}$$

Note that $P[X_i = 0] = 0$.

The statistic $W = \sum_{i=1}^{n} Z_i R_i$ is the Wilcoxon statistic.

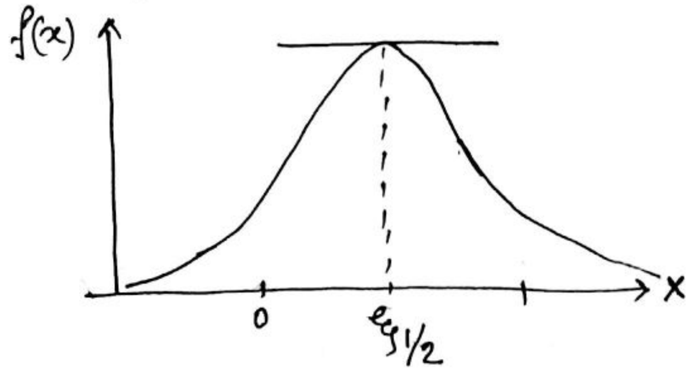
[If T^+ : sum of the ranks of the $(+)$ X_i's.
T^- : sum of the ranks of the $(-)$ X_i's.

Clearly $T^+ + T^- = \frac{n(n+1)}{2}$ and $W = T^+ - T^- = 2\left\{T^+ - \frac{n(n+1)}{4}\right\}$

or $2\left\{\frac{n(n+1)}{4} - T^-\right\}$

Here, W, T^+ or T^- are linearly related.]

A large $(+)$ value of W indicates the most of the large deviations from zero are $(+)$ and the number of $(+)$ signs is also large.



Then we suspect $H_0: \xi_{1/2} = 0$ and support H_1.

$\therefore$ We reject H_0 in favour of $H_1: \xi_{1/2} > 0$ at level α iff the observed value of $W \geq c$, where $P_{H_0}\{W \geq c\} = \alpha$.

Alternatively, the p-value $P_{H_0}\{W \geq W_0\}$ can be computed.

Distribution of W under H_0 (Null distribution of W) :- To compute probabilities like $P\{W \geq c\}$, $P_{H_0}\{W \leq c_1\}$, etc. of W under H_0, when We need to determine the distribution of $H_0: \xi_{1/2} = 0$ or $F(0) = 1/2$ is true, we note the following facts:

(i) The assumption that $F(-x) = 1 - F(x)$ ensures that $P\{X_i < 0\} = P\{X_i > 0\} = \frac{1}{2}$ $\forall\, i = 1(1)n$. Hence we have,

$P\{Z_i = -1\} = P\{Z_i = +1\} = \frac{1}{2}$ $\forall\, i = 1(1)n$.

Moreover, Z_i's are all i.i.d. as X_i's are all iid.

(ii) Due to symmetry the rank R_i of $|X_i|$ doesn't depend on the sign Z_i of X_i, $i = 1(1)n$. Hence R_i's are stochastically independent of Z_i's.

Write $W = \sum Z_i R_i = \sum V_i$, where $V_1, V_2, \ldots, V_n$ is one and only one of $Z_1R_1, \ldots, Z_nR_n$ such that

$P\{V_i = -i\} = P\{V_i = i\} = \frac{1}{2}$ and V_i's are independent

Exact Distribution :- The MGF of W is

$$\begin{aligned} M_W(t) &= E\left(e^{tW}\right) \\ &= E\left(e^{t\sum_i V_i}\right) \\ &= \prod_{i=1}^{n} E\left[e^{tV_i}\right] \\ &= \prod_{i=1}^{n} \frac{e^{-ti} + e^{ti}}{2} \\ &= M_{(-W)}(t). \end{aligned}$$

Hence W and $-W$ have the same distribution.

$\Rightarrow$ W is symmetric about '0'.

Now, $P_{H_0}\{W = i\}$ = the coefficient of e^{ti} in the expansion of $M_W(t)$.

