A Handout on Statistical Approach to

Nonparametric Methods

By Ankan Chakraborty, Akash Das, Anubhab Biswas and Debanjan Dutta

Presidency University, Kolkata; November 2020

Non-Parametric Methods

Presidency University

November 17, 2020

Ankan Chakraborty

Akash Das

Anubhab Biswas

Debanjan Dutta

Table of Contents

A. BINOMIAL TEST

Data: We observe the outcomes of $\boldsymbol{n}$ independent repeated Bernoulli Trials.

Assumptions:

1. The outcome of each trial can be classified as a **success** or a **failure**.
2. The **probability of a success**, denoted by $\boldsymbol{p}$, remains **constant** from trial to trial.
3. The $\boldsymbol{n}$ trials are **independent**.

Target: To make inference about "$\boldsymbol{p}$".

1) To Test:

$$\begin{cases} \boldsymbol{H_0:} & \boldsymbol{p = p_0} \\ \boldsymbol{Vs.} & \boldsymbol{,where\ 0 < p_0 < 1} \\ \boldsymbol{H_0:} & \boldsymbol{p > p_0} \end{cases}$$

Let $\boldsymbol{B}$ =No. of successes.

We may use $\boldsymbol{B}$ as our test statistic because the statistic $\dfrac{\boldsymbol{B}}{\boldsymbol{n}}$ is an estimator of the True unknown parameter $\boldsymbol{p}$. Thus, if $\boldsymbol{p > p_0}$, $\dfrac{\boldsymbol{B}}{\boldsymbol{n}}$ will tend to be larger than $\boldsymbol{p_0}$. thus suggests rejecting $\boldsymbol{H_0 : p > p_0}$ in favoir of $\boldsymbol{p > p_0}$ for large values of $\boldsymbol{B}$.

▶ **Exact Distribution of $\boldsymbol{B}$:**

$$\boldsymbol{B = \sum_{i=1}^{n} d_i}$$

where, $\boldsymbol{d_i} = \begin{cases} \boldsymbol{1} & \boldsymbol{,if\ the\ i^{th}\ Bernoulli\ trial\ is\ a\ success} \\ \boldsymbol{0} & \boldsymbol{,if\ the\ i^{th}\ Bernoulli\ trial\ is\ a\ failure} \end{cases}$

Total no. of possible outcomes $\boldsymbol{(d_1, d_2, d_3, \ldots, d_n)}$ is $\boldsymbol{= 2^n}$.

Any outcomes with $\boldsymbol{b}$ times **1's** and $\boldsymbol{(n - b)}$ times **0's** has probability $\boldsymbol{p^b(1-p)^{n-b}}$.

[Here the position of **1's** and **0's** are fixed]

$$\boldsymbol{P_p[B = b]} = \begin{cases} \boldsymbol{\binom{n}{b} p^b(1-p)^{n-b}} & \boldsymbol{,for\ b = 0, 1, 2, \ldots, n} \\ \boldsymbol{0} & \boldsymbol{,otherwise} \end{cases}$$

$$\begin{aligned} \boldsymbol{E_p(B)} &= \boldsymbol{E_p(\sum_{i=1}^{n} d_i)} \\ &= \boldsymbol{\sum_{i=1}^{n} E_p(d_i)} \\ &= \boldsymbol{np} \end{aligned}$$

Also it can be shown that,

$$\boldsymbol{Var_p(B) = np(1-p)}$$

► **The Asymptotic Distribution of B:**

The random variable $\boldsymbol{B}$ is a sum of independent and identically distributed random variables and hence the central limit theorem establishes that as $\boldsymbol{n \to \infty}$, $\boldsymbol{\frac{B - np}{\sqrt{np(1-p)}}}$ has a limiting $\boldsymbol{N(0,1)}$ distribution.

► **Testing Criterion:**

Reject $\boldsymbol{H_0}$ at the $\boldsymbol{\alpha}$ level of significance if $\boldsymbol{B \geq b_\alpha}$; otherwise do not reject, where the constant $\boldsymbol{b_\alpha}$ is chosen such that,

$$\begin{aligned} \boldsymbol{P(B \geq b_\alpha | H_0) = \alpha} \\ \boldsymbol{i.e.\ P(B \geq b_\alpha | B \sim Bin(n, p_0)) = \alpha} \quad \boldsymbol{\cdots\cdots\cdots (1)} \end{aligned}$$

i.e. $\boldsymbol{b_\alpha}$ is the upper $\boldsymbol{\alpha}$ percentile point of the binomial distribution with sample size $\boldsymbol{n}$ and success probability $\boldsymbol{p}$.

Due to the discreteness of binomial distribution, not all the values of $\boldsymbol{\alpha}$ are available i.e. we can't always find a $\boldsymbol{b_\alpha}$which satisfies **(1)** for any arbitrary choice of $\boldsymbol{\alpha}$.

So, we focus on finding a $\boldsymbol{b_\alpha}$ such that

$$\boldsymbol{P(B \geq b_\alpha | H_0) \leq \alpha} \qquad \boldsymbol{\cdots\cdots\cdots (2)}$$

[choose the smallest $\boldsymbol{b_\alpha}$which satisfies **(2)** **]**

This is a **one-sided upper tail test**.

2) To Test:

$$\begin{cases} \boldsymbol{H_0 :} & \boldsymbol{p = p_0} \\ \boldsymbol{Vs.} & \boldsymbol{,where\ 0 < p_0 < 1} \\ \boldsymbol{H_0 :} & \boldsymbol{p < p_0} \end{cases}$$

► **Testing Criterion:**

Reject $\boldsymbol{H_0}$ at the $\boldsymbol{\alpha}$ level of significance in favour of $\boldsymbol{H_1}$ if $\boldsymbol{B \leq c_\alpha}$; otherwise do not reject, where $\boldsymbol{c_\alpha}$ is chosen such that,

$$\boldsymbol{P(B \leq c_\alpha | H_0) = \alpha}$$
$$\boldsymbol{i.e.\ P(B \leq c_\alpha | B \sim Bin(n, p_0)) = \alpha} \quad \cdots\cdots\cdots (3)$$

i.e. $\boldsymbol{b_\alpha}$ is the upper $\boldsymbol{\alpha}$ percentile point of the binomial distribution with sample size $\boldsymbol{n}$ and success probability $\boldsymbol{p}$.

Again, due to some discreteness issue, we focus on the following rather than **(3)**, i.e. choose $\boldsymbol{c_\alpha}$ such that

$$\boldsymbol{P(B \leq c_\alpha | H_0) \leq \alpha} \qquad \cdots\cdots\cdots (4)$$

[choose the largest $\boldsymbol{c_\alpha}$which satisfies **(4)]**

- If $\boldsymbol{p_0 = \frac{1}{2}}$, then $\boldsymbol{c_\alpha = n - b_\alpha}$.

This is a **one-sided lower tail test**.

3) Similary for testing,

$$\begin{cases} \boldsymbol{H_0 :} & \boldsymbol{p = p_0} \\ \boldsymbol{Vs.} & \boldsymbol{,where\ 0 < p_0 < 1} \\ \boldsymbol{H_0 :} & \boldsymbol{p \neq p_0} \end{cases}$$

Reject $\boldsymbol{H_0}$ at the $\boldsymbol{\alpha}$ level of significance if $\boldsymbol{B \geq b_{\alpha_1}}$ or $\boldsymbol{B \leq c_{\alpha_2}}$; otherwise do not reject.
Where $\boldsymbol{b_{\alpha_1}}$ is the upper $\boldsymbol{\alpha_1}$ percentile point and $\boldsymbol{c_{\alpha_2}}$ is the lower $\boldsymbol{\alpha_2}$ percentile point and $\boldsymbol{\alpha_1 + \alpha_2 = \alpha}$.

Remarks :

1. Binomial test is a distribution free test.

   Reason : Apart from the mild assumptions **(1) - (3)** , the probability distribution of $\boldsymbol{B}$ does not depend on the underlying population from which the dichotomous data comes.

Applications :

Target: To test hypothesis about the unknown median $\boldsymbol{\theta}$, of a population. The application of binomial theory to this problem leads to a test statistic $\boldsymbol{B}$, that counts the number of sample observation larger than a specified null hypothesis value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta_0}$.

For this particular special case, the statistic $\boldsymbol{B}$ is referred to as the **sign statistic** and the associated test procedures are referred to as **sign test** procedures.

B. Wilcoxon Signed-Rank Test

The ordinary sign test does not make use of the magnitude of the difference between the observed value and the assumed value of the quantile. The Wilcoxon signed rank test provides an alternative test of location by taking into account the magnitude of the difference as well as their sign and as such is more efficient than ordinary sign test.

▶ Kinds of data we deal with :

1. Paired replicates data represent pair of "**pre-treatment**" and "**post-treatment**" observations; here we are concerned with a shift in location due to the application of the "**treatment**".
2. One sample data, counts of observations from a single population whose location we wish to make inferences.

▶ Paired Replicates Analysis :

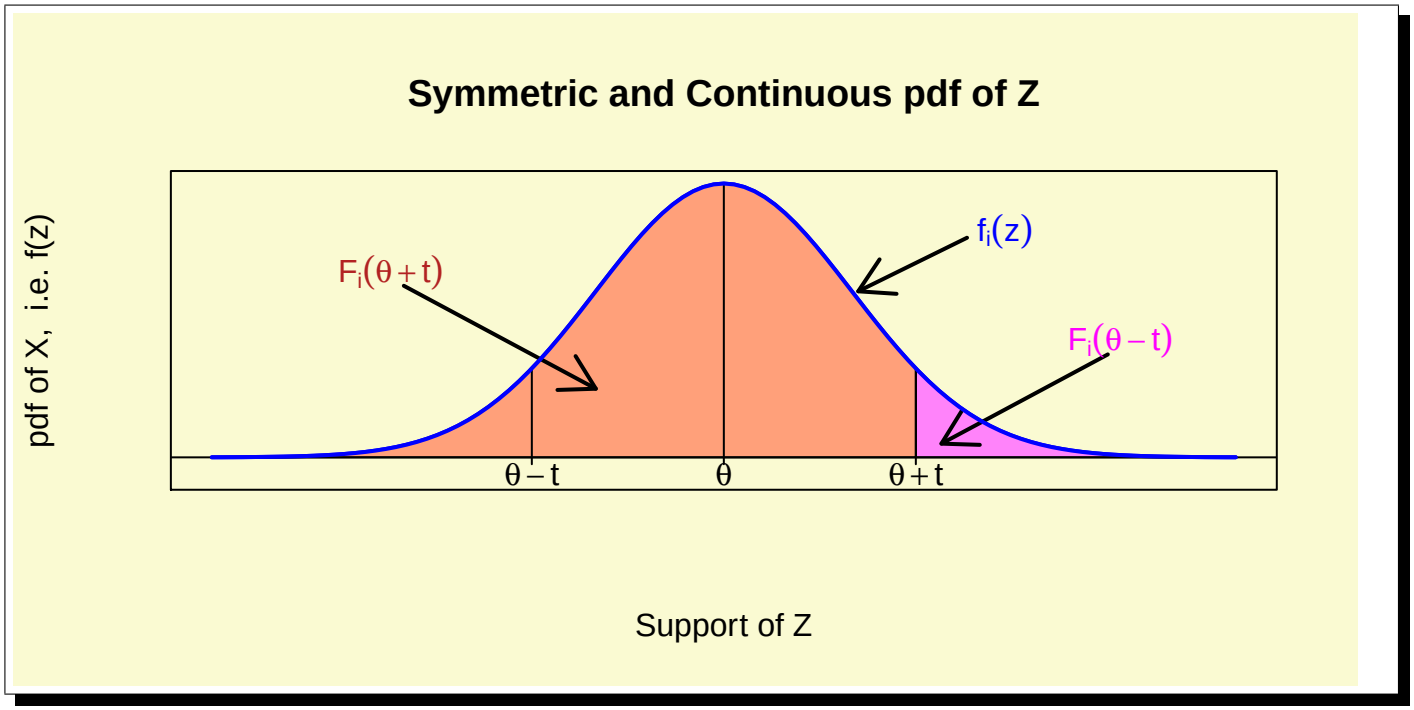
Data: We obtain $\boldsymbol{2n}$ observations, two observations on each of $\boldsymbol{n}$ subjects.

Subject i:	**1**	**2**	**3**	$\cdots$	n
X_i :	X_1	X_2	X_3	$\cdots$	X_n
Y_i :	Y_1	Y_2	Y_3	$\cdots$	Y_n

Assumptions:

1. We let $\boldsymbol{Z_i = Y_i - X_i}$ for $\boldsymbol{i = 1(1)n}$. The $\boldsymbol{Z_i}$'s are mutually independent.
2. Each $\boldsymbol{Z_i}$, $\boldsymbol{i = 1(1)n}$, comes from a continuous population (not necessarily the same are) that is symmetric about a common median $\boldsymbol{\theta}$; i.e.

$$\boldsymbol{i.e.\ F_i(\theta + t) = 1 - F_i(\theta - t)} \qquad \forall\, t \in \mathbb{R}\ \boldsymbol{and}\ \forall\, i = 1(1)n$$
$$\boldsymbol{\iff F_i(\theta + t) + F_i(\theta - t) = 1} \qquad \forall\, t \in \mathbb{R}\ \boldsymbol{and}\ \forall\, i = 1(1)n$$



The parameter $\boldsymbol{\theta}$ is refered to as the treatment effect.

Target :

We have to test,

$$\boldsymbol{H_0 : \theta = 0} \qquad \boldsymbol{Vs.} \quad \begin{cases} \boldsymbol{a)} & \boldsymbol{H_1 : \theta > 0} \\ \boldsymbol{b)} & \boldsymbol{H_2 : \theta < 0} \\ \boldsymbol{c)} & \boldsymbol{H_3 : \theta \neq 0} \end{cases}$$

$\boldsymbol{H_0}$ implies there is **zero shift** in location due to treatment. That is, each of the distributions (not necessarily same) for the differences $\boldsymbol{(Y_i - X_i)}$ is symmetrically distributed about $\mathbf{0}$.

Procedure :

1. Find the absolute values $\boldsymbol{|Z_1|, |Z_2|, \ldots, |Z_n|}$.
2. Order them from least to greatest.
3. Define, $\boldsymbol{R_i = Rank\ of\ |Z_i|,\ i = 1(1)n}$ in the ordering.
4. Define, $\boldsymbol{d_i} = \begin{cases} \mathbf{1} & \boldsymbol{if\ Z_i > 0} \\ \mathbf{0} & \boldsymbol{if\ Z_i < 0} \end{cases}$

   Theoretically, $\boldsymbol{Pr[Z_i = 0] = 0}$

   $\boldsymbol{R_i d_i}$ is known as the **positive signed rank** of $\boldsymbol{|Z_i|}$.
5. **Test Statistic :**

   The **Wilcoxon signed rank statistic** $\boldsymbol{T^+}$ is that the sum of the positive signed ranks, namely,

$$\boldsymbol{T^+ = \sum_{i=1}^{n} R_i d_i}$$

(a) To test $\boldsymbol{H_0 : \theta = 0}$ Vs. $\boldsymbol{H_1 : \theta > 0}$.
Reject $\boldsymbol{H_0}$ at the $\boldsymbol{\alpha}$ level of significance if $\boldsymbol{T^+ \geq t_\alpha}$; otherwise do not reject $\boldsymbol{H_0}$.

> ***Q.) How do we determine t_α?***
>
> Choose $\boldsymbol{t_\alpha}$ such that $\boldsymbol{P_{H_0}[T^+ \geq t_\alpha] = \alpha}$ $\boldsymbol{\cdots\cdots\cdots\cdots\cdots}$ $\mathbf{(1)}$.
> But due to the discreteness of the distribution of $\boldsymbol{T^+}$, it is **not** always possible to find $\boldsymbol{t_\alpha}$ for an arbitrary choice of $\boldsymbol{\alpha}$ that satisfies $\mathbf{(1)}$.
> Therefore we choose $\boldsymbol{t_\alpha}$ such that $\boldsymbol{P_{H_0}[T^+ \geq t_\alpha] \leq \alpha}$.
>
> This is nothing but a **one-sided upper tail test**.

(b) To test $\boldsymbol{H_0 : \theta = 0}$ Vs. $\boldsymbol{H_2 : \theta < 0}$.

Reject $\boldsymbol{H_0}$ at the $\boldsymbol{\alpha}$ level of significance if $\boldsymbol{T^+ \leq \frac{n(n+1)}{2} - t_\alpha}$; otherwise do not reject $\boldsymbol{H_0}$.

(c) To test $\boldsymbol{H_0 : \theta = 0}$ Vs. $\boldsymbol{H_1 : \theta \neq 0}$.

Reject $\boldsymbol{H_0}$ at the $\boldsymbol{\alpha}$ level of significance if $\boldsymbol{T^+ \geq t_{\frac{\alpha}{2}}}$ or $\boldsymbol{T^+ \leq \frac{n(n+1)}{2} - t_{\frac{\alpha}{2}}}$; otherwise do not reject $\boldsymbol{H_0}$.

Q.) How do we determine $t_{\frac{\alpha}{2}}$?

Choose $\boldsymbol{t_{\frac{\alpha}{2}}}$ such that $\boldsymbol{P_{H_0}[T^+ \geq t_{\frac{\alpha}{2}}] \leq \frac{\alpha}{2}}$ ..... **(2)**.