Asymptotic Distribution:- Under H_0, $E(V_i)=0$ and $Var(V_i)=i^2$, $i=1(1)n$.

$$E(W)=\sum E(V_i)=0$$

$$Var(W)=\sum Var(V_i)=\sum i^2=\frac{n(n+1)(2n+1)}{6}$$

By Liapunov's CLT, $\dfrac{W-0}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \sim N(0,1)$, under H_0 for large n.

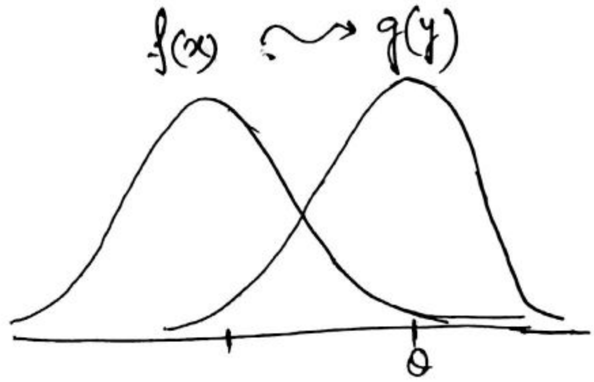
For large n,
$$\alpha = P_{H_0}\{W \geq c\}$$
$$= P_{H_0}\left\{\frac{W}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \geq \frac{c}{\sqrt{\frac{n(n+1)(2n+1)}{6}}}\right\}$$

(II) **Two Samples Problems:-** Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent samples from two univariate absolutely continuous distribution function $F(x)$ and $G(y)$.

To test $H_0: F(u)=G(u)\ \forall\ u\in\mathbb{R}$

against the usual one and two-sided alternatives.

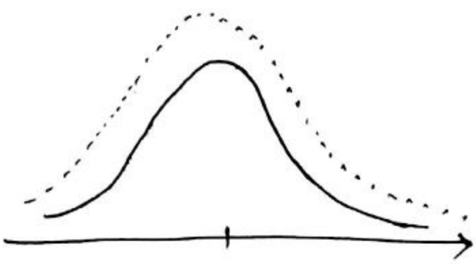
1. **Location Alternative:-**



$$F(x)=G(x-\theta)$$

Then H_0 reduces to $H_0: \theta=0$.

2. **Scale Alternative:-**



$$F(x)=G\left(\frac{x}{\sigma}\right),\ \sigma>0$$

Then H_0 reduces to $H_0: \sigma=1$.

3. **General Alternative:-**

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

We first consider a simple test for location.

—

- Median Test :- Combine the two samples into one sample $(m+n)$ and order the $(m+n)$ values in ascending order of magnitude.

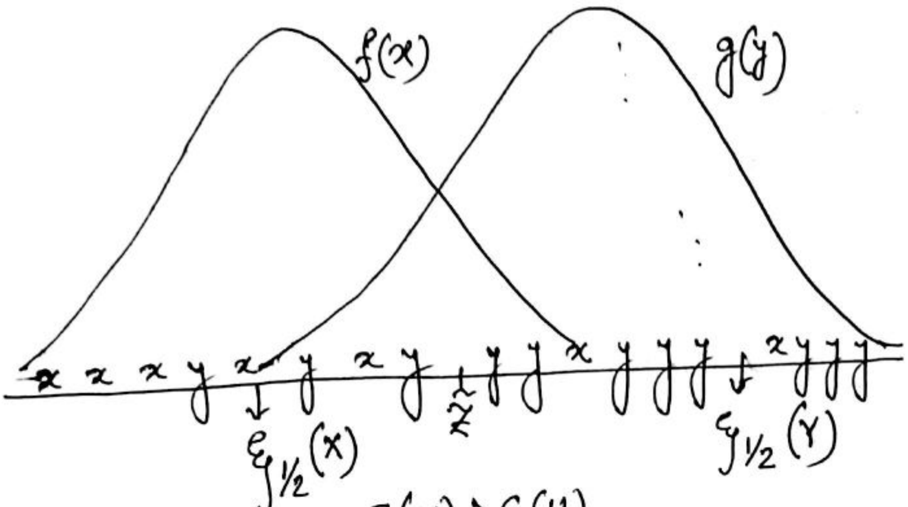
Let $Z_1 < Z_2 < \cdots < Z_{m+n}$.

Let $\tilde{Z}$ be the median of the combined sample.

Let V be the number of X_i's which are $< \tilde{Z}$ in the combined sample.

If the value of V is quite large one might suspect that

$$\xi_{1/2}(X) < \xi_{1/2}(Y)$$



$$F(U) > G(U).$$

Hence, we reject $H_0: F(U) = G(U)\ \forall\ U \in \mathbb{R}$ in favour of
$H_1: F(U) \geq G(U)$ but $F(U) > G(U)$ for some U if

V is quite large, i.e. $V \geq c$.

If median of X and Y is equal, then cdf of X and Y is not equal.

Here c is such that $P_{H_0}[V \geq c] = \alpha$ or one can compute the p-value, $P_{H_0}[V \geq v_0]$, where v_0 is the observed value of V.

This is called Median Test.

Limitation / Difficulties :- The median test will tend to accept $H_0: F(U) = G(U)\ \forall\ U \in \mathbb{R}$, even if the shapes of $F(\cdot)$ and $G(\cdot)$ are different as long as their medians are equal.

Scanned by CamScanner

Null Distribution of V [Distribution of V under H_0] :-

To compute $P_{H_0}[V = v] = P_{H_0}$[exactly v of the X_i's are $< \tilde{z}$ in the combined sample]

Case-I :- $m+n = \text{even} = 2p$, say

There are exactly $p = \frac{m+n}{2}$ values $< \tilde{z}$ in the combined sample. and these p values can be selected in $\binom{m+n}{p}$ ways. Each of the cases has the same probabilities under H_0: $F(u) = G(u)$ $\forall\, u \in \mathbb{R}$.

For the favourable cases, exactly v of the m values of X (and hence $p-v$ of the n values of Y) are $< \tilde{z}$ and the number of such cases is $\binom{m}{v}\binom{n}{p-v}$.

Hence,
$$P_{H_0}[V = v] = \begin{cases} \dfrac{\binom{m}{v}\binom{n}{p-v}}{\binom{m+n}{p}} & ; \; v = 0, 1, 2, \ldots, \min(m, p) \\ 0 & ; \; \text{OW} \end{cases}$$

Case-II :- $m+n = \text{odd} = 2p+1$.

Here $(p+1)^{th}$ value is the median of the combined sample.

Now, $P_{H_0}[V = v] = P_{H_0}$[exactly v of the m values of X are below $\tilde{z}$ of the $(p+1)^{th}$ value]

Hence,
$$P_{H_0}[V = v] = \begin{cases} \dfrac{\binom{m}{v}\binom{n}{p-v}}{\binom{m+n}{p}} & ; \; v = 0, 1, 2, \ldots, \min(m, p) \\ 0 & ; \; \text{OW} \end{cases}$$

Asymptotic Distribution :- V has Hypergeometric distribution,

$$E(V) = p \cdot \frac{m}{m+n}, \quad V(V) = p \cdot \frac{m}{m+n} \cdot \frac{n}{m+n} \cdot \frac{m+n-p}{m+n-1}.$$

where $p = \left[\frac{m+n}{2}\right]$ under H_0.

For large m, n; $E(V) \approx \frac{m}{2}$; $V(V) \approx \frac{mn}{4(m+n)}$.

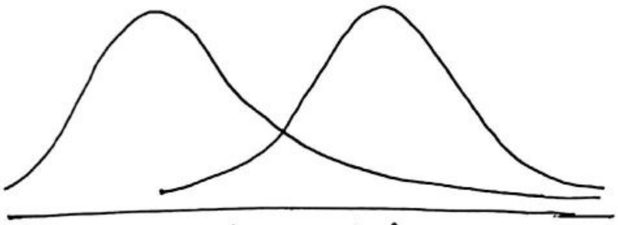
$$\therefore \quad \frac{V - \frac{m}{2}}{\sqrt{\frac{mn}{4(m+n)}}} \overset{a}{\sim} N(0,1) \text{ under } H_0.$$

- <u>Wald Wolfowitz Run Test</u> :— Let $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ be independent random samples from an absolutely continuous distribution functions F and G. It is a simple test of the hypothesis $H_0: F(Z) = G(Z)\ \forall z$ based on the notation of runs of the values of X and the values of Y. We shall now explain what we mean by runs.