But due to the discreteness of the distribution of $\boldsymbol{T^+}$, it is **not** always possible to find $\boldsymbol{t_{\frac{\alpha}{2}}}$ for an arbitrary choice of $\boldsymbol{\alpha}$ that satisfies **(2)**.

Therefore we choose $\boldsymbol{t_{\frac{\alpha}{2}}}$ such that $\boldsymbol{P_{H_0}[T^+ \geq t_{\frac{\alpha}{2}}] \leq \frac{\alpha}{2}}$.

This is nothing but a **upper tail test of two-sided test**.

Null Distribution of T^+ (No ties case) :

Define, $\boldsymbol{B}$ = **No. of positive $\boldsymbol{Z_i}$'s.**

Let $\boldsymbol{r_1 < r_2 < \ldots < r_B}$ denote the ordered ranks of the absolute values of these **positive $\boldsymbol{Z_i}$'s**.

Then the null distribution of $\boldsymbol{T^+}$ can be obtained directly from the *representation* $\boldsymbol{T^+ = \sum_{i=1}^{B} r_i}$.

Under the assumption that the underlying $\boldsymbol{Z_i}$ distributions are all continuous, the probabilities are **zero** that there are ties among the absolute values of $\boldsymbol{Z_i}$'s or that any of the $\boldsymbol{Z_i}$ are exactly **zero**.

In addition, under $\boldsymbol{H_0}$, these underlying $\boldsymbol{Z_i}$ distributions are all symmetric about $\boldsymbol{\theta = 0}$.

It follows that undser $\boldsymbol{H_0}$, each of the $\boldsymbol{2^n}$ possible outcomes for the ordered configuration $\boldsymbol{(r_1, r_2, \ldots, r_B)}$ occurs with probability $\boldsymbol{\frac{1}{2^n}}$.

E.g. $n = 3$; therefore $2^3 = 8$ possible outcomes for $(r_1, r_2, \ldots, r_B)$.

B	$(r_1, r_2, \ldots, r_B)$	Probability under H_0	$T^+ = \sum_{i=1}^{B} r_i$
0	$-$	1/8	0
1	$r_1 = 1$	1/8	1
1	$r_1 = 2$	1/8	2
1	$r_1 = 3$	1/8	3
2	$r_1 = 1, r_2 = 2$	1/8	3
2	$r_1 = 1, r_2 = 3$	1/8	4
2	$r_1 = 2, r_2 = 3$	1/8	5
3	$r_1 = 1, r_2 = 2, r_3 = 3$	1/8	6

So, clearly,

$T^+ = t$:	0	1	2	3	4	5	6
$P_{H_0}[T^+ = t]$:	1/8	1/8	1/8	2/8	1/8	1/8	1/8

The event $\boldsymbol{T^+ = 3}$ corresponds to $\begin{cases} \boldsymbol{B = 1} & \boldsymbol{\&} \quad \boldsymbol{r_1 = 3} \\ \boldsymbol{or,} \\ \boldsymbol{B = 2} & \boldsymbol{\&} \quad \boldsymbol{r_1 = 1, r_2 = 2} \end{cases}$

Note :

- The test procedure is based on $\boldsymbol{T^+}$ are called **distribution free procedure**.

  Reason ?

  Answer: We have derived the null distribution of $\boldsymbol{T^+}$ without specifying the forms of the underlying $\boldsymbol{Z}$ populations under $\boldsymbol{H_0}$ beyond the point of requiring that they are continuous and symmetric about "**0**".

▶ **Mean and Variance of $\boldsymbol{T^+}$ under $\boldsymbol{H_0}$:**

$$\boldsymbol{T^+ = \sum_{i=1}^{B} r_i}$$

$$\boldsymbol{T^+ \overset{d}{=} \sum_{i=1}^{B} V_i}$$

where $\boldsymbol{V_1, V_2, \ldots V_n}$ are mutually independent dichotomous random variables, with probability distribution,

$$\boldsymbol{P[V_i = i] = P[V_i = 0] = 1/2, i = 1(1)n}$$

$$\begin{aligned} \therefore \boldsymbol{E_{H_0}(T^+)} &= \boldsymbol{E\left(\sum_{i=1}^{n} V_i\right)} \\ &= \boldsymbol{\sum_{i=1}^{n} E(V_i)} \\ &= \boldsymbol{\sum_{i=1}^{n} \left[i.(1/2) + 0.(1/2)\right]} \\ &= \boldsymbol{\sum_{i=1}^{n} \frac{i}{2}} \\ &= \boldsymbol{\frac{n(n+1)}{4}} \end{aligned}$$

Also,

$$\begin{aligned} \boldsymbol{Var_{H_0}(T^+)} &= \boldsymbol{Var\left(\sum_{i=1}^{n} V_i\right)} \\ &= \boldsymbol{\sum_{i=1}^{n} Var(V_i)} \end{aligned}$$

But,

$$\begin{aligned} \boldsymbol{Var(V_i)} &= \boldsymbol{E\left(V_i^2\right) - \left(E(V_i)\right)^2} \\ &= \boldsymbol{\left(i^2.\frac{1}{2} + 0^2.\frac{1}{2}\right) - \left(\frac{i}{2}\right)^2} \\ &= \boldsymbol{\frac{i^2}{4}} \end{aligned}$$

$$\begin{aligned} \boldsymbol{\therefore Var_{H_0}(T^+)} &= \boldsymbol{\sum_{i=1}^{n} \frac{i^2}{4}} \\ &= \boldsymbol{\frac{n(n+1)(2n+1)}{24}} \end{aligned}$$

Note :
The assymptotic normality of the standardized form

$$\boldsymbol{T^* = \frac{T^+ - E_{H_0}(T^+)}{\left(Var_{H_0}(T^+)\right)^{1/2}}}$$

follows **Lyapunov Central Limit Theorem**.

Remarks :

1. $\boldsymbol{T^+ + T^- = \sum_{i=1}^{n} \frac{n(n+1)}{2}}$

   Test statistics based on $\boldsymbol{T^+}$ only, $\boldsymbol{T^-}$ only, or $\boldsymbol{T^+ + T^-}$ are linearly related and therefore equivalent in terms.

2. To use the signed rank statistic in hypothesis testing, the entire null distribution is not necessary. Infact, one set of critical values is sufficient for even a two sided test, because of the relations $\boldsymbol{T^+ + T^- = \sum_{i=1}^{n} \frac{n(n+1)}{2}}$ and the symmetry of $\boldsymbol{T^+}$ about $\boldsymbol{\frac{n(n+1)}{4}}$.

3. Large values of $\boldsymbol{T^+}$ corresponds to small values of $\boldsymbol{T^-}$ and furthermore $\boldsymbol{T^+ \overset{d}{=} T^-}$ under $\boldsymbol{H_0}$.

   Proof :

$$\begin{aligned} \boldsymbol{P_{H_0}[T^+ \geq c]} &= \boldsymbol{P_{H_0}[T^+ - \frac{n(n+1)}{4} \geq c - \frac{n(n+1)}{4}]} \\ &= \boldsymbol{P_{H_0}[\frac{n(n+1)}{4} - T^+ \geq c - \frac{n(n+1)}{4}]} \\ &= \boldsymbol{P_{H_0}[\frac{n(n+1)}{2} - T^+ \geq c]} \\ &= \boldsymbol{P_{H_0}[T^- \geq c]} \end{aligned}$$

   So, we can write $\boldsymbol{T^+ \overset{d}{=} T^-}$ under $\boldsymbol{H_0}$.

In the case where two or more absolute values of differences are equal, i.e. $|Z_i| = |Z_j|$ for at least one $i \neq j$, the observations are tied. the first and most possibility is to discard all tied observations and reduce the sample size accordingly. This method certainly leads to a loss of information, but if the number of observations to be omitted is small relative to the sample size, the loss may be minimal.

Another approach is to use **mid-rank method**.

Q.) What is mid-rank method?

The **mid-rank method** assigns to each member of a group of tied observations the simple average of the ranks they would have if distinguishable. Using this approach, tied observations are given tied ranks.

Remarks :

1. The mid-rank method affects the null distribution of ranks. The mean rank is unchanged, but the variance of the ranks is redeuced.
2. Coming back to Wilcoxon signed rank test, the probability distribution of T^+ is clearly not the same in the presence of tied ranks, but the effect is generally slight and no correction needs to be made unless the ties are quite extensive.

▶ Correction of Variance according to ties :

Suppose that "t" observations are tied for a given rank, say "$s+1$" and that if they would be given the ranks $s+1, s+2, \cdots, s+t$. The mid-rank is they $s + \frac{t+1}{2}$ and the **sum of squares** of these **ranks** is,

$$\underbrace{(s+\frac{t+1}{2})^2 + (s+\frac{t+1}{2})^2 + \cdots + (s+\frac{t+1}{2})^2}_{(t\ times)}$$

$$= t(s+\frac{t+1}{2})^2$$

$$= t\{s^2 + s(t+1) + \frac{(t+1)^2}{4}\}$$

If these ranks had not been tied, their **sum of squares** would have been ,

$$\sum_{i=1}^{t}(s+i)^2 = ts^2 + ts(t+1) + \frac{t(t+1)(2t+1)}{6}$$

The presence of these "t" ties then decrease the **sum of squares** by,

$$\frac{t(t+1)(2t+1)}{6} - \frac{t(t+1)^2}{4}$$

$$= \frac{t(t+1)(t-1)}{12}$$

$$= \frac{t(t^2-1)}{12}$$

A bit of algebra leads to a reduced variance,

$$\boldsymbol{Var(T^+|H_0) = \frac{n(n+t)(2n+1)}{24} - \sum_t \frac{t(t^2-1)}{48}}$$

,where the sum is extended over all sets of "$\boldsymbol{t}$" ties. This is called the ***correction for ties***.

■ Confidence-Interval Procedures :

As with the ordinary one-sample sign test, the Wilcoxon signed-rank procedure lends itself to confidence-interval estimation of the unknown population median M. The confidence limits are those values of M which do not lead to rejection of the null hypothesis. To find these limits for any sample size N, we first find the critical value $t_{\alpha/2}$ such that if the true population median is M and T is calculated for the derived sample values $X_i - M$, then

$$P(T^+ \leq t_{\alpha/2}) = \alpha/2 \quad and \quad P(T^- \leq t_{\alpha/2}) = \alpha/2$$

The null hypothesis will not be rejected for all numbers M which make $T^+ > t_{\alpha/2}$ and $T^- > t_{\alpha/2}$. The confidence-interval technique is to use trial and error to find those two numbers, say M_1 and M_2 where $M_1 < M_2$, such that when T is calculated for the two sets of differences $X_i - M_1$ and $X_i - M_2$, at significance level α, T^+ or T^-, whichever is smaller, is just short of significance, i.e., slightly larger than $t_{\alpha/2}$. This generally does not lead to a unique interval, and the manipulations can be tedious even for moderate sample sizes.

This technique is best illustrated by an example. The following eight observations are drawn from a continuous, symmetric population:

$$-1, 6, 13, 4, 2, 3, 5, 9 \ \ldots\ldots\ldots\ldots\ (a)$$

For $N = 8$ the two-sided rejection region of nominal size 0.05 was found earlier to be $t_{\alpha/2} = 3$ with exact significance level

$$\alpha = P(T^+ \leq 3) + P(T^- \leq 3) = {}^{10}/_{256} = 0.039$$

We try six different values for M and calculate T^+ or T^-, whichever is smaller, for the differences $X_i - M$. The example illustrates a number of difficulties which arise. In the first trial choice of M, the number 4 was subtracted and the resulting differences contained three sets of tied pairs and one zero even though the original sample contained neither ties nor zeros. If the zero difference is ignored, N must be reduced to 7 and then the $t_{\alpha/2} = 3$ is no longer accurate for $\alpha = 0.039$.

Table :

Trial-and-Error Determination of Endpoints :

X_i	$X_i - 4$	$X_i - 1.1$	$X_i - 1.5$	$X_i - 9.1$	$X_i - 8.9$	$X_i - 8.95$
-1	-5	-2.1	-2.5	-10.1	-9.9	-9.95
6	2	4.9	4.5	-3.1	-2.9	-2.95
13	9	11.9	11.5	3.9	4.1	4.05
4	0	2.9	2.5	-5.1	-4.9	-4.95
2	-2	0.9	0.5	-7.1	-6.9	-6.95
3	-1	1.9	1.5	-6.1	-5.9	-5.95
5	1	3.9	3.5	-4.1	-3.9	-3.95
9	5	7.9	7.5	-0.1	0.0	0.05
T^+ or T^-		3	3.5	3	5	5

The midrank method could be used to handle the ties, but this also disturbs the accuracy of $t_{\alpha/2}$. Since there seems to be no real solution to these problems, we try to avoid zeros and ties by judicious choices for our M values for subtraction. These data are all integers, and hence a choice for M which is not an integer obviously reduces the likelihood of ties and makes zero values impossible. Since T^- for the differences $X_i - 1.5$ yields $T^- = 3.5$ using the midrank method, we will choose $M_1 = 1.5$. The next three columns represent an attempt to find an M which makes T^+ around 4. These calculations illustrate the fact that M_1 and M_2 are far from being unique. Clearly M_2 is in the vicinity of 9, but the differences $X_i - 9$ yield a zero. We conclude there is no need to go further. An approximate 96.1% confidence interval on M is given by $1.5 < M < 9$. The interpretation is that hypothesized values of M within this range will lead to acceptance of the null hypothesis for an exact significance level of 0.039.

This procedure is undoubtedly tedious, but the limits obtained are reasonably accurate. The numbers should be tried systematically to narrow down the range of possibilities. Thoughtful study of the intermediate results usually reduces the additional number of trials required.

A different method of construction which leads to a unique interval and is much easier to apply is described in Noether [(1967), pp. 57-58]. The procedure is to convert the interval $T^+ > t_{\alpha/2}$ and $T^- > t_{\alpha/2}$ to an equivalent statement on M whose end points are functions of the observations X_i. For this purpose we must analyze the comparisons involved in determining the ranks of the differences $r(|X_i - M_0|)$ and the signs of the differences $X_i = M_0$ since T^+ and T^- are functions of these comparisons. Note that the rank of any random variable in a set $\{V_1, V_2, ..., V_w\}$ can be written symbolically as

$$r(V_i) = \sum_{k=1}^{N} S(V_i - V_k) + 1$$

where

$$S(u) = \begin{cases} 1 & \text{if } u > 0 \\ 0 & \text{if } u \leq 0 \end{cases}$$

To compute a rank, then we make $\binom{N}{2}$ comparisons of pairs of different numbers and one comparison of a number with itself. To compute the sets of all ranks, we make $\binom{N}{2}$ comparisons of pairs and N identity comparisons, a total of $\binom{N}{2} + N = N(N+1)/2$ comparisons. Substituting the rank function in (7.1), we obtain

$$T^+ = \sum_{i=1}^{N} Z_i r$$
$$= \sum_{i=1}^{N} Z_i + \sum_{i=1}^{N} \sum_{k \neq i} Z_i S(|X_i - M_0| - |X_k - M_0|) \ldots\ldots\ldots\ldots\ldots (b)$$

Therefore these comparisons affects T^+ as follows :

1. A comparison of $|X_i - M_0|$ with itself adds 1 to T^+ if $X_i - M_0 > 0$.
2. A comparison of $|X_i - M_o| > 0$ with $|X_k - M_0|$ for any $i \neq k$ adds 1 to T^+ if $|X_i - M_0| > |X_k - M_0|$ and $X_i - M_0 > 0$, that is, $X_i - M_0 > |X_k - M_0|$. If $X_k - M_0 > 0$, this occurs when $X_i > X_k$, and if $X_k - M_0 < 0$, we have $X_i + X_k > 2M_0$ or $(X_i + X_k)/2 > M_0$. But when $X_i - M_0 > 0$ and $X_k - M_0 > 0$, we have $(X_i + X_k)/2 > M_0$ also.

Combining these two results, then, $(X_i + X_k)/2 > M_0$ is a necessary condition for adding 1 to T^+ for all i, k. Similarly, if $(X_i + X_k)/2 < M_0$, then this comparison adds 1 to T^-. The relative magnitudes of the $N(N+1)/2$ averages of pairs $(X_i + X_k)/2$ for all $i \leq k$, called the **Walsh averages**, then determine the range of values for hypothesized numbers M_0 which will not lead to rejection of H_0. If these $N(N+1)/2$ averages are arranged as order statistics, the two numbers which are in the $(t_{\alpha/2} + 1)$ position from either end are the endpoints of the $100(1-\alpha)\%$ confidence interval on M. Note that this procedure is exactly analogous to the ordinary sign-test confidence interval except that here the order statistics are for the averages of all pairs of observations instead of the original observations.

The data in **(a)** for $N = 8$ arranged in order of magnitude are -1,2,3,4,5,6,9,13, and the 36 Walsh averages are given in Table 7.5. For exact $\alpha = 0.039$, we found before that $t_{\alpha/2} = 3$. Since the fourth largest numbers from either end are 1.5 and 9.0, the confidence interval is $1.5 < M < 9$ with exact confidence coefficient $\gamma = 1 - 2(0.039) = 0.922$. This result agrees exactly with that obtained by the previous8 method, but this will not always be the case since the trial-and-error procedure does not yield unique endpoints.

The process of determining a confidence interval on M by the above method is much facilitated by using the graphical method of construction, which can be described as follows.