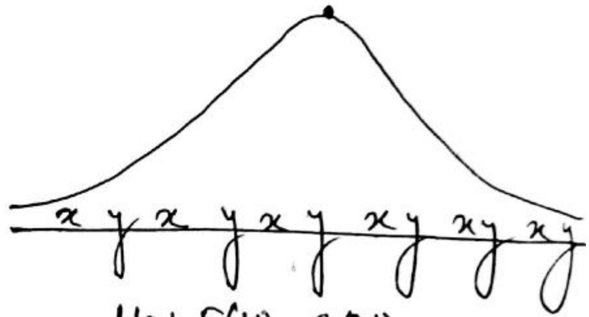
[For example, if $m=4$ and $n=5$, one might obtained: yxxyxyyyx. A run is a sequence of letters of the same kind founded by letters of another kind except for the first and last position. In our example, there are total of $3+3 = 6$ runs.

" Of what can runs be suggested ?"

— Suppose that with $m=7$, $n=8$, we have the following ordering: xxxxxyxxyyyyyyy



$F(U) > G(U)$



$H_0: F(U) = G(U)$

To us this strongly suggests that $F(U) > G(U)$. For if $F(U) = G(U)\ \forall\ U \in \mathbb{R}$, we would anticipate a greater number of runs.]

Let us combine the sample of m values of X and the sample of n values of Y. There are one collection of $(m+n)$ ordered values arranged in accending order of magnitude.

It is obvious that, if the two samples are taken from the same population, the X_i's and Y_i's will be ordinary and well mixed and number of runs will be large.

In general, differences between two population will tend to reduce the number of runs. Let R be the number of runs in the combined samples. A test is then performed by observing R and rejecting $H_0 : F(U) = G(U) \ \forall \ U$, if $R \le c$, (R is small)
The constant 'c' is determined from the restriction $P_{H_0}[R \le c] = \alpha$

Otherwise, p-value $= P_{H_0}[R \le r_0]$, where r_0 is the observed value of R, that can be computed.

Null Distribution of R [The distn. of R under H_0] :—

Note that, we can select m places for m values of X [and n places for n values from Y] from $(m+n)$ values in $\binom{m+n}{m}$ ways, under H_0. The all possible arrangement of m values of X and n values of Y are equally probable. To find $P_{H_0}[R = r]$

Case – I :– $r = 2K+1 =$ odd, this means that —

{ There must be $K+1$ runs of the order values of X and K runs of the order values of Y.

or { K runs of order value of X and $(K+1)$ runs of the order value of Y.

To get $(K+1)$ runs of the m values of X, we have to insert K dividers into the $(m-1)$ spaces between the m values of X and these can be done in $\binom{m-1}{K}$ ways.

Similarly, K runs of n values of Y, we have $\binom{n-K}{K-1}$ ways.

Hence, $$P_{H_0}[R = r = 2K+1] = \frac{\binom{m-1}{K}\binom{n-1}{K-1} + \binom{m-1}{K-1}\binom{n-1}{K}}{\binom{m+n}{m}}.$$

Case – II :– $r = 2K =$ even.

Here, the ordered values X and Y must have K runs each. We may begin with either run of the values of X or run of the values of Y.

Hence, $$P_{H_0}[R = r = 2K] = \frac{2\binom{m-1}{K-1}\binom{n-1}{K-1}}{\binom{m+n}{n}}.$$

Asymptotic Distribution:-

It can be shown that, $E(R) = 1 + \frac{2mn}{m+n} = \mu$, say, and

$$V(R) = \frac{2mn(2mn - m - n)}{(m+n)^2(m+n-1)} = \frac{\frac{2mn}{m+n}\left(\frac{2mn}{m+n} - 1\right)}{(m+n-1)}$$

$$= \frac{(\mu-1)(\mu-2)}{(m+n-1)} = \sigma^2, \text{ say.}$$

The distribution of R can be approximated with large sample sizes m and n, by a normal distn. with mean μ and variance σ^2, i.e., $\frac{R-\mu}{\sigma} \overset{a}{\sim} N(0,1)$.

In fact, this variance approximation is good enough for practical purposes when both m and n exceed 10.

Run test as a Test for Randomness:- Run test can be used as a check to see if it is reasonable to treat $X_1, X_2, \ldots, X_K$ as a random sample of size K from some continuous distribution. We are given that the K values of X to be the observed values $x_1, x_2, \ldots, x_K$ which are not ordered by magnitude but by order in which they are observed.

In the sequence, $x_1, x_2, \ldots, x_K$ replace each value that is below the sample median B and each value that is above the sample median A, example, K = 10 (let), sequence is such as

BBBBABAAAA

may suggest towards increasing value of X, i.e., these values of X may not reasonably be looked upon as a r.s.. If the trend is the only alternative to randomness, then we reject the null hypothesis of randomness. In favour of alternative hypothesis of trend if $R \le c$.

To make this test, we would use the pmf of R with

$$m = n = \frac{K}{2} \; ; \; K = \text{even}.$$

- **Mann-Whitney-Wilcoxon Test :-** Let $(X_1, X_2, \ldots, X_m)$ and $(Y_1, Y_2, \ldots, Y_n)$ be independent random samples from continuous distribution functions F and G, respectively.

Hypothesis of MWW Test is

$H_0: F(z) = G(z) \;\forall\, z$

Let us define, $Z_{ij} = \begin{cases} 1 & \text{if } x_i < y_j \\ 0 & \text{o/w} \end{cases}$

where $Z_{ij} = (X_i, Y_j) = (X_1, Y_1), \ldots, (X_n, Y_n)$

and the test statistic $U = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} Z_{ij} \right) = \sum_{j=1}^{n} U_j$,

We note that, $U_j = \sum_{j=1}^{m} Z_{ij}$ counts the number of X_i that less than Y_j, $j = 1(1)n$.

Thus U is the sum of these m counts.

The statistic U is called Mann-Whitney-Wicoxon Test statistic.

Clearly, $U = 0$ iff all the X_i's are larger than all the Y_j's.

and $U = mn$ iff all the X_i's are smaller than all the Y_j's.

If U is large, then the values of Y tend to be larger than the values of X and this supports the alternative:

$H_1: F(U) > G(U) \;\forall\, U$ and $F(U) > G(U)$ for some U.

On the other hand, a small values of U supports $H_1: F(U) < G(U) \;\forall\, U$.

H_0	H_1	Critical Region
$F = G$	$F \geq G$	$U \geq c_1$
	$F \leq G$	$U \leq c_2$
	$F \neq G$	$U \leq c_3$ or $U \geq c_4$

To determine the critical value or the p-value, we need the distribution of U, under H_0.

The null distribution of U:- Let, $P_{H_0}[U=u] = P_{m;n}(U)$

If the observations are arranged in increasing order of magnitude, the largest value can be either x value or y value, under H_0. The place can be filled out any one of $(m+n)$ equally likely ways, m of which are favourable to X values, n of which are favourable to Y values.

Hence the prob. that an arrangement ends with X values $= \frac{m}{m+n}$ and it ends with Y values $\frac{n}{m+n}$.

$$P_{m;n}(U) = P_{H_0}[U=u] = P_{H_0}[U=U \mid \text{the largest value of } X] \times P[\text{the largest value of } X] + P_{H_0}[U=u \mid \text{the largest value of } Y] \times P[\text{the largest value of } Y]$$

$$= P_{m-1,n}(U)\cdot\frac{m}{m+n} + P_{m,n-1}(U-m)\cdot\frac{n}{m+n}.$$

Here if the largest value is X, it does not contribute to U and the remaining $m-1$ values of X, n values of Y can be arranged to produce $U=u$ with probability $P_{m-1,n}(U)$.