Table :

Walsh averages for data in (a) :

-1.0	0.5	1.0	1.5	2.0	2.5	4.0	6.0
2.0	2.5	3.0	3.5	4.0	5.5	7.5	
3.0	3.5	4.0	4.5	6.0	8.0		
4.0	4.5	5.0	6.5	8.5			
5.0	5.5	7.0	9.0				
6.0	7.5	9.5					
9.0	11.0						
13.0							

H_0 will not be rejected $\forall M$ which make $T^+ > t_{\frac{\alpha}{2}}$ and $T^- > t_{\frac{\alpha}{2}}$.
Now what is $t_{\frac{\alpha}{2}}$?
$P(T^+ \leq t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and $P(T^- \leq t_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

■ Trial and error method :

Find those 2 numbers, say M_1 and M_2 where $M_1 < M_2$, such that when T is calculated for the two sets of difference, $X_i - M_1$ and $X_i - M_2$, at α level of significance, $min\{T^+, T^-\}$ is just short of significance, i.e., slightly larger than $t_{\frac{\alpha}{2}}$.

This generelly does not lead to a unique interval, and the manipulations can be tedious even for moderate sample sizes.

eg: $-1, 6, 13, 4, 2, 3, 5, 9$

$m = 8$

$\alpha = 0.05$

$t_{\alpha/2} = 3$ from table.

Exact significance level $\alpha = P_{H_0}(T^+ \leq 3) + P_{H_0}(T^- \leq 3) = 0.02 + 0.02 = 0.04$

We try six different values of M and calculate T^+ and T^- whichever, for the difference $X_i - M$.

	$X_i - 4$	$X_i - 1.1$	$X_i - 1.5$	$X_i - 9.1$	$X_i - 8.9$	$X_i - 8.95$
Here, $n = 7$	3		3.5	3	5	5

$\Downarrow$

$t_{\alpha/2} = 3$ is no longer accurate for $\alpha = 0.039$

- Try to avoid zeroes and ties by judicious choices for our M values for substitution.

These data are all integers, and hence a choice for M which is not an integar obviously reduces the likelihood of ties and makes zero values impossible.

$H_0 : M = 2$ Vs. $H_1 : M \neq 2$

Data : $-3 \quad -6 \quad 1 \quad 9 \quad 4 \quad 10 \quad 12$

$\alpha = 0.10$

Target : To find confidence interval for M.

Recall: The confidence limits are those values of M which do not lead to rejection of H_0.

H_0 will not be rejected $\forall M$ which make $T^+ > t_{\alpha/2}$ and $T^- > t_{\alpha/2}$

i.e. , $min\{T^+, T^-\} > t_{\alpha/2}$

Here $\frac{\alpha}{2} = 0.05$; $t_{\frac{\alpha}{2}} =?\ Pr\{T^+ \leq t_{\frac{\alpha}{2}}\} = \frac{\alpha}{2} \implies t_{\frac{\alpha}{2}} = 3$

Let's apply trial and error method.

Table :

Trial-and-Error Determination of Endpoints :

X_i	$X_i - 2.9$	$X_i - 0.9$	$X_i - 0.4$	$X_i + 2.1$	$X_i + 3.1$	$X_i + 2.9$
-3	-5.9	-3.9	-3.4	-0.9	0.1	-0.1
-6	-8.9	-6.9	-6.4	-4.9	-2.9	-3.1
1	-1.9	0.1	0.6	3.1	4.1	3.9
9	6.1	8.1	8.6	11.1	12.1	11.9
4	1.1	3.1	3.6	6.1	7.1	6.9
10	7.1	9.1	9.6	12.1	13.1	12.9
12	9.1	11.9	11.6	14.1	15.1	14.9
$T^+ or T^-$	11	7	6	4	2	3

Here $T^-$for the differnces "$X_i - (-2.1)$" yields $T^- = 4$, we will choose $M_1 = -2.1$
Now , let's focus on finding an M which makes T^+around 4.

X_i	$X_i - 9.9$
-3	-12.9
-6	-15.9
1	-8.9
9	-0.9
4	-5.9
10	0.1
12	2.1
$T^+ or T^-$	4

$\therefore$ We may take $M_2 = 9.9$

Here exact significance level $\alpha = 2 \times 0.039 = 0.078$
$\Rightarrow 1 - \alpha = 1 - 0.078 = 0.922$
$\therefore$ An approximate 92.2 C.I. on M is given by $(-2.1, 9.9)$.

► **Interpretation :**

The hypothesized values of M within the range $(-2.1, 9.9)$ will lead to acceptance of the H_0 for exact significance level 0.078.

C. Order Statistics

Let $X_1, X_2, \ldots, X_n$ denote a random sample from a population with continuous cdf F_X. Here, there exists a unique ordered arrangement within the sample. Suppose, $X_{(1)}$ denotes the samallest of $X_1, X_2, \ldots, X_n$; $X_{(2)}$ denotes the second smallest ; ... and $X_{(n)}$ denotes the largest. Then,

$$X_{(1)} < X_{(2)} < ... < X_{(r)} < ... < X_{(n)}, \textit{ where } 2 < r < n$$

denotes the original random sample after arrangement in increasing order of magnitude, and these are collectively termed the ordered statistics of the random sample $X_1, X_2, \ldots, X_n$. The r^{th} smallest , $1 \leq r \leq n$, $X_{(r)}$ is called the r^{th} order statistic.

- **Sample median :**

$$\begin{cases} X_{[\frac{n+1}{2}]} & , for\ n\ odd \\ Any\ number\ between\ X_{(\frac{n}{2})}\ and\ X_{(\frac{n}{2}+1)} & , for\ n\ even \end{cases}$$

- **Sample midrange :** $\dfrac{(X_{(1)} + X_{(n)})}{2}$
- **Sample Range :** $X_{(n)} - X_{(1)}$

► **Quantile function : (κ_p or $Q_X(p)$ or X_p)**

A quantile of a distribution is that value of X such that a specific percentage of the probability is at or below it. Thus a quantile divides the area under the pdf into two parts of specific amounts. Only the area to the left of the number need to be specified since the entire area is equal to 1.

The p^{th} quantile (or the $100p^{th}$ percentile) is that value of the random varible X, say X_p, such that $100p\%$ of the values of X in the population are less than or equal to X_p , for any positive fraction p $(0 < p < 1)$

$$i.e.\ P(X \leq X_p) = p$$
$$\therefore F_X(X_p) = p$$

Moreover, if F_X is strictly increasing, the p^{th} quantile is the unique solution to the equation :

$$X_p = F_X^{-1}(p) = Q_X(p)\ say,$$

for a given p and the inverse of the cdf $Q_X(p)$, $0 < p < 1$, is called the **quantile function of the random variable X**.

Thus the p^{th} quantile is the solution to the equation $F_X(x) = p$. Since the cdf may not be increasing for all values, we define p^{th} quantile $Q_X(p)$ as the smallest value at which the cdf is atleast equal to p, or,

$$Q_X(p) = F_X^{-1}(p) = inf\{x : F_X(x) \geq p\},\ 0 < p < 1$$

This definition gives a unique value for the quantile $Q_X(p)$ even when F_X is flat or is step function.

▶ **CDF of $X_{(r)}$:**

$$\boldsymbol{P(X_{(r)} \leq t) = \sum_{i=r}^{n} \binom{n}{i} [F_X(t)]^i [1 - F_x(t)]^{n-i},\ -\infty < t < \infty}$$

▶ **pdf of $X_{(r)}$:**

$$\boldsymbol{f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!} [F_X(x)]^{r-1} [1 - F_x(x)]^{n-r} f_X(x),\ -\infty < x < \infty}$$

- For a random sample of size $\boldsymbol{n}$ from the $\boldsymbol{U(0,1)}$ the $\boldsymbol{r^{th}}$ order statistic $\boldsymbol{X_{(r)}}$ follows a $\boldsymbol{beta(r, n-r+1)}$ **distribution**.
- For $\boldsymbol{U(0,1)}$ distribution,

$$\boldsymbol{P(X_{(r)} \leq t) = \sum_{i=r}^{n} \binom{n}{i} t^i (1-t)^{n-i}}$$

$$\boldsymbol{= \underbrace{\frac{1}{B(r, n-r+1)} \int_0^t x^{r-1}(1-x)^{n-r} dx}_{= I_t(r, n-r+1)}}$$

- One key reason why the order statistics are so important in nonparametric statistics is that for any order statistic $\boldsymbol{X_{(r)}}$ from a continuous cdf $\boldsymbol{F}$, the transformed random variable $\boldsymbol{U_r = F(X_{(r)})}$ has the same distribution as that of the $\boldsymbol{r^{th}}$ ordered statistic from the $\boldsymbol{U(0,1)}$, regardless of the shape of $\boldsymbol{F}$ as long as it is continuous; in this sense $\boldsymbol{F(X_{(r)})}$ may be viewed as **distribution free**.

  This property of continuous ordered statistics is called the **probability-integral transformation (PIT)**.

▶ **Confidence interval for a population quantile :**

$$\boldsymbol{F_X(\kappa_p) = p\ \ldots\ldots\ldots\ (1) \qquad ; \kappa_{0.50} = The\ median\ of\ the\ distribution}$$
$$\boldsymbol{\Rightarrow \kappa_p = Q_X(p) = F_X^{-1}(p)}$$

Assumption : *Unique solution to the equation* ***(1)***.

A natural point estimate of $\boldsymbol{\kappa_p}$ is the $\boldsymbol{p^{th}}$ sample quantile, which is the $\boldsymbol{(np)^{th}}$ order statistic, provided of course $\boldsymbol{np}$ is an integer.

We define the order statistic $\boldsymbol{X_{(r)}}$ to be the $\boldsymbol{p^{th}}$ sample quantile where $\boldsymbol{r}$ is defined by,

$$\boldsymbol{r = \begin{cases} np & if\ np\ is\ an\ integer \\ [np+1] & if\ np\ is\ \ not\ an\ integer \end{cases}}$$

$$\boldsymbol{\therefore \quad \underbrace{X_{(r)}}_{\substack{p^{th}\ sample \\ quantile}} = \begin{cases} X_{(np)} & if\ np\ is\ an\ integer \\ X_{([np+1])} & if\ np\ is\ not\ an\ integer \end{cases}}$$

A logical choice for the CI endpoints are the two order statistics, say $X_{(r)}$ and $X_{(s)}$, $r < s$, from the random sample drawn from the population F_X.

To find the $100(1-\alpha)\%$ CI, we must then find the two integers r and s, $1 \leq r \leq s \leq n$, such that ,

$$P(X_{(r)} < \kappa_p < X_{(s)}) = 1 - \alpha \textit{ for some given number } 0 < \alpha < 1$$

Alternative notation for $1-\alpha$ *:* "γ" $\rightarrow$ **Confidence level or Confidence co-efficient.**

The event $\{X_{(r)} < \kappa_p\}$ happens iff either $\{X_{(r)} < \kappa_p < X_{(s)}\}$ or $\{X_{(s)} < \kappa_p\}$, and these later two events are clearly mutually exclusive.

$$\therefore \ \forall r < s, \ P(X_{(r)} < \kappa_p) = P(X_{(r)} < \kappa_p < X_{(s)}) + P(X_{(s)} < \kappa_p)$$

Equivalently,

$$P(X_{(r)} < \kappa_p < X_{(s)}) = P(X_{(r)} < \kappa_p) - P(X_{(s)} < \kappa_p) \ \ldots\ldots\ldots (2)$$

$\therefore F_X$ is strictly increasing function, $X_{(r)} < \kappa_p$ iff $F_X(X_{(r)}) < F_X(\kappa_p) = p$.

But when F_X is continuous, the **PIT** implies that the probability distribution of the random variable X, i.e. $F_X(X_{(r)})$ is the same as that of $U_{(r)}$, the r^{th} order statistic from the uniform distribution over the interval $(0, 1)$.

$$\begin{aligned} P[X_{(r)} < \kappa_p] &= P[F_X(X_{(r)}) < F_X(\kappa_p)] \\ &= P[F_X(X_{(r)}) < p] \quad [\because F_X(\kappa_p) = \kappa_p] \\ &= P[U_{(r)} < p] \\ &= \int_0^p \frac{n!}{(r-1)!(n-r)!} x^{r-1}(1-x)^{n-r}dx \\ &= \int_0^p n\binom{n-1}{r-1} x^{r-1}(1-x)^{n-r}dx \ \ldots\ldots\ldots\ldots (3) \end{aligned}$$

Clearly, this probability does not depend on F_X. A confidence interval based on **(2)** is therefore **distribution free**.

In order to find the interval estimate of κ_p, we substitute **(3)** into **(2)** and find r and s such that,

$$\begin{aligned} P(X_{(r)} < \kappa_p < X_{(s)}) &= \int_0^p n\binom{n-1}{r-1} x^{r-1}(1-x)^{n-r}dx - \int_0^p n\binom{n-1}{s-1} x^{s-1}(1-x)^{n-s}dx \\ &= 1-\alpha \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (4) \end{aligned}$$

Clearly, this equation will not give a unique solution for the two unknowns, r and s, and the additional conditions are needed.

For the nearest possible interval for a fixed confidence co-efficient, r and s would be chosen such that (iv) is satisfied and $X_{(s)} - X_{(r)}$, or $E(X_{(s)} - X_{(r)})$, is as small as possible. Alternatively, we could minimize $s - r$.

Applying integration by parts, leads to,

$$\boldsymbol{P(X_{(r)} < \kappa_p) = \sum_{j=0}^{n-r} \binom{n}{r+j} p^{r+j}(1-p)^{n-r-j}}$$

or after substituting $\boldsymbol{r+j=i}$,

$$\boldsymbol{P(X_{(r)} < \kappa_p) = \sum_{i=r}^{n} \binom{n}{i} p^{i}(1-p)^{n-i}}$$

$$\begin{aligned}
\therefore\ \boldsymbol{P(X_{(r)} < \kappa_p < X_{(s)})} &= \boldsymbol{\sum_{i=r}^{n} \binom{n}{i} p^{i}(1-p)^{n-i} - \sum_{i=s}^{n} \binom{n}{i} p^{i}(1-p)^{n-i}} \\
&= \boldsymbol{\sum_{i=r}^{s-1} \binom{n}{i} p^{i}(1-p)^{n-i}} \\
&= \boldsymbol{P(r \le K \le s-1)} \\
&\boldsymbol{where\ K \sim Bin(n,p)\ \ldots\ldots\ldots\ldots\ (5)}
\end{aligned}$$

Choose $\boldsymbol{r}$ and $\boldsymbol{s}$, such that $\boldsymbol{(s-r)}$ is **minimum** for fixed $\boldsymbol{\alpha}$.
Discreteness issue arises here too.
So, choose $\boldsymbol{r}$ and $\boldsymbol{s}$ such that,

$$\begin{aligned}
\boldsymbol{P(X_{(r)} < \kappa_p < X_{(s)})} &= \boldsymbol{P(r \le K \le s-1)} \\
&\boldsymbol{\ge 1-\alpha}
\end{aligned}$$

► **Alternative way (simple way) :**
The event $\boldsymbol{\{X_{(r)} < \kappa_p\}}$ occurs iff at least $\boldsymbol{r}$ of the $\boldsymbol{n}$ sample values, $\boldsymbol{X_1, ..., X_n}$ are less than $\boldsymbol{\kappa_p}$. Thus,

$$\begin{aligned}
\boldsymbol{P[X_{(r)} < \kappa_p]} &= \boldsymbol{P[exactly\ r\ of\ the\ n\ observations < \kappa_p]} \\
&\quad \boldsymbol{+P[exactly\ (r+1)\ of\ the\ n\ observations < \kappa_p]} \\
&\quad \vdots \\
&\quad \boldsymbol{+P[exactly\ n\ of\ the\ n\ observations < \kappa_p]} \\
\boldsymbol{P[X_{(r)} < \kappa_p]} &= \boldsymbol{\sum_{i=r}^{n} P[exactly\ i\ of\ the\ n\ observations < \kappa_p]}
\end{aligned}$$

The probability that exactly $\boldsymbol{i}$ of the $\boldsymbol{n}$ observations are less than $\boldsymbol{\kappa_p}$ can be found as the probability of $\boldsymbol{i}$ successes in $\boldsymbol{n}$ independent Bernoulli trials, since the sample observations are all independent and each observation can be classified as either a success or a failure, where a success is defined as an observation less than $\boldsymbol{\kappa_p}$.
In other words,

$$\begin{aligned}
&\boldsymbol{P[exactly\ i\ of\ the\ n\ sample\ values < \kappa_p]} \\
&= \boldsymbol{\binom{n}{i} p^{i}(1-p)^{n-i}}
\end{aligned}$$

$$\therefore\ \boldsymbol{P(X_{(r)} < \kappa_p) = \sum_{i=r}^{n} \binom{n}{i} p^i (1-p)^{n-i}}$$

▶ **Summary :**

The $\mathbf{100(1-\alpha)\%}$ CI for the $\boldsymbol{p^{th}}$ quantile is given by $\boldsymbol{(X_{(r)}, X_{(s)})}$, where $\boldsymbol{r}$ and $\boldsymbol{s}$ are integars such that $\boldsymbol{1 \le r \le s \le n}$ and

$$\begin{aligned} \boldsymbol{P(X_{(r)} < \kappa_p < X_{(s)})} &= \boldsymbol{\sum_{i=r}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}} \\ &\boldsymbol{\ge 1-\alpha} \ \boldsymbol{\ldots\ldots\ldots\ldots\ldots\ (6)} \end{aligned}$$

Choose $\boldsymbol{r}$ and $\boldsymbol{s}$ such that $\boldsymbol{(s-r)}$is **minimum**.

One common approach : Assign the probability $\frac{\alpha}{2}$ in each tail. This yields the so called **"equal tails"** interval, where $\boldsymbol{r}$ and $\boldsymbol{s}$ are the largest and least integers $\boldsymbol{1 \le r \le s \le n}$ such that,

$$\left.\begin{aligned} &\boldsymbol{\sum_{i=1}^{r} \binom{n}{i} p^i (1-p)^{n-i} \le \frac{\alpha}{2}} \\ \&\ &\boldsymbol{\sum_{i=0}^{s-1} \binom{n}{i} p^i (1-p)^{n-i} \ge 1-\frac{\alpha}{2}} \end{aligned}\right\} \ \boldsymbol{\ldots\ldots\ldots\ (7)}$$

Remarks :

1. In some cases there may be no $\boldsymbol{r-1}$, $\boldsymbol{r \ge 1}$ that satisfies **1st inequality** of **(7)**. In this case we take $\boldsymbol{X_{(r)} = -\infty}$. This means that for the given $\boldsymbol{n}$, $\boldsymbol{p}$, $\boldsymbol{\alpha}$, we obtain a one sided (upper) CI $\boldsymbol{(-\infty, X_{(s)})}$ with exact confidence level $\boldsymbol{\sum_{i=0}^{s-1} \binom{n}{i} p^i (1-p)^{n-i}}$ and we may want to choose $\boldsymbol{s}$ such that this level is atleast $\boldsymbol{1-\alpha}$, rather than $\boldsymbol{1-\frac{\alpha}{2}}$.
2. Similarly, there may be no $\boldsymbol{s-1 \le n}$, which satisfies the **2nd inequality** of **(7)** and in that case we take the right hand CI end point $\boldsymbol{X_{(s)} = \infty}$, so that we obtain a one sided (lower) CI $\boldsymbol{(X_{(r)}, \infty)}$ with exact confidence level $\boldsymbol{1 - \sum_{i=0}^{r-1} \binom{n}{i} p^i (1-p)^{n-i} \ge 1-\alpha}$.

D. Test For Randomness

Introduction :

Run: Given an ordered sequence of one or more types of symbols, a run is defined to be a succession of one or more types of symbols which are followed and preceded by a different symbol or no symbol at all.

Suppose we have an ordered sequence of two types of symbols, T_1 and T_2. We may want to know whether the symbols are occurring randomly or following some pattern. The number of runs in a sequence may give clue of lack of randomness.

Let, total number of runs of 1^{st} type of symbol $(T_1) = R_1$,
Total number of runs of 2^{nd} type of symbol $(T_2) = R_2$,
Total number of runs $R = R_1 + R_2$
Then the total number of runs can be used to test for randomness

▶ Hypothesis :

Here our null hypothesis is H_0 :The arrangement is random against the alternative, H_1 : The null hypothesis is not true.

In this case both too few runs and both too many runs suggest lack of randomness.

- Too few runs suggests that symbols are clustered in the arrangements and they are following some trend
- Again too many runs suggests that the symbols are positioned pairwise in arrangement and they are following some kind of cyclic pattern

▶ Test based on the total number of runs :

Assume an ordered sequence of n elements with n_1 elements of 1^{st} type of symbol T_1 and n_2 elements of 2^{nd} type of symbol T_2. Let there are r_1 runs of T_1 and r_2 runs of T_2. Then the total number of runs $r = r_1 + r_2$. In order to derive a test of randomness based on variable R we need to find the null distribution of R i.e., the distribution of R under the null hypothesis.

▶ Exact null distribution of *R*:

The distribution of R will be found by first determining the joint probability distribution of R_1 and R_2 and then the distribution of their sum. Since under the null hypothesis every arrangement of the n_1+n_2 objects is equiprobable, the probability that $R_1 = r_1$ and $R_2 = r_2$ is the number of distinguishable arrangements, which is $\dfrac{n!}{n_1!n_2!}$. For the numerator quantity, the following counting lemma can be used.

The number of distinguishable ways of distributing n-like objects into r distinguishable cells with no cell empty is $\binom{n-1}{r-1}$, $n \geqslant r$.

All possible cases :

1. $r_1 = r_2$ and $c = 2$
2. $r_1 = r_2 + 1$ and $c = 1$

3. $r_1 = r_2 - 1$ and $c = 1$

Therefore the joint probability distribution of R_1 and R_2 is

$$f_{R_1,R_2}(r_1, r_2) = \frac{c\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n}} \quad r_1 = 1, 2, ..., n_1;\ r_2 = 1, 2, ..., n_2 \text{ —(1)}$$

The marginal probability distribution of R_1 and R_2 is,

$$f_{R_1}(r_1) = \frac{\binom{n_1-1}{r_1-1}\binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n}} \qquad r_1 = 1, 2, ..., n_1$$

$$f_{R_2}(r_2) = \frac{\binom{n_2-1}{r_2-1}\binom{n_2-1}{r_2-1}}{\binom{n_1+n_2}{n}} \qquad r_2 = 1, 2, ..., n_2$$

The probability distribution of R is

$$f_R(r) = \begin{cases} 2\frac{\binom{n_1-1}{r/2-1}\binom{n_2-1}{r/2-1}}{\binom{n_1+n_2}{n}} & \text{,when r is even} \\ \frac{\binom{n_1-1}{(r-1)/2}\binom{n_2-1}{(r-3)/2}+\binom{n_2-1}{(r-3)/2}\binom{n_1-1}{(r-1)/2}}{\binom{n_1+n_2}{n_1}} & \text{,when r is odd} \end{cases}$$

because r is even implies $r_1 = r_2 = r/2$ and (1) is summed over this pair. If $r_1 = r_2 + 1$ or $r_1 = r_2 - 1$,r is odd. In this case (1) is summed over the two pairs of values $r_1 = (r-1)/2$ and $r_2 = (r+1)/2$, $r_1 = (r+1)/2$ and $r_2 = (r-1)/2$, obtaining the given result.

Alternative Way :

Case 1: $r = 2k + 1$

- If the sample starts with the symbol $T_1$and ends with the symbol T_1, then we get the case $r_1 = r_2+1$ i.e. $r_2 = k$ and $r_1 = k + 1$
- If the sample starts with the symbol $T_2$and ends with the symbol T_2, we get the case $r_1 = r_2 - 1$ i.e $r_1 = k$ and $r_2 = k + 1$.

For the first case total number of arrangements is $\begin{pmatrix} n_1 - 1 \\ k \end{pmatrix}\begin{pmatrix} n_2 - 1 \\ k - 1 \end{pmatrix}$ and in the second case total number of arrangements is $\begin{pmatrix} n_2 - 1 \\ k \end{pmatrix}\begin{pmatrix} n_1 - 1 \\ k - 1 \end{pmatrix}$. So total number of possible arrangements is $\begin{pmatrix} n_1 - 1 \\ k \end{pmatrix}\begin{pmatrix} n_2 - 1 \\ k - 1 \end{pmatrix} + \begin{pmatrix} n_2 - 1 \\ k \end{pmatrix}\begin{pmatrix} n_1 - 1 \\ k - 1 \end{pmatrix}$.

Hence, $P(R = r) = \dfrac{\begin{pmatrix} n_1 - 1 \\ k \end{pmatrix}\begin{pmatrix} n_2 - 1 \\ k - 1 \end{pmatrix} + \begin{pmatrix} n_2 - 1 \\ k \end{pmatrix}\begin{pmatrix} n_1 - 1 \\ k - 1 \end{pmatrix}}{\begin{pmatrix} n_1 + n_2 \\ n_1 \end{pmatrix}}$, when $r = 2k + 1$

Case 2: $r = 2k$

If the sample starts with the symbol T_1 and end with the sybol T_2or vice-versa, we get the above case. i.e $r_1 = r_2 = k$

So total number of arrangements is $2\left(\begin{pmatrix} n_1 - 1 \\ k - 1 \end{pmatrix}\begin{pmatrix} n_2 - 1 \\ k - 1 \end{pmatrix}\right)$.

Hence, $P(R=r) = \dfrac{2\left(\begin{pmatrix} n_1-1 \\ k-1 \end{pmatrix}\begin{pmatrix} n_2-1 \\ k-1 \end{pmatrix}\right)}{\begin{pmatrix} n_1+n_2 \\ n_1 \end{pmatrix}}$, when $r = 2k$

Hence the probability distribution of R is -

$$f_R(r) = \begin{cases} 2\dfrac{\begin{pmatrix} n_1-1 \\ r/2-1 \end{pmatrix}\begin{pmatrix} n_1-1 \\ r/2-1 \end{pmatrix}}{\begin{pmatrix} n_1+n_2 \\ n_1 \end{pmatrix}} & \text{when } r \text{ is even} \\ \dfrac{\begin{pmatrix} n_1-1 \\ (r-1)/2 \end{pmatrix}\begin{pmatrix} n_2-1 \\ (r-3)/2 \end{pmatrix}+\begin{pmatrix} n_2-1 \\ (r-3)/2 \end{pmatrix}\begin{pmatrix} n_1-1 \\ (r-1)/2 \end{pmatrix}}{\begin{pmatrix} n_1+n_2 \\ n_1 \end{pmatrix}} & \text{when } r \text{ is odd} \end{cases}$$

► **Rejection criteria :**

- H_1 : The symbols are following some "trend" pattern

Reject H_0 if there exists a c_1 such that
$P(R \leqslant c_1) \leqslant \alpha$

- H_1: The symbols are following some "cyclical" pattern

Reject H_0 if there exists a c_2 such that
$P(R \geqslant c_2) \leqslant \alpha$

- H_1 : The arrangement is non random

Reject H_0 if there exists a c_1 and c_2 such that
$P(R \leqslant c_1) + P(R \geqslant c_2) \leqslant \alpha$

► **Importance :**

Test for randomness is a very important addition to the statistical theory. Because most of the statistical analysis is started with the assumption of having a random sample. If the assumption is valid then every sequential order is of no consequence. However if the randomness is suspected then the information about order, which is almost always available, can be used to test a hypothesis of randomness. This type of testing is helpful in time series and quality control analysis.

► **Remark :**

The run test is applicable in both qualitative and quantitative data. In the latter case, the values are compared with a focal point, often the mean or median and notingwhether they exceedor is exceeded by this value. If any observation is equal to the focal point then it is ignored in analysis and n_1,n_2 and n are reduced accordingly.

E. Kolmogorov-Smirnov

Tests of Goodness of Fit :

In classical statistics, information about the form generally must be postulated in the null hypothesis to perform an exact parametric type of inference. For example, suppose we have a small number of observations from an unknown population with unknown variance and the hypothesis of interest concerns the value of the population mean. The traditional parametric test,, based on Student's t-distribution,is derived under the assumption of a normal population. Therefore, it muist be desirable to check on the reasonableness of the normality assumption before forming any conclusions based on t-distribution.

- ***How do we check ?***

  Ans : *Goodness of Fit tests.*

▶ Types of Goodness of fit tests :

The first type is designed for null hypothesis concerning a discrete distribution and compares the **observed frequencies** with the **frequencies expected** under the null hypothesis. This is the chi-square test provided by Karl Pearson.

Thew second type of goodness of fit test is designed for null hypothesis concerning a continuous distribution and compares the **observed cumulative relative frequencies** with **those of expected** under the null hypothesis. This group includes-

1. **Kolmogorov-Smirnov (K-S) Test**
2. **Lilliefors Test**
3. **Anderson-Darling (A-D) Test**

Remark : One may use graphical approaches too.

■ The Kolmogorov-Smirnov (K-S) One Sample Statistic :

(**Recall:** *Emperical Distribution Function*)

For a random sample from theb distribution with cdf $F_X(x)$, the emperical distribution function or edf, denoted by $S_n(x)$, is imply the proportion of sample values less than or equal to the specified value of x, that is -

$$S_n(x) = \frac{no.\ of\ sample\ values \leq x}{n}$$

In terms of order statistics,

$$S_n(x) = \begin{cases} 0 & if\ x < x_{(1)} \\ \frac{i}{n} & if\ X_{(i)} \leq x \leq x_{(i+1)},\ i = 1, 2, \ldots, n-1 \\ 1 & if\ x \geq x_{(n)} \end{cases}$$

In case of tied observations, the edf is still a step function but it jumps only at the distinct observed sample values $X_{(j)}$ and the height of the jump is equal to $\frac{k}{n}$, where k is number of values tied at $X_{(j)}$.

► **Statistical Properties of $S_n(x)$:**

Let $T_n(x) = n.S_n(x)$

Result-1 :- For any fixed real value of x, the random variable $T_n(x) \sim Bin(n, F_X(x))$.

$$\therefore E(S_n(x)) = F_X(x)$$
$$\& \; Var(S_n(x)) = \frac{F_X(x)(1 - F_X(x))}{n}$$

For any fixed real value of x, $S_n(x)$ is a consistent estimator of $F_X(x)$, or, in other words, $S_n(x) \xrightarrow{P} F_X(x)$.

$$E(T_n(x)T_n(y)) = n.F_X(x) + n(n-1)F_X(x)F_Y(y) for\ x < y$$

♦ Glivenko-Cantelli Theorem :

$S_n(\cdot)$ converges uniformly to $F_X(\cdot)$ with probability 1, that is

$$Pr\Big\{ \lim_{n\to\infty} sup\big[|S_n(x) - F_X(x)|\big] \Big\} = 1$$

As $n \to \infty$, the limiting distribution of the standardized $S_n(x)$ is standard normal or

$$\lim_{n\to\infty} Pr\left\{ \frac{\sqrt{n}[S_n(x) - F_X(x)}{\sqrt{F_X(x)(1 - F_X(x))}} \leq t \right\} = \Phi(t)$$

Lets come back to KS one sample statistic.

Assumption : $X_1, X_2, ..., X_n$ be a sample from a population that is continuous. Let $F(.)$ be the corresponding cdf.

Target : To test the hypothesis that the sample comes from a specified cdf F_0 against the alternative that it is from some other cdf F_1 where $F_1(x) \neq F_0(x)$ for some $x \in \mathbb{R}$

- ***How to deal with the problem ?***

  Comparison can be made between observed and expected cumulative relative frequencies for each of the observed values. Several goodness of fit test statistics are function of the derivation between the edf and population cdf specified under the null hypothesis. The function of these deviations used to perform a goodness of fit test might be the sum of a square or absolute values, or the maximum deviations, to name only a few. The best known test is the K-S one sample statistic.

► **Test Statistic :**

According to Glivenko-Cantelli theorem as $n \to\infty$, $S_N(x)$ approaches the cdf $F_\theta(x)$ for all x. Therefore for large n the deviations between the true function and its statistical image, $|S_N(x) - F_0(x)|$ should be small for all values of x.This suggests that if H_0 is true the statistic

$D_n = |S_n(x) - F_0(x)|$ is for any x seasonable measure of our estimate.

This D_n statistic, called the K-S one sample statistic, is particularly useful in nonparametric statistical inference because the probability distribution of D_n does not depend on $F_\theta(x)$ as long as F_0 is continous.

Therefore D_n is a distribution free statistics.

$D_n^+ = \sup_x(S_n(x) - F_0(x))$ & $D_n^- = \sup_x(F_0(x) - S_n(x))$ are called the one sided K-S statistics.

Result : ***The statistics*** D_n, D_n^+, D_n^- ***are completely distribution-free for any specified continuous cdf*** F_θ.

Proof :

Defining $X_{(0)} = -\infty$ and $X_{(n+1)} = \infty$, we can write $S_n(x) = \frac{i}{n}$ for $X_{(i)} \leq x < X_{(i+1)}$ for $i = 0, 1, 2, ..., n$.

$$\begin{aligned}
D_n^+ &= \sup_x [S_n(x) - F_0(x)] \\
&= \max_{0\leq i\leq n} \sup_{X_{(i)}\leq x<X_{(i+1)}} [S_n(x) - F_0(x)] \\
&= \max_{0\leq i\leq n} \sup_{X_{(i)}\leq x<X_{(i+1)}} \left[\frac{i}{n} - F_0(x)\right] \\
&= \max_{0\leq i\leq n} [\frac{i}{n} - \inf_{X_{(i)}\leq x<X_{(i+1)}} F_0(x)] \\
&= \max_{0\leq i\leq n} \left[\frac{i}{n} - F_0(X_{(i)})\right] \\
&= max(\max_{0\leq i\leq n} \left[\frac{i}{n} - F_0(X_{(i)})\right], 0)
\end{aligned}$$

Now, $D_n = \underset{x}{sup}|S_n(x) - F_0(x)| = \underset{x}{max}(D_n^+, D_n^-)$

Similarly, we can show that,

$$D_n^- = max(\underset{1\leq i\leq n}{max}[F_0(X_{(i)}) - \frac{i-1}{n}], 0)$$

.

Also, we know that,

$$\begin{aligned}
D_n &= \underset{x}{max}(D_n^+, D_n^-) \\
&= max\left\{\max_{1\leq i\leq n}\left[\frac{i}{n} - F_0(X_{(i)})\right], \underset{1\leq i\leq n}{max}\left[F_0(X_{(i)}) - \frac{i-1}{n}\right], 0\right\}
\end{aligned}$$

Observe the probability distribution of D_n, D_n^+ and D_n^- depend only on the random variables $F_0(X_{(i)})$, $i = 1, 2, ..., n$ under H_0, these are order statistics from $U(0, 1)$, regardless of the original F_0 as long as it is continuous and completely specified.

Thus D_n, D_n^+ and D_n^- have distributions which are independent of the particular F_0.

Result :

For $D_n = \underset{x}{sup}|S_n(x) - F_0(x)|$, where $F_0(x)$ is only specific continuous cdf, we have under H_0,

$$P(D_n < \frac{1}{2n} + \nu) = \begin{cases} 0 & for\ \nu \leq 0 \\ \int_{\frac{1}{2n}-\nu}^{\frac{1}{2n}+\nu} \int_{\frac{1}{3n}-\nu}^{\frac{1}{3n}+\nu} \dots \int_{\frac{2n-1}{2n}-\nu}^{\frac{2n-1}{2n}+\nu} f(u_1, u_2, ..., u_n) du_1 ... du_n & for\ 0 < \nu < \frac{2n-1}{2n} \\ 1 & for\ \nu \geq \frac{2n-1}{2n} \end{cases}$$

where $f(u_1, u_2, ..., u_n) = \begin{cases} n! & o < u_1 < u_2 < \ldots < u_n < 1 \\ o & otherwise \end{cases}$

- Numerical values of $D_{n,\alpha}$ are given for $n \leq 40$ and selected tail probabilities α.
- For larger sample sizes, Kolmogorov (1933) observed the following convenient approximation to the sample data of D_n.

  "If F_X is any continuous df, then for any $d > 0$,

$$\lim_{n\to\infty} P\left\{D_n \leq \frac{d}{\sqrt{n}}\right\} = L(d)$$

  where ,

$$L(d) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2 d^2}$$

  "

Result :

$$P_{H_0}(D_n^+ < c) = \begin{cases} 0 & for\ c \leq 0 \\ \int_{1-c}^{1} \int_{\frac{n-1}{n}-c}^{u_n} \dots \int_{\frac{1}{n}-c}^{u_2} f(u_1, u_2, ..., u_n) du_1 ... du_n & for\ 0 < c < 1 \\ 1 & for\ c \geq 1 \end{cases}$$

where $f(u_1, u_2, ..., u_n) = \begin{cases} n! & o < u_1 < u_2 < \ldots < u_n < 1 \\ o & otherwise \end{cases}$

- D_n^+ and D_n^- have identical distributions because of symmetry.
- For large n, $\forall d \geq 0$, $\lim_{n\to\infty} P\left\{D_n \leq \frac{d}{\sqrt{n}}\right\} = 1 - e^{-2d^2}$
- If F_0 is any specified continuous cdf, then for every $d \geq 0$, the limiting null distribution of $V = 4nD_n^{+^2}$, as $n \to \infty$, is the $\chi^2_{(2)}$.

■ Applications of the K-S one sample statistics :

Assume that we have the random sample $X_1, X_2, ..., X_n$ and the hypothesis,

$H_0 : F_X(x) = F_0(x)\ \forall x$ where $F_0(x)$ is completely specified continuous cdf .

The differences between $S_n(x)$ & $F_0(x)$ should be small for all x except for sampling variation , if $H_0$is true.

For $H_1 : F_X(x) \neq F_0(x)$ for some x , large absolute values of thet deviations tend to discredit the H_0
Therefore ,the K-S goodness - of - fit test with significance level α is to reject $H_0$when $D_n > D_{n_1\alpha}$
The following expression is considerably easier for algebra calculations & applies when ties are present
:

$$\begin{aligned} D_x &= \underset{x}{sup}|S_n(x) - F_0(x)| \\ &= \underset{x}{sup}\,[|S_n(x) - F_0(x)|, |S_n(x-\epsilon) - F_0(x)|] \end{aligned}$$

,where ϵ denotes any small positive number.

► **One - Sided Tests :**

Spse $H_1 : F_X(x) \geq F_0(x)\ \forall x$
the appropriate rejection-region is $D_n^+ > D_{n_1\alpha}^+$
Suppose $H_1 : F_\alpha(x) \leq F_0(x), \forall x$, $H_0$is rejected when $D_n^- > D_{n,\alpha}^-$

- Most test of the goodness of fit are two-sided.
- The tail probabilities for the the one-sided statistic are approx. one-half of the corresponding tail proabilities for the two sided statistic.

► **Confidence Bounds :**

Recall ,

$$\begin{aligned} &Pr\{D_n > D_{n,\alpha}\} = \alpha \\ \Leftrightarrow &Pr\{D_n < D_{n,\alpha}\} = 1 - \alpha \\ \Leftrightarrow &Pr\{\underset{x}{Sup}|S_n(x) - F_X(x)| < D_{n,\alpha}\} = 1 - \alpha \\ \Leftrightarrow &Pr\{S_n(x) - D_{n,\alpha} < F_X(x) < S_n(x) + D_{n,\alpha}, \forall x\} = 1 - \alpha \end{aligned}$$

Thus we define

$$\begin{aligned} L_n(x) &= max(S_n(x) - D_{n_1\alpha}, 0) \\ \&\ U_n(x) &= min(S_n(x) + D_{n_1\alpha}, 1) \end{aligned}$$

as lower & upper confidence bounds assosiated with confidence coefficient $1 - \alpha$.

■ **Determination of sample Size :**

The statistics $D_n$enables us to determine the minimum sample size required to garuntee with a certain probabilty $1 - \alpha$, that the error in the estimate never exceeds a fixed value c
i.e., We want to find the minimum value of n that satisfies

$$\begin{aligned} &Pr\{D_n < c\} = 1 - \alpha \\ \Leftrightarrow &1 - Pr\{D_n < c\} = Pr\{D_n > c\} = \alpha \end{aligned}$$

$\therefore$ c equals $D_{n,\alpha}$
$'n'$ can read directly from tables as that sample sixe corresponding to $D_{n,\alpha} = c$

E.g. Spse error should be less 0.25 with probabaility 0.98, we look down the $0.02 = 1 - 0.98$ column of tbales until we find the largest $c \leq 0.25$. This entry is 0.247 which corresponds to $n = 36$.

SOURCE: Gibbons & Chakrabarti

F. The Wald-Wolfowitz Runs

► **Some Two Sample Problems :**

So far we observed problem related to either single set of observations or two dependent samples. Therefore, all these problems can be legitimately classified as one sample problems.

From here onwards, we will be concerned with the data consisting of two mutually independent random samples, that is, random samples drawn independently from each of two populations. Not only are the elements within each sample independent, but also every element in the first sample is independent of every element in the second sample.

Assumptions :

1. The observations $X_1, X_2, \ldots, X_m$ are a random sample from population 1 with C.D.F. F_X. The observations $Y_1, Y_2, \ldots, Y_n$ are a random sample from population 2 with C.D.F. F_Y.
2. The X's and Y's are mutually independent.
3. Populations 1 and 2 are continuous populations.

Target :

To test whether the two samples are drawn from identical populations,

$$i.e.\ H_0 : F_X(x) = F_Y(x) \forall x \in \mathbb{R}$$

[Recall t-test for equality of means]

Tests of H_0 depend on the type of alternative specified.

Some of the alternatives :

1. **Location alternative :** $F_Y(x) = F_X(x - \theta),\ \theta \neq 0$ i.e. $Y \overset{D}{=} X + \theta,\ \theta \neq 0.$
2. **Scale alternative :** $F_Y(x) = F_X(x\theta),\ \theta \neq 1$ i.e. $Y \overset{D}{=} \dfrac{X}{\theta},\ \theta \neq 1.$
3. **Lehmann alternative :** $F_Y(x) = 1 - \big(1 - F_X(x)\big)^{\theta+1},\ \theta + 1 > 0.$
4. **Stochastic alternative :** $F_Y(x) \geq F_X(x),\ \forall x$ and $F_Y(x) > F_X(x),\ for\ atleast\ one\ x.$
5. **General alternative :** $F_Y(x) \neq F_X(x),\ for\ some\ x.$

Alternatives 1. and **2.** show differences in F_X and F_Y in location and scale respectively.

Alternative 3. states that $Pr\{Y > x\} = \big[Pr\{X > x\}\big]^{\theta+1}$. In the special case when θ is an integer, it states that Y has the same distribution as the smallest of the $\theta + 1$ of X-variables.

$$i.e. Y \stackrel{d}{=} X_{1:\theta+1}$$

A similar alternative to test that is sometimes used is $F_Y(x) = \big(F_X(x)\big)^{\alpha}$ for some $\alpha > 0$ and for all x. When α is an integer, this states that Y is distributed as the largest of α X-variables.

$$i.e. Y \stackrel{d}{=} X_{\alpha:\alpha}$$

Alternative 4. refers to the relative magnitudes of X's and Y's. It states that

$$Pr\{Y \leq x\} \geq Pr\{X \leq x\}$$
$$So\ that,\ Pr\{Y > x\} < Pr\{X > x\}$$

In other words, X's tend to be larger than Y's.

Under H_0,the two random sample can be considered a single random sample of size $N = m+n$ drawn from the common, continuous, but unspecified population. Then the combined ordered configuration of the m X and n Y random variables in the sample is one of the $\binom{m+n}{m}$ possible equally likely arrangements.

Eg. $m = 3$, $n = 2$

Under H_0,each of the $\binom{m+n}{m}$ possible equally likely arrangements.

Eg. $m = 3$, $n = 2$

Under H_0, each of the $\binom{5}{2} = 10$ possible arrangements of the combined single shown below is equally likely.

1. $XXXYY$	2. $XXYXY$	3. $YXYXY$
4. $XXYYX$	5. $XYXXY$	6. $XYXYX$
7. $YXXXY$	8. $YXXYX$	9. $XYYXX$
10. $YYXXX$		

Remark : The sample pattern of arrangement of X's & Y's provides information about the type of the difference which may exist in the populations.

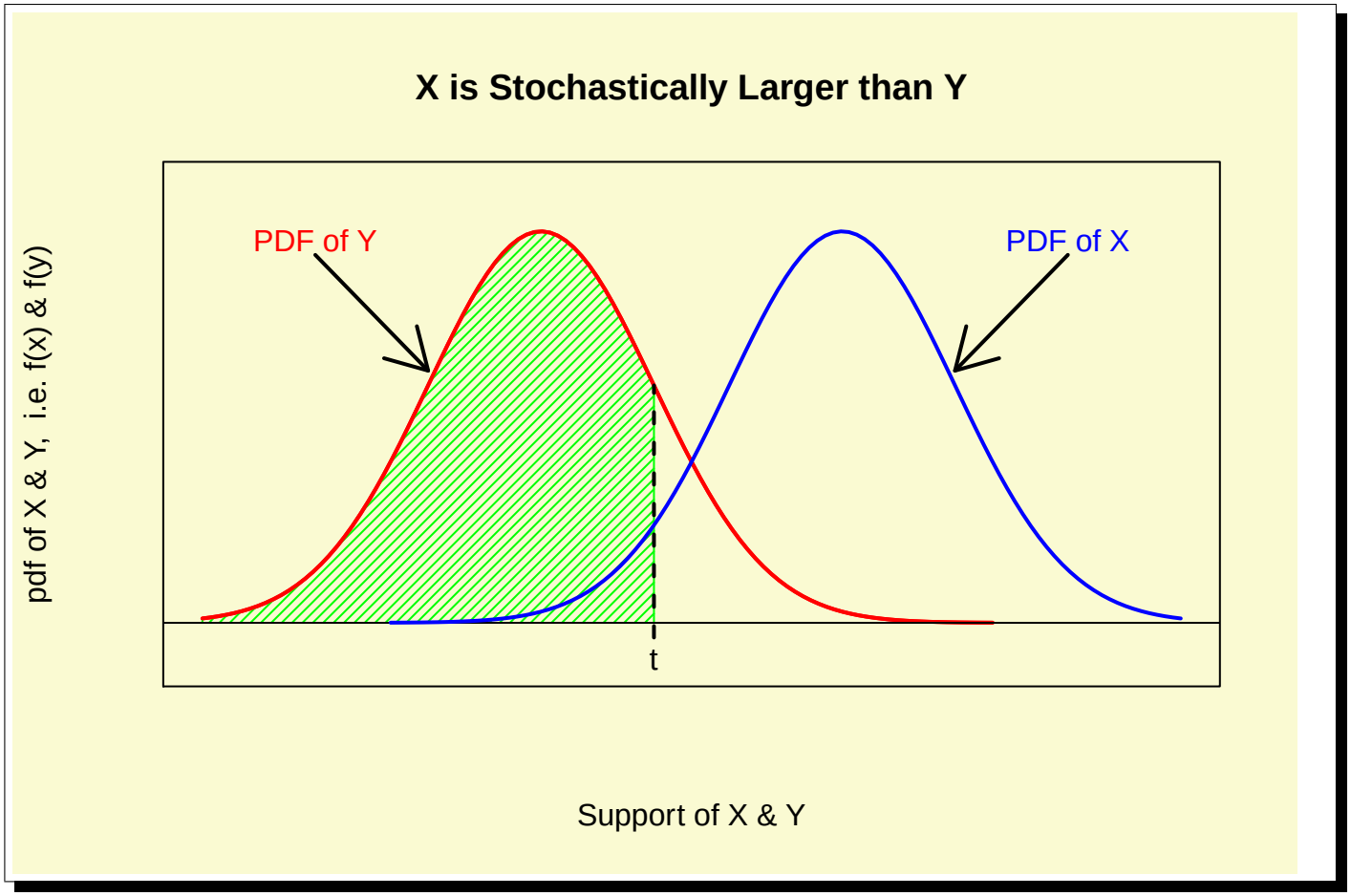
Many statistical tests are based on same function of this combined arrangement. The type of function which is most appropriate depends on the the type of the difference one hopes to detect, which is indicated by the alternative hypothesis.

(recall the type of alterations discussed)

Definition : $X \geq_{st} Y$

We may say that a continuous random variable X is stochastically larger than a continuous random variable Y if

$$\begin{aligned} & P(Y \leq x) \ \geq P(X \leq x) \quad \forall x \\ \text{and} \quad & P(Y \leq x) \ > P(X \leq x) \quad \text{for at least one } x \in \mathbb{R} \end{aligned}$$



$$\boldsymbol{X \geq_{st} Y}$$

Recall location alternative,

$$H_A : F_Y(x) = F_X(x - \theta)\ \forall x \in \mathbb{R}\ \&\ some\ \theta \neq 0$$
$$\Rightarrow Y \stackrel{d}{=} X + \theta$$
$$So\ that, Y \geq_{st} X\ (\text{or } Y \leq_{st} X)\ iff\ \theta > 0\ (\theta < 0)$$

Recall scale alternative,

$$H_A : F_Y(x) = F_X(\theta x)\ \forall x \in \mathbb{R}\ \&\ some\ \theta \neq 1$$
$$\Rightarrow Y \stackrel{d}{=} \frac{X}{\theta}$$
$$So\ that, Y \geq_{st} X\ (\text{or } Y \leq_{st} X)\ iff\ \theta < 1\ (\theta > 1)$$

Recall,
$H_A : F_Y(x) = (F_X(x))^{\alpha}$, for some positive integer α & $\forall x$.
This is called **Lehman alternative**.
Hence, $Y \stackrel{d}{=} X_{\alpha}$
Under this alternative, $Y \underset{st}{\geq} X (wY \underset{st}{\leq} X)$ iff $k > 1 (k < 1)$

▶ The Wald-Wolfowitz Runs Test :

Combine the two sets of random samples $X_1, X_2, ..., X_m$ and $Y_1, Y_2, ..., Y_n$ into a single ordered sequence from smallest to largest, keeping track of which observations correspond to the X sample and which to the Y.

Assuming that their probability distribution are continuous, a unique ordering is always possible, since theoretically ties do not exist.

Eg: For $m = 4$, $n = 5$, a typical arrangement might be $XXYXYYXYY$

Under H_0 of identical distributions.

$H_0 : F_Y(x) = F_X(x) \forall x$ we expect the X and Y random variables to be well mixed in the ordered configuration, since the $m + n = N$ random variables constitute a single random sample of size N from the population.

(Recall the definition of "run".

A run is a squence of identical letters proceeded and followed by a different letter or no letter.)

Point to note : The total number of runs in the ordered pooled sample is an indication of the degree of mixing.

A pattern of arrangement with too few runs would suggest that this group of N is not a single random sample but instead is composed of two samples from two distinguishable populations.

Eg:

1. $XXXXYYYYY$. Here $R = 2$. May be $Y \geq_{st} X$
2. $YYYYYXXXX$. Here $R = 2$. May be $Y \leq_{st} X$

Remark : Test criterion based solely on the total number of runs cannot distinguish this above two cases.

The runs test is appropriate primarily when the alternative is completely general and two-sided as in

$H_A : F_Y(x) \neq F_X(x)$ for some x

Define $R \coloneqq$ The total no. of runs in the combined arrangement of mX and nY random variables.

▶ Rejection criteria :

Since too many few runs tend to discredit the H_0 when the alternative is H_A, the Wald-Wolfowitz(1940) runs test for significance level α generally has the rejection region in the lower tail as

$$R \leq c_\alpha$$

where c_α is chosen to be the largest integar satisfying $P_{H_0}(R \leq c_\alpha) \leq \alpha$.

The p value for the runs test is then given by $Pr\{R \leq R_0\}$, where R_0 is the observed value of the runs test statsitic R.

Remark:

1. Under H_0, the probability distribution of R is exactly the same as we found for the runs test for randomness.

2. The other properties of R including the moemnts & assymptotic null distribution are also unchanged.

3. The only difference here is that the assymptotic critical region for the alternative of different population is too few runs.

▶ **The problem of ties :**

Ties do not present a problem in counting the number of runs unless the tie is across the samples; that is , two or more observations from different samples have exactly the same magnitude.

We can break all the ties in all possible ways & compute the total no. of runs for each resolution of all ties . The values of the test statistic R is the largest computed value , since that is the one least likely to lead to rejection of H_0.

For each groups of ties across samples , where there are s x's and y's of equal magnitude for some $s \geq 1, t \geq 1$, there are $\binom{s+t}{s}$ ways to break the ties. Thus ,if there are k groups of ties, the total no. of values of R to be computed is the product $\prod_{i=1}^{k} \binom{s_i+t_i}{s_i}$.

G. The Kolmogorov-Smirnov (K-S) Two Sample Test

Here the compsrison is made between the emperical distribution functions of the two samples.

Data :

Two independent random samples of size n from continuous populations with CdFs F_X and F_Y respectively.

$$X_1, X_2, X_3, \ldots, X_m$$
$$Y_1, Y_2, Y_3, \ldots, Y_n$$

The respective emperical distribution function, denoted by $S_m(x)$ and $S_n(x)$ are defined as before :

$$S_m(x) = \begin{cases} 0 & if\ x < X_{(1)} \\ \dfrac{k}{m} & if\ X_{(k)} \le x < X_{(k+1)}\ for\ k = 1, 2, \ldots, m-1 \\ 1 & if\ x \ge X_{(m)} \end{cases}$$

$$\&\quad S_n(x) = \begin{cases} 0 & if\ x < Y_{(1)} \\ \dfrac{k}{n} & if\ Y_{(k)} \le x < Y_{(k+1)} \quad for\ k = 1, 2, \ldots, n-1 \\ 1 & if\ x \ge Y_{(n)} \end{cases}$$

In the combined ordered arrangement of the $m + n$ sample observations, $S_m(x)$ and $S_n(x)$ are the respective proportions of X and Y observations which do not exceed the specified value of x.

Here, $H_0 : F_Y(x) = F_X(x)\ \forall x$

If H_0 is true the population distributions are identical and wee have two samples from the same population.

The empirical distribution function for X and Y sample are reasonable estimates of their respective population CDFs. Therefore, allowing for sampling variation, there should be reasonable agreement between the two empirical distributions if H_0 is true; otherwise the data suggests that H_0 is not true & therefore should be rejected.

- ***Q. How dose do the two empirical cdf's have to be so that they could be viewed as not significantly different, taking account of the sampling versatility ?***

The two sided K-S two sample test criterion, denoted by $D_{m,n}$, is based on the absolute difference between the two empirical distributions.

$$D_{m,n} = \max_x |S_m(x) - S_n(x)| \qquad \left[\begin{array}{l} \text{As } \{|S_n(x) - S_m(x)| : x \in \mathbb{R}\} \text{ is a finite set} \\ \text{we can take maximum instead of supremum} \end{array} \right]$$

Since have only;y the magnitude, & not the direction, of the deviations are considered, $D_{m,n}$ is appropriate for a general two-sided alternative.

$$H_A : F_Y(x) \neq F_X(x) \text{ for some } x$$

• ***Rejection criteria :*** Reject H_0 at level of significance α if $D_{m,n} \geq c_\alpha$, where c_α is chosen such that $P_{H_0}(D_{m,n} \geq c_\alpha) \leq \alpha$

• ***p-value :*** $P_{H_0}(D_{m,n} \geq D_0)$, where D_0 is the observed value of the two sample KS statistic

▶ A method to compute $Pr_{H_0}(D_{m,n} \geq d)$ (where d is the observed value of K-S Statistic) :

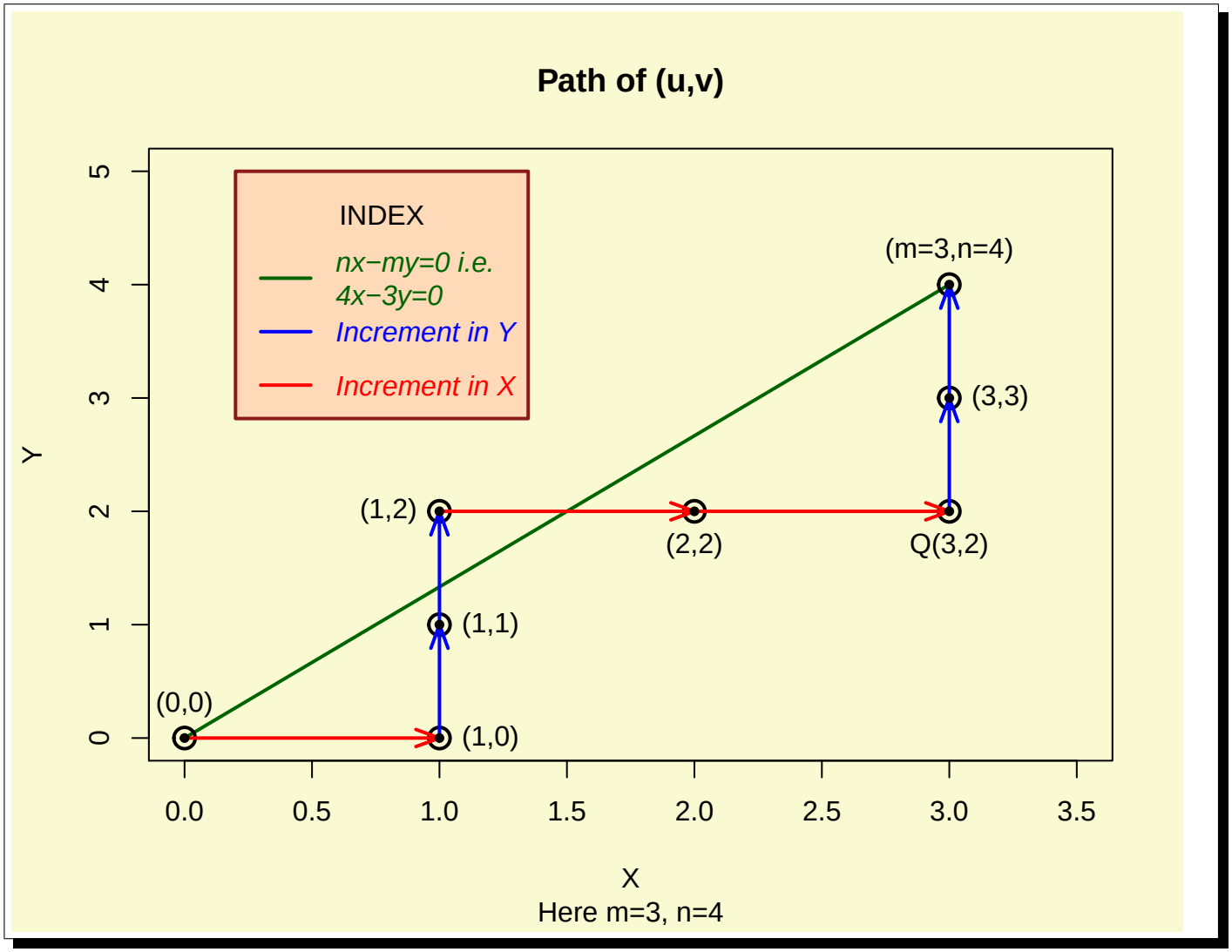
Arrange the combined sample of $m + n$ observations in increasing order of magnitude.

The arrangement can be depicted graphically on a Cartesian coordinate system by path which starts at the origin & moves one step to the right for an x observation and one step upward for an y observation, ending at (m, n).

E.g. Let the sample arrangement be-

$$\boldsymbol{xyyxxyy}$$

The observed values of $mS_m(x)$ and $nS_n(x)$ are respectively, the coordinates of all parts (u, v) on the path where u and v are integers.



The number d is the largest of the differences $|\frac{u}{m} - \frac{v}{n}| = \frac{|nu - mv|}{mn}$.

The equation of the line joining the points $(0,0)$ and (m,n) is $nx - my = 0$.

The vertical difference from any point (u,v) on the path to this line is $|v - \frac{nu}{m}|$.

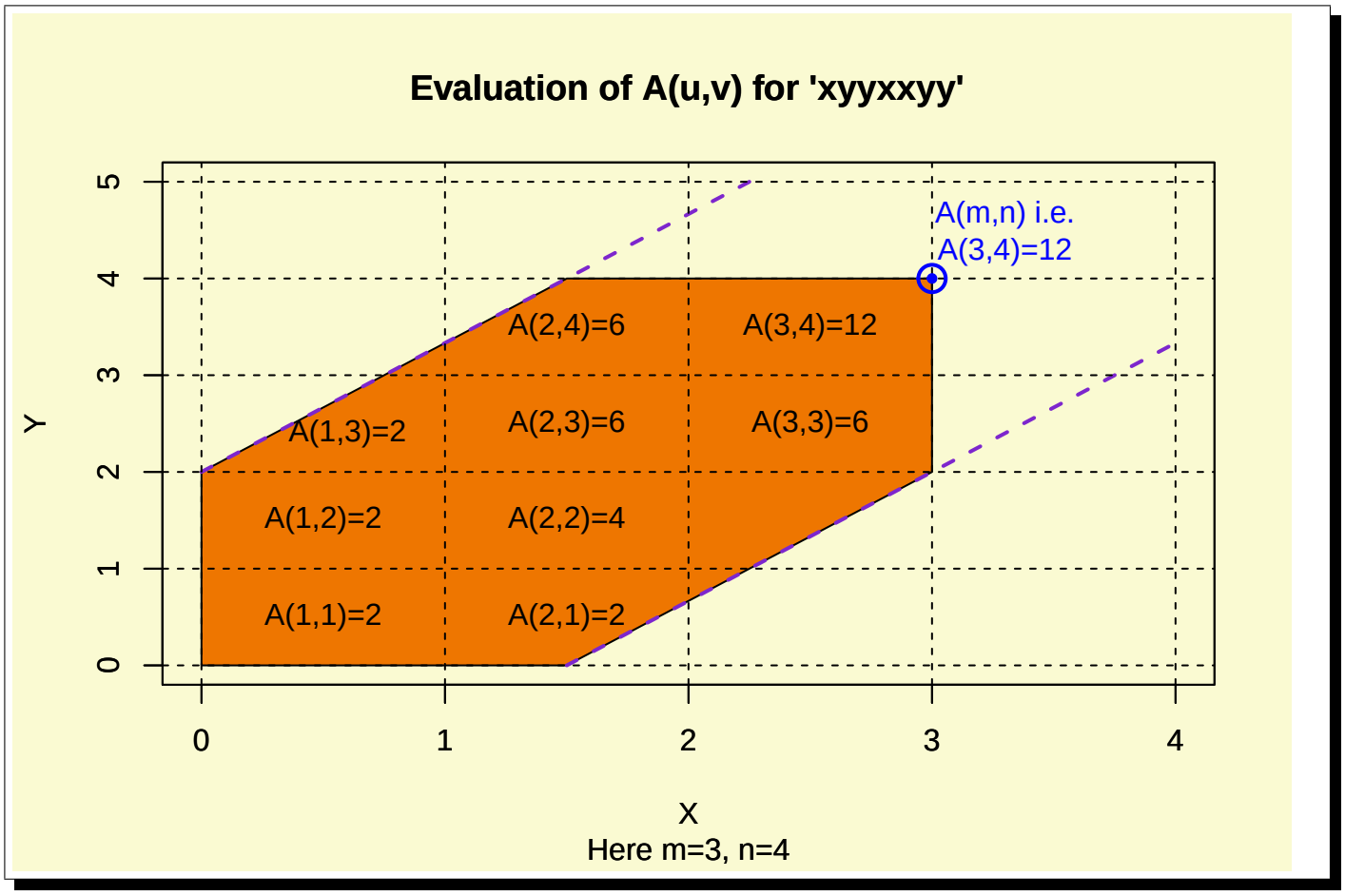
Therefore, nd_ϕ for the observed sample is the distance from thev diagonal line. The furthest point is labeled as Q , and the value of d is $\frac{2}{4}$. i.e., $nd = 2$ at Q.

The total number of arrangements of mX and nY random variables is $\binom{m+n}{n}$, and under H_0each of the corresponding paths is equally likely.

$$P_{H_0}(D_{m,n} \geq d) = \frac{\text{No. of paths which have points at a distance of not less than } nd \text{ from the diagonal line}}{\binom{m+n}{n}}$$

- ***Q. How do we count this number?***

  Ans : We draw another figure of the same dimension as before and make off two lines at vertical distance and for the diagonal.



Denote by $A(m,n)$ the number of paths from $(0,0)$ to (m,n) which lie entirely within (not on) these boundary lines. Then the desired probability is -

$$\begin{aligned} P_{H_0}(D_{m,n} \geq d) &= 1 - P_{H_0}(D_{m,n} < d) \\ &= 1 - \frac{A(m,n)}{\binom{m+n}{n}} \end{aligned}$$

The no. $A(u,v)$ at any intersection (u,v) clearly satisfies the recursion relation

$$A(u,v) = A(u-1,v) + A(u,v-1)$$

with boundary conditions,

$$A(0,v) = A(u,0) = 1$$

Thus, $A(u,v)$ is the sum of the numbers at the intersections where the previous point on the path could have been whie still in within boundaries.

Since here $A(3,4) = 12$, we have

$$Pr\{D_{3,4} \geq 0.5\} = 1 - \frac{12}{\binom{7}{4}} = 0.65714$$

As $m, n \to \infty$ in such a way that $\frac{m}{n}$ remains constant, Smirnov (1939) proved the result-

$$\lim_{m,n\to\infty} Pr\left(\sqrt{\frac{mn}{m+n}} D_{m,n} \leq d\right) = L(d)$$
$$where,\ L(d) = 1 - 2\sum_{i=1}^{\infty} (-1)^{i-1} e^{-2i^2d^2}$$

► **One sided alternative :**

$$D^+_{m,n} = \max_x (S_m(x) - S_n(x))$$

One may use this statistic to test

$$\begin{aligned} H_0 :& F_Y(x) = F_X(x) \qquad \forall x \in \mathbb{R} \\ Vs.\ H_1 :& F_Y(x) \leq F_X(x) \qquad \forall x \in \mathbb{R} \\ & F_Y(x) < F_x(x) \ \ for\ some\ x \end{aligned}$$

• **Rejection Criteria :** $D^+_{m,n} \geq c_\alpha$

Remarks :

1. The one sided test based on $D^+_{m,n}$ is also distribution free.
2. The grgaphic method described for $D_{m,n}$ can be applied here to calculate $Pr_{H_0}\{D^+_{m,n} \geq d\}$. The point Q^+, corresponding to Q , would be the point farthest below the diagonal line , and $A(m,n)$ is the no. of points lying entirely above the lower boundary line.
3. $\lim_{m,n\to\infty} Pr(\sqrt{\frac{mn}{m+n}} D^+_{m,n} \leq d) = 1 - e^{-2d^2}$

► **Ties:**

Ties within and across samples can be handled by considering only the r distinct ordered observations in the combined sample as values of x in computing $S_m(x)$ and $S_m(x)$ for $r \leq m$ & $r \leq n$. Then w efind the empirical cdf for each different x and their difference at these observations and calculate the statistic in the usual way.

H. Median Test

Let $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ be independent samples from two absolutely continuous distribution function $F_X(\cdot)$ and $F_Y(\cdot)$ respectively.

► **Target :**

To test $H_0 : F_X(x) = F_Y(x)\ \forall x \in \mathbb{R}$ *or*, $M_Y = M_X$ against $H_{1A}/H_{2A}/H_{3A}$ where,

$$\begin{aligned} where, H_{1A} &: X \geq_{st} Y && \text{or } M_X > M_Y \\ H_{2A} &: Y \geq_{st} X && \text{or } M_Y > M_X \\ H_{3A} &: F_X(x) \neq F_Y(x) \text{ for some } x \in \mathbb{R} && \text{or } M_Y \neq M_X \end{aligned}$$

where M_X and M_Y are respective medians of the populations from where the samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ are drawn from.

► **Method :**

First, we form combined ordered sample of $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$. Let "δ" be the median of combined sample.

If $m + n$ is odd, the median is the $(\frac{m+n+1}{2})^{th}$ value in the ordered arrangement.

If $m + n$ is even, the median is any number between the two middle values.

Let V be the number of observed values of X that are less than "δ".

- ***Q. What dose large value of*** V ***indicates?***

  Ans : It indicates that the actual median of X is smaller than the median of Y. One therefore rejects $H_0 : F_X = F_Y$ in favour of $H_{2A} : Y \geq_{st} X$

  If, however, the alternative is

  $H_{1A} : X \geq_{st} Y$, then the median test reject H_0 for small values of V.

  For two sided alternative we use two sided test.

Alternative	**Rejection Region**	**p-value**
$H_{1A}: \quad X \geq_{st} Y$ or $M_X > M_Y$	$V \leq c_\alpha$	$P_{H_0}(V \leq V_0)$
$H_{2A}: \quad X \leq_{st} Y$ or $M_X < M_Y$	$V \geq c'_\alpha$	$P_{H_0}(V \geq V_0)$
$H_{3A}: \quad F_X(x) \neq F_Y(x)$ for some $x \in \mathbb{R}$ or $M_Y \neq M_X$	$V \leq c'_\alpha$ or $V \geq c'_\alpha$	$2 \times$ (smallest of the above)

Where c_α and c'_α are respectively, the largest and the smallest integers such that $P_{H_0}(V \leq c_\alpha) \leq \alpha$ and $P_{H_0}(V \geq c'_\alpha) \leq \alpha$, c and c' are two integers, $c < c'$ such that $P_{H_0}(V \leq c) + P_{H_0}(V \geq c') \leq \alpha$ and V_0 is the observed value of the median test statistic V.

Null Distribution of V :

Case (i) : $m + n = 2p$, $p \in \mathbb{N}$

$$P_{H_0}(V = v) = P_{H_0}(\ exactly \quad v \quad of \quad the \quad x'_i s \leq \ combined \quad median\ )$$
$$= \begin{cases} \frac{\binom{m}{v}\binom{n}{p-v}}{\binom{m+n}{p}} & v = 0, 1, 2, ..., min(m, p) \\ 0 & otherwise \end{cases}$$

Case (ii) : $m + n = 2p + 1$, $p \in \mathbb{N}$

Here $(\frac{m+n+1}{2})^{th}$ value is the median in the combined sample, and

$$P_{H_0}(V = v) = P_{H_0}(\text{exactly } v \text{ of the } x_i\text{'s are below } (p+1)^{th}\text{value in the ordered arrangement})$$

Remark:

1. Under H_0, we accept $[m/2]$ values of x above "δ" and $[m/2]$ values of x below "δ" . Similar is the for y. One can, therefore use the χ^2- test of significance with 1 d.f. for testing against the both sided alternative.

	No. of X's	**No. of Y's**	**Total**
$> \delta$	m_1; expected: (m/2)	n_1; excpected:(n/2)	$m_1 + n_1$
$< \delta$	m_2; expected: (m/2)	n_2; expected : (n/2)	$m_2 + n_2$
	m	n	$m + n$

2. The test can be easily generalised to test for $H_0 : p^{th}$ order percentile of the two distributions are equal. Under H_0, one would expect $[mp]$ obserevations of x below the p^{th} percentile and $m - [mp]$ observations above the p^{th} percentile.

Similar for y.

▶ Confidence Interval (Median Test) :

$H_0 : F_X(x) = F_Y(x)\ \forall x \in \mathbb{R}$ Vs. $H_1 : F_X(x) \neq F_Y(x)$ for some $x \in \mathbb{R}$
Equivalently,
$H_0 : \theta = 0$ Vs. $H_1 : \theta \neq 0$, where θ is the shift in the location parameter.
(Because of the assumption that the only difference could be due to the location parameter)

▶ Test Statistic :

V =No. of observed values of X that are less than
the sample median of the combined sample.

► **Rejection Region : (At α level of significance)**

Reject H_0 in favour of H_1 at α level of significance if either $V \leq c$ or $V \geq c'$, when c and c' are chosen such that,

$$Pr_{H_0}\{V \leq c\} + Pr_{H_0}\{V \geq c'\} =\leq \alpha$$

- ***Q.*** $\boldsymbol{Pr_{H_0}(V = v) =?}$

For $m + n = 2p$, $p \in \mathbb{N}$

The median is any value between the p^{th} and $(p+1)^{th}$ ordered values.

$$P_{H_o}(V = v) = \begin{cases} \dfrac{\binom{m}{v}\binom{n}{p-v}}{\binom{m+n}{p}} & \text{for } 0 \leq v \leq \min\{m, p\} \\ 0 & \text{otherwise} \end{cases}$$

For $m + n = 2p + 1, p \in \mathbb{N}$, the median is the $(p-1)^{th}$ value.

$$P_{H_0}(V = v) = \begin{cases} \dfrac{\binom{m}{v}\binom{n}{p-v}}{\binom{m+n}{p}} & \text{for } v = 0, 1, 2, ..., \min\{m, p\} \\ 0 & \text{otherwise} \end{cases}$$

Data : $X_1, X_2, ..., X_n$; $Y_1, Y_2, ..., Y_n$

Target : To find CI for the shift parameter.

If θ were known we could form the derived random variables $X_1, X_2, ..., X_n$ and $Y_1-\theta, Y_2-\theta, ..., Y_n-\theta$ and these would constitute samples from identical populations.

From (1) it is clear that for α level of significance, the corresponding acceptance region for μ is $[c+1, c'-1]$. Making use of this fact we shall find $100(1-\alpha)\%$ C.I. for θ.

Remember $100(1-\alpha)$ CI for θ is all those values of θ for which H_0 will be accepted at significance level α.

Method :

Order the two derived samples respectively from smallest to largest as $X_{(1)}, X_{(2)}, ..., X_{(n)}$ and $Y_{(1)} - \theta, Y_{(2)} - \theta, ..., Y_{(n)} - \theta$

Let $p = \dfrac{m+n}{2}$ or $\dfrac{m+n-1}{2}$ according as $(m+n)$ is even or odd.

The p smallest observations of $N = m + n$ total are made up of exactly iX and $(p-i)Y$ variables if each of the set $X_{(1)}, X_{(2)}, ..., X_{(i)}, Y_{(1)} - \theta, Y_{(2)} - \theta, ..., Y_{(p-i)} - \theta$ is less than each observation of the set $X_{(i+1)}, ..., X_{(m)}, y_{(p-i+1)} - \theta, ..., y_{(n)} - \theta$.

Eg:

Let $m = 7$, $n = 8$

here $m + n = 15$ (an odd no.)

$\therefore p = \frac{m+n-2}{2} = \frac{15-1}{2} = 7$

$\Rightarrow$ 7 observations in the combined sample are less than the sample median of the combined sample.

Let us assume that these 3 observations from "$(Y - \theta)$" sample less tha =n the combined sample median. i.e., $i = 3$, $p - i = 4$

The corresponding observations of first set are $X_{(1)}, X_{(2)}, X_{(3)}, y_{(1)} - \theta, y_{(2)} - \theta, y_{(3)} - \theta$.

Second set : $X_{(4)}, X_{(5)}, X_{(6)}, X_{(7)}, y_{(5)} - \theta, y_{(6)} - \theta, y_{(7)} - \theta, y_{(8)} - \theta$.

The value of "i" is atleast $(c+1)$ iff for $i = c+1$, the largest X in the first set is less than the smalesst Y in the second set. i.e.,

$X_{(c+1)} < Y_{(p-c)} - \theta$

Proof :

Let us assume that $i \geq c + 1$

Target :

To prove that for $i = c + 1$,

$$X_{(c+1)} < Y_{(p-c)} - \theta$$

If $i = c + 1$,

First set : $X_{(1)}, X_{(2)}, ..., X_{(c)}, X_{(c+1)}, Y_{(1)}, Y_{(2)}, ..., Y_{(p-(c+1))} - \theta$

Second set : $X_{(c+2)}, X_{(c+3)}, ..., X_{(m)}, Y_{(p-c)} - \theta, Y_{(p-(c-1))} - \theta, ..., Y_{(n)} - \theta$

Clearly, $X_{(c+1)} < Y_{(p-c)} - \theta$

Also if, $i = c + 2$,

First set : $X_{(1)}, X_{(2)}, ..., X_{(c)}, X_{(c+1)}, X_{(c+2)}, Y_{(1)}, Y_{(2)}, ..., Y_{(p-(c+2))} - \theta$

Second set : $X_{(c+3)}, X_{(c+4)}, ..., X_{(m)}, Y_{(p-(c+1))} - \theta, Y_{(p-c)} - \theta, ..., Y_{(n)} - \theta$

Clearly, $X_{(c+2)} < Y_{(p-(c+1))} - \theta$

$< Y_{(p-c)} - \theta$ ..... (3)

Also, $X_{(c+1)} < X_{(c+2)}$ ..... (4)

Combining (iii) and (iv),

$X_{(c+1)} < Y_{(p-c)} - \theta$

Conclusion : Whenever $i \geq c + 1$, $X_{(c+1)} < Y_{(p-c)} - \theta$

Now, suppose assume that $X_{(c+1)} < Y_{(p-c)} - \theta$

Target :

To prove that, $i \geq c + 1$

Suppose not. i.e., assume $i < c + 1$

In particular, take $i = c$.

Then, first set : $X_{(1)}, X_{(2)}, ..., X_{(c-1)}, X_{(c)}$

$Y_{(1)} - \theta, Y_{(2)} - \theta, ..., Y_{(p-1)} - \theta$

Second set : $X_{(c+1)}, X_{(c+2)}, ..., X_{(m)}$
$Y_{p-(c-1)} - \theta, Y_{p-(c-2)} - \theta, ..., Y_{(n)} - \theta$

Clearly, $X_{(c+1)} > Y_{(p-c)} - \theta$

Which is a contradiction to the fact that $X_{(c+1)} < Y_{(p-c)} - \theta$

$\therefore$ Our assumption that $i = c$ is false.

Similarly , we can S.T. the assumption "$i < c$" is false too.

"i" should at least $c + 1$

Hence the result.

Similarly, $X_{(i)} > Y_{(p-c'+1)} - \theta$ can be seen to be a n.a.s.c. for having atmost $(c' - 1)$ X observations among the p smallest of the total $(m + n)$ (Exercise)

$\therefore$ We accept H_0 at significance level α if

$$\begin{aligned} & X_{(c+1)} < Y_{(p-c)} - \theta \text{ and} \\ & X_{(c')} > Y_{(p-c'+1)} - \theta \\ \text{or equivalently, } & Y_{(p-c)} - X_{(c+1)} > \theta \text{ and} \\ & Y_{(p-c'+1)} - X_{(c')} < \theta \end{aligned}$$

$\therefore$ The desired confidence interval is $\left(Y_{(p-c'+1)} - X_{(c')}, Y_{(p-c'+1)} - X_{(c')}\right)$

I. Linear Rank Statistic and General Two Sample Problem

$X_1, X_2, \ldots, X_m$ $\searrow$
$Y_1, Y_2, \ldots, Y_n$ $\nearrow$ Two independent random samples from populations with continuous CDFs F_X and F_Y respectively.

$H_0 : F_X(x) = F_Y(x)\ \forall x \in \mathbb{R}$
Let $F(\cdot)$ be their common but nunspecified CDF.
Let $N = m + n$

▶ **Definition : (Rank of an observation in the combined sample)**

Assumption : No Ties.

$$r_{XY}(x_i) = \sum_{k=1}^{m} S(x_i - x_k) + \sum_{k=1}^{n} S(x_i - y_k)$$
$$\& \ r_{XY}(x_i) = \sum_{k=1}^{m} S(y_i - x_k) + \sum_{k=1}^{n} S(y_i - y_k)$$
$$where, S(u) = \begin{cases} 0, & if\ u < 0 \\ 1, & if\ u \geq 0 \end{cases}$$

Combined ordered sample can be indicated by a vector of indicator variables, $\underset{\sim}{Z} = (Z_1, Z_2, \ldots, Z_N)$ where

$$Z_i = \begin{cases} 1, & \text{if } i'th \text{ ordered observation in combined} \\ & \text{sample comes from } X \text{ sample} \\ 0, & \text{otherwise} \end{cases} \quad , \forall i = 1(1)N$$

The rank of an observation for which Z_i is indicator is "i". Thereforethe vector $\underset{\sim}{Z}$ indicates the rank-order statistics of the combined samples and in addition identifies the sample to which the observation belongs.

E.g. Let $(X_1, X_2, X_3, X_4) = (2, 9, 3, 4)$ and $(Y_1, Y_2, Y_3) = (1, 6, 10)$. Here $m = 4$ and $n = 3$.
The combined ordered sample $(1, 2, 3, 4, 6, 9, 10)$ or $(Y_1, X_1, X_3, X_4, Y_2, X_2, Y_3)$.
$\therefore \underset{\sim}{Z} = (0, 1, 1, 1, 0, 1, 0)$
Here $Z_6 = 1$.
$\therefore$ The corresponding observations belongs to Xsamples and it is X_2.

$$\therefore r_{XY}(X_2) = 6$$

▶ **Definition : (Linear Rank Statistic)**

A **linear rank statistic** $T_N(\cdot)$ is a linear function of the indicator variables $\underset{\sim}{Z}$, i.e. $T_N(\underset{\sim}{Z}) = \sum_{i=1}^{N} a_i Z_i$, where a_i are given constants called **weights** or **scores**.

Remark : In order to study test based on linear rank statistic $T_N(\cdot)$, one needs to know their distributional properties.

Result : Under $H_0 : F_X(x) = F_Y(x) = F(x)\ \forall x \in \mathbb{R}$, we have for $i, j = 1, 2, 3, \ldots, N$,

$$
\begin{aligned}
(i)\ E(Z_i) &= \frac{m}{N} \\
(ii)\ Var(Z_i) &= \frac{mn}{N^2} \\
(iii)\ cov(Z_i, Z_j) &= \frac{-mn}{N^2(N-1)}\ \forall i \neq j
\end{aligned}
$$

Proof :
Clearly, $Z_i \sim Ber(1, \frac{m}{N})$ under H_0 for $i = 1, 2, 3, \ldots, N$.

$$
\begin{aligned}
\therefore E_{H_0}(Z_i) &= Pr_{H_0}\{Z_i = 1\} \\
&= \frac{m}{N} \\
and\ Var_{H_0}(Z_i) &= \frac{mn}{N^2}\ for\ i \neq j
\end{aligned}
$$

$$
\begin{aligned}
E_{H_0}(Z_i Z_j) &= P_{H_0}(Z_i = 1 \cap Z_j = 1) \\
&= \frac{\binom{m}{2}}{\binom{N}{2}} \\
&= \frac{m(m-1)}{N(N-1)}
\end{aligned}
$$

$$
\begin{aligned}
\therefore cov(Z_i, Z_j) &= E(Z_i E_j) - (E(Z_i))^2 \\
&= \frac{m(m-1)}{N(N-1)} - (\frac{m}{N})^2 \\
&= \frac{-mn}{N^2(N-1)}
\end{aligned}
$$

Result : Under $H_0 : F_X(x) = F_Y(x) = F(x)$, say $\forall\, x \in \mathbb{R}$

$$E(T_N) = m\sum_{i=1}^{N}\frac{a_i}{N}$$

$$\& \; Var(T_N) = \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}a_i^2 - \left(\sum_{i=1}^{N}a_i\right)^2\right]$$

$$= \frac{mn}{N(N-1)}\sum_{i=1}^{N}(a_i-\bar{a})^2 \quad \text{where } \bar{a} = \frac{1}{N}\sum_{i=1}^{N}a_i$$

Proof :

$$\begin{aligned} E_{H_0}(T_N) &= E_{H_0}\left(\sum_{i=1}^{N}a_iZ_i\right) \\ &= \sum_{i=1}^{N}a_iE_{H_0}(Z_i) \\ &= \sum_{i=1}^{N}a_i\frac{m}{N} \\ &= \tfrac{m}{N}\sum_{i=1}^{N}a_i \end{aligned}$$

$$\begin{aligned} Var_{H_0}(T_N) &= Var_{H_0}\left(\sum_{i=1}^{N}a_iZ_i\right) \\ &= \sum_{i=1}^{N}Var_{H_0}(a_iZ_i) + \sum\sum_{i\neq j}a_ia_jcov_{H_0}(Z_i,Z_j) \\ &= \sum_{i=1}^{N}a_i^2Var_{H_0}(Z_i) + \sum\sum_{i\neq j}a_ia_jcov_{H_0}(Z_i,Z_j) \\ &= \frac{mn\sum_{i=1}^{N}a_i^2}{N^2} - \frac{mn\sum\sum_{i\neq j}a_i^2}{N^2(N-1)} \\ &= \frac{mn}{N^2(N-1)}\left[(N-1)\sum_{i=1}^{N}a_i^2 - \sum\sum_{i\neq j}a_ia_j\right] \\ &= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}a_i^2 - \sum_{i=1}^{N}a_i^2 - \sum\sum_{i\neq j}a_ia_j\right] \\ &= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N}a_i^2 - \left(\sum_{i=1}^{N}a_i\right)^2\right] \end{aligned}$$

$$= \frac{mn}{N^2(N-1)}\left[N\sum_{i=1}^{N} a_i^2 - (N\bar{a})^2\right]$$
$$= \frac{mn}{N(N-1)}\left[\sum_{i=1}^{N} a_i^2 - N\bar{a}^2\right]$$
$$= \frac{mn}{N(N-1)}\left[\sum_{i=1}^{N} (a_i - \bar{a})^2\right]$$

Result :

If $B_N = \sum_{i=1}^{N} b_i Z_i$ and $T_N = \sum_{i=1}^{N} a_i Z_i$ are two linear rank statistics, under H_0, then

$$cov(B_N, T_N) = \frac{mn}{N^2(N-1)}\left(N\sum_{i=1}^{N} a_i b_i - \sum_{i=1}^{N} a_i \sum_{i=1}^{N} b_i\right)$$

Remarks :

1. The result discussed so far help us in finding the exact moment under H_0 for any linear rank statistic.
2. The exact null distribution of T_N depends on the probability of the vector Z.

▶ **Null Distribution of T_N:**

There are $\binom{N}{m}$ many distinguishable $\underset{\sim}{Z}$ vectors (such that in each such vector there are m ones and n zeros) and all those vectors are equally likely under H_0.

Therefore, the probability of any such vector (specific vector with particular arrangement of m ones and n zeros) is $\dfrac{1}{\binom{N}{m}}$.

- ***Q. What is*** $P_{H_0}\{T_N(Z) = k\}$ ***for any*** $k \in \mathbb{R}$***?***

  Ans : First we need to find the number of 2 vectors that lead to the constant "$T_N(Z) = k$". Let $t(k)$ be the total no of arrangements of mX and nY random variables such that $T_N(Z) = k$. All these arrangements are equally likely and obviously mutually eclusive.

$$\therefore P_{H_0}(T_N(Z) = k) = \frac{t(k)}{\binom{N}{m}}$$

Remarks :

1. The tediousness of enumerations increases rapidly as m and n increases.
2. When the null distribution of LRS is symmetric, only one half of the distribution needs to be generated.
3. The statistic $T_N(Z)$ is symmetric about its mean μ if $\forall k \neq 0$, $P(T_N(Z) - \mu = k) = P(T_N(Z) - \mu = -k)$.

► **Important results :**

1. The null distribution of $T_N(Z)$ is symmetric about its mean $\mu = \frac{m}{n}\sum_{i=1}^{n} a_i$ whenever the weights satisfy the relation $a_i + a_{N-i+1} = c$ where c is constant $\forall i = 1, 2, ..., N$
2. The null distribution of $T_N(Z)$ is symmetric about its mean for any set of weights if $m = n = \frac{N}{2}$.
3. The null distribution of $T_N(Z)$ is symmetric about its mean μ if N is even and the weights are $a_i = i$ for $i \leq \frac{N}{2}$ and $a_i = N - i + 1$ for $i > \frac{N}{2}$.

► **The Wilcoxon Rank-Sum Test :**

Data :
Consists of two random samples. A sample from the control population and independent sample from the treatment population.

$$X_1, X_2, ..., X_m$$
$$Y_1, Y_2, ..., Y_n$$

Target :
To investigate the presence of a treatment effect "θ", that results in a shift of location.

Assumptions :

1. The observations $X_1, X_2, ..., X_m$ are a random sample from population 1; the observations $Y_1, Y_2, ..., Y_n$ are a random sample from population 2.
2. The $X's$ and $Y's$ are mutually independent.
3. Population 1 and Population 2 are continuous populations.

Let $F_X(.)$ be the cdf of population 1 and let $F_Y(.)$ be the cdf of population 2 .
$H_0 : F_X(x) = F_Y(x) = F(x)\ \forall x \epsilon \mathbb{R}$
i.e., there is no treatment effect
i.e., the samples can be thought of as a single sample from one population.
i.e., $\theta = 0$.
$H_1 : F_Y(x) = F_X(x - \theta)\ \forall x \epsilon \mathbb{R}$ and some $\theta \neq 0$
Functionally same, but shifted to the left if $\theta < 0$ and shifted to the right if $\theta > 0$.
$Y \geq_{st} X$ when $\theta > 0$
$Y \leq_{st} X$ when $\theta < 0$
X and Y are not identically distributed when $\theta \neq 0$.

- ***Q. What if F_x is CDF of normal?***

  Think

So,
$H_{1A} : \theta < 0$ or $X \geq_{st} Y$
$H_{2A} : \theta > 0$ or $X \leq_{st} Y$
$H_{3A} : \theta \neq 0$

Logic: The ranks of X's in the combined ordered arrangement of the two samples will generally be larger than the ranks of Y's if the median of X population exceeds the median of Y population.

Wilcoxon(1945) proposed a test where we accept $H_{1A} : \theta < 0$ (or $X \geq_{st} Y$) , if the sums of the ranks of the X's is too large, or $H_{2A} : \theta > 0$ ($X \leq_{st} Y$) if the sums of the ranks of X's is too samll, and the two sided alternative $H_{3A} : \theta \neq 0$, if the sums of the ranks of the X's is either too large or too small.

► **Test Statistic :**

$$W_N = \sum_{i=1}^{N} i z_i$$

(i.e., here $a_i = i,\ \forall i = 1, ..., N$)
(Recall the definition of z_i)

If there are no ties , the mean and variance of W_N under $H_0$are

$$E_{H_0}(W_N) = \frac{m}{n}\sum_{i=1}^{N} i = \frac{m(N+1)}{2}$$

$$Var_{H_0}(W_N) = \frac{mn}{N(N+1)}[\sum_{i=1}^{N}(a_i - \bar{a})^2]$$

Here, $a_i = i\ \forall i = 1, 2, \ldots$

$$\therefore Var_{H_0}(W_N) = \frac{mn(N+1)}{12} \text{ (Verify)}$$

Also , $a_i + a_{N-i+1} = i + N - i + 1 = \underbrace{N+1}_{Constant}, \forall i = 1, 2, ...$

$\therefore$ The distribution of W_N is symmetric about its mean under H_0.

If $m \leq n$, W_n has a minimum value of $\sum_{i=1}^{M} i = \frac{m(m+1)}{2}$ and a maximum value of $\sum_{i=N-m+1}^{N} \frac{m(2N-m+1)}{2}$

$m = 3$ and $n = 4$

- ***Q. S.T. the range of W_N will be between 6 and 18.***
- ***Q. Is the distribution of W_N symmetric about 12.***
- ***Q. Find the null distribution of W_N.***

The appropriate rejection required and p values for $m \leq n \geq 10$ tables are provided-

Alternative	**Rejection Criteria**	**p- values**
$\theta < 0 (X \geq_{st} Y)$	$W_N \geq W_\alpha$	$P_{H_0}\{W_N > W_\alpha\}$
$\theta > 0 (X \leq_{st} Y)$	$W_N \leq W_\alpha$	$P_{H_0}\{W_N \leq W_\alpha\}$
$\theta \neq 0$	$W_N \geq W_\alpha$or $W_N \leq W_\alpha$	2(smaller of the above)

J. The Mann-Whitney U Test

Based on the idea that the particular pattern exhibited by the combined arrangement of the X and Y random variables in increasing order of magnitude provides information about the relationship between their populations.

Here we consider the magnitudes of, say, the Y's in relation to the X's, that is, the position of the Y's in the combined ordered sequence.

A sample pattern of arrangement where most of the Y's are greater than most of the X's, or vice versa, would be evidence against a random mixing and thus tend to discredit the null hypothesis of identical distributions.

► **Test Statistic :**

$U =$ The no. of times a Y precedes an X in the combined ordered arrangement of the two independent random samples $X_1, X_2, \ldots, X_m$ & $Y_1, Y_2, \ldots, Y_n$ in a single sequence of $m + n = N$ variables increasing in magnitude.

Assumptions : Both the populations are continuous. Therefore $Pr\{X_i = Y_j\} = 0 \quad \forall i \neq j$.

Let

$$D_{ij} = \begin{cases} 1 & if\ Y_j < X_i \\ 0 & if\ Y_j > X_i \end{cases}$$
$$for\ i = 1, 2, \ldots, m$$
$$j = 1, 2, \ldots, n$$

$$\therefore U = \sum_{i=1}^{m} \sum_{j=1}^{n} D_{ij}$$

Recall :

$$H_0 : F_X(x) = F_Y(x)\ \forall x \in \mathbb{R}$$

$$H_{1A} : F_X(x) \leq F_Y(x)\ \forall x \in \mathbb{R}$$
$$\& \ F_X(x) \leq F_Y(x)\ for\ some\ x \in \mathbb{R}$$
$$i.e.\ X \geq_{st} Y$$

$$H_{2A} : F_X(x) \geq F_Y(x)\ \forall x \in \mathbb{R}$$
$$\& \ F_X(x) > F_Y(x)\ for\ some\ x \in \mathbb{R}$$
$$i.e.\ Y \geq_{st} X$$

$$H_{3A} : F_X(x) \neq F_Y(x)\ for\ some\ x \in \mathbb{R}$$

► **Rejection Criteria :**

We reject H_0 in favour of H_{1A} for larger values of U. Similarly we reject H_0 in favour of H_{2A} for smaller values of U.

For H_0Vs. H_{3A}, we reject H_0 if either U is too small or too large.

Let us consider the problem

$$\begin{aligned} H_0 : & F_X(x) = F_Y(x) \ \forall x \in \mathbb{R} \\ Vs.\ H_{2A} : & F_Y(x) \leq F_X(x) \ \forall x \in \mathbb{R} \\ & F_Y(x) < F_X(x) \text{ for some } x \in \mathbb{R} \end{aligned}$$

Let

$$\begin{aligned} p &= P(Y < X) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_{XY}(x,y)dydx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_X(x) f_Y(y)dydx \\ &= \int_{-\infty}^{\infty} F_Y(x) f_X(x)dx \\ &= \int_{-\infty}^{\infty} F_Y(x) dF_X(x) \end{aligned}$$

Under H_0, $p = 0.5$

i.e., $p = \int_{-\infty}^{\infty} F_Y(x) dF_X(x) = 0.5$

Under H_{2A}, $p < 0.5$

Thus the hypothesis can be reqritten as $H_0 : p = 0.5$ vs. $H_{2A} : p < 0.5$
Clearly

$$\begin{aligned} D_{ij} &\sim Ber(p) \\ \therefore E(D_{ij}) &= p = E(D_{ij}^2) \\ and\ Var(D_{ij}) &= p(1-p) \end{aligned}$$

Also

$$\begin{aligned} cov(D_{ij}, D_{ik}) &= 0 \text{ for } i \neq j \\ \underset{j \neq k}{cov}(D_{ij}, D_{ik}) &= p_1 - p^2 \\ \underset{i \neq h}{cov}(D_{ij}, D_{hj}) &= p_2 - p^2 \end{aligned}$$

where

$$\begin{aligned} p_1 &= P(Y_j < X_i \cap Y_k < X_i) \\ &= P(Y_j \& Y_k < X_i) \\ &= \int_{-\infty}^{\infty} [F_Y(x)]^2 dF_X(x) \end{aligned}$$

[Reason :

$$\begin{aligned} & P(Y_j \& Y_k < X_i) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x}\int_{-\infty}^{x} f_{X_i Y_j Y_k}(x, y_j, y_k) dy_j dy_k dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x}\int_{-\infty}^{x} f_{X_i}(x) f_{Y_j}(y_j) f_{Y_k}(y_k) dy_j dy_k dx \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{x} F_Y(x) f_{Y_k}(y_k) f_{X_i}(x) dy_k dx \\ &= \int_{-\infty}^{\infty} F_Y(x) F_Y(x) f_{X_i}(x) dx \\ &= \int_{-\infty}^{\infty} [F_Y(x)]^2 f_{X_i}(x) dx \\ &= \int_{-\infty}^{\infty} [F_Y(x)]^2 f_X(x) dx \quad] \end{aligned}$$

Similarly $p_2 = \int_{-\infty}^{\infty} (1 - F_X(y))^2 dF_Y(y)$

Recall :

$$\begin{aligned} U &= \sum_{i=1}^{m}\sum_{j=i}^{n} D_{ij} \\ \therefore E(U) &= \sum_{i=1}^{m}\sum_{j=i}^{n} E(D_{ij}) \\ &= mnp \end{aligned}$$

Also

$$\begin{aligned} Var(U) = & \sum_{i=1}^{m}\sum_{j=i}^{n} Var(D_{ij}) + \sum_{i=1}^{m}\sum_{1\le j\ne k\le n}\sum cov(D_{ij}, D_{ik}) \\ & + \sum_{j=1}^{n}\sum_{1\le i\ne h\le m}\sum cov(D_{ij}, D_{hj}) + \sum_{1\le i\ne h\le m}\sum\sum_{1\le j\ne k\le n}\sum cov(D_{ij}, D_{hk}) \\ &= mnp(1-p) + mn(n-1)(p_1 - p^2) + nm(m-1)(p_2 - p^2) \\ &= mn[p - p^2(N-1) + (n-1)p_1 + (m-1)p_2] \end{aligned}$$

Null Distribution of U :

under H_0, each of the $\begin{pmatrix} m+n \\ m \end{pmatrix}$ arrangements of the random variables into a combined sequences occurs with equal probability, so that

$$f_U(u) = P(U = u) = \frac{r_{m,n}(u)}{\begin{pmatrix} m+n \\ m \end{pmatrix}}$$

where $r_{m,n}(u)$ is the number of distinguishable arrangements of the mX and nY random variables such that in each sequence the number of times a Y proceeds on X is exactly u.

Range of $U : 0, 1, ..., mn$

► **Claim :** The null distribution of U is symmetric about mean $\frac{mn}{2}$.

Proof : For every particular arrangement of z of the mx and ny letters, define the conjugate arrangements of z' as the sequence z written backward. In other words, if z denotes a set of numbers written from smallest to largest, z' denotes the same numbers written from largest to smallest.

Every y that proceeds an x in z then follows that x in z', so that if u is the value of the Mann-Whitney statistic for z, $mn - u$ is the value for z'.

$\therefore$ Under H_0, $r_{m,n}(u) = r_{n,m}(mn - u)$ or equivalently,

$P(U - \frac{mn}{2} = u) = P(U = \frac{mn}{2} + u) = P(U = mn - (\frac{mn}{2} - u)) = P(U = \frac{mn}{2} - u) = P(U - \frac{mn}{2} = u))$

$\therefore$ U is symmetric about $\dfrac{mn}{2}$ under H_0.

Benefit : Only lower tail critical values need to be found for either a one or two-sided test.

Let $U' = \sum_{i=1}^{m} \sum_{j=1}^{n} (1 - D_{ij})$

Alternative	Rejection region	p-value
$p < 0.5$ or $Y \underset{st}{\geq} X$	$U \leq c_\alpha$	$P_{H_0}(U \leq u_0)$
$p > 0.5$ or $Y \underset{st}{\leq} X$	$U' \leq c_\alpha$	$P_{H_0}(U' \leq u_0)$
$p \neq 0.5$ or $F_Y(x) \neq F_X(x)$ for some $x \in \mathbb{R}$	$U \leq c_{\alpha/2}$ w $U' \leq c_{\alpha/2}$	$2(smaller\ of\ \ the\ \ above\ )$

► The problem of ties :

If ties occur within one or both samples, a unique value of U is obtained. However, if one or more X is tied with one or more Y, our definition requires that the ties be broken in some way.

The conservative approach may be adapted, which means that all ties are broken in all possible ways and the largest resulting value of U (or U') is used in reaching the decision.

Another approach :

Define,

$$U_T = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}$$

$$\text{where, } D_{ij} = \begin{cases} 1 & X_i > Y_j \\ 0.5 & X_i = Y_j \\ 0 & X_i < Y_j \end{cases}$$

Let

$$p^+ = Pr\{x > Y\}$$
$$\& \; p^- = Pr\{X < Y\}$$
$$E(U_T) = mn(p^+ - p^-) \; (verify)$$

$$Under \; H_0, p^+ = p^-$$
$$\therefore E_{H_0}(U_T) = 0$$

Also

$$Var(U_T|H_0) = \frac{mn(N+1)}{12}\left[1 - \frac{\sum t(t^2+1)}{N(N^2-1)}\right]$$

where + denoted the multiplicity of a tie and the sum is extended over all t ties,

▶ Confidence Interval for θ :

$$F_Y(x) = F_X(x-\theta) \; \forall x \; \& \; some \; \theta \in \mathbb{R}$$

Under this assumption , the sample observations $X_1, X_2, \ldots, X_m$ and $Y_1 - \theta, Y_2 - \theta, \ldots, Y_n - \theta$ come from identical populations.

A CI for θ with confidence coefficient $1 - \alpha$ consists of all those values of θ from which the H_0 will be accepted at significance level α.

The random variable U denotes the number of times a $Y - \theta$precedes an X, that is, the number of pairs $(X_i, Y_j - \theta),\ i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$ for which $X_i > Y_j - \theta$ or equivalently $Y_j - X_i < \theta$.

If a table of critical values for a two sided U test at level α gives a rejection region $U \leq k$, say, we reject H_0 when no more than k differences are less than θ. The total number of differences $Y_j - X_i$ is mn. If these differences are ordered from smallest to largest according to actual maginitude, denoted by $D_{(1)}, D_{(2)}, ..., D_{(mn)}$, there are exactly k differences less than θ if θ is the $(k+1)$st - ordered difference, $D_{(k+1)}$. Any number exceeding this $(k+1)$ st difference will produce more than k differences less than θ. Therefore, the lower limit of the confidence interval for θ is $D_{(k+1)}$. Similarly , since the probability distribution of U is symmetric, an upper confidence limit is given by that difference which is $(k+1)^{th}$ from the largest, that is $D_{(mn-k)}$.

$\therefore$The confidence interval with confidence coefficient $(1-\alpha)$ is $\left(D_{(k+1)}, D_{(mn-k)}\right)$

E.g. -

$m = 3$, $n = 5$, $\alpha = 0.1$

$$Pr\{u < 1\} = 2/56 = 0.036$$
$$Pr\{u < 2\} = \frac{4}{56} = 0.071$$

$\therefore$ The critical value when $\alpha/2 = 0.05$, is 1 with type-I error probability (exact) $= 2 \times 0.036 = 0.072$
The confidence interval is $\left(D_{(2)}, D_{(14)}\right)$.

Example :

Data :

$$X : 1, 6, 7$$
$$Y : 2, 4, 9, 10, 12$$

$\alpha = 0.10$, $\theta_1 = 1$, $\theta_2 = 6$, $\theta_3 = 7$
Show that exact type-I error probability is 0.928
Also CI is (-4,9).

▶ U and W are equivalent test statistics.

Proof :
Let

$$D_{ij} = \begin{cases} 1 & ,if.Y_j < X_i \\ 0 & ,if, Y_j > X_i \end{cases}$$
$$U = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij}$$

Also $W = \sum_{i=1}^{m} Q