If the largest value is Y, then this Y values is greater than the m values of X and the remaining $(n-1)$ values of Y, m values of X to contribute $U' = U-m$ with prob. $P_{m,n-1}(U-m)$

Asymptotic Null Distribution of U:-

Under H_0: $G(U) = F(U)\ \forall\ U$ [i.e., $X_1,\dots,X_m$ & $Y_1, Y_2, \dots, Y_n$ are from same population]

i) $P[X_i < Y_j] = \frac{1}{2} = P[X_i > Y_j]$

ii) $P[X_i < Y_j, X_i < Y_K] = \frac{2!}{3!} = \frac{1}{3}\ \forall\ j \neq K$

iii) $P[X_i < Y_j, X_n < Y_j] = \frac{2!}{3!} = \frac{1}{3}\ \forall\ i \neq n$

iv) $P[X_i < Y_j, X_n < Y_K] = P[X_i < Y_j]\,P[X_n < Y_K]$

$= \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}\ \forall\ i \neq n, j \neq K.$

Now, $Z_{ij} = \begin{cases} 1 & \text{if } X_i < Y_j \\ 0 & \text{ow} \end{cases}$

So, $E(Z_{ij}) = 1 . P[X_i < Y_j] = \frac{1}{2}$

So, $E(U) = \sum_{i=1}^{m} \sum_{j=1}^{n} E(Z_{ij}) = \frac{mn}{2} = \mu .$

Now, $E(U^2) = E\left(\sum_{i=1}^{m} \sum_{j=1}^{n} Z_{ij}\right)\left(\sum_{h=1}^{m} \sum_{k=1}^{n} Z_{hk}\right)$

$$= E\left[\sum_{i=1}^{m}\sum_{j=1}^{n} Z_{ij}^2\right] + E\left[\sum\sum_{j \neq k} \sum_{i=1}^{m} Z_{ij} Z_{ik}\right]$$

$$+ E\left[\sum_{j=1}^{n} \sum\sum_{h \neq i} Z_{ij} Z_{hj}\right] + E\left[\sum\sum_{j \neq k} \sum\sum_{h \neq i} Z_{ij} Z_{hk}\right]$$

$$= \frac{mn}{2} + \frac{mn(n-1)}{3} + mn(m-1) + m(m-1)(n-1).n.\frac{1}{4}$$

When Ho is true, we know that X_i, X_h, X_j, X_k, $i \neq h, j \neq k$ are mutually stochastically independent and have same distribution of continuous type. Moreover, $P[X_i < Y_j, X_i < Y_k] = \frac{1}{3}$.

$E(Z_{ij} Z_{ik}) = P[X_i < Y_j, X_i < Y_k] = \frac{1}{3}$, $j \neq k$

$E(Z_{ij} Z_{hj}) = P[X_i < Y_j, X_h < Y_j] = \frac{1}{3}$, $i \neq h$

$E(Z_{ij} Z_{hk}) = P[X_i < Y_j, X_h < Y_k] = \frac{1}{4}$, $i \neq h, j \neq k$.

$E(Z_{ij}^2) = 1^2 P[X_i < Y_j] = \frac{1}{2}$.

$$V(U) = E(U^2) - E^2(U) = mn\left[\frac{1}{2} + \frac{n-1}{3} + \frac{m-1}{3} + \frac{n-1}{2}\cdot\frac{m-1}{2} - \frac{m}{2}\cdot\frac{n}{2}\right]$$

$$= \frac{1}{12} mn (m+n+1).$$

If m, n are both large enough, it can be shown that —

$$\frac{U - \frac{mn}{2}}{\sqrt{\frac{mn}{2}(m+n+1)}} \overset{a}{\sim} N(0,1), \text{ under Ho}.$$

This approximation is fairly good for $m, n \geq 8$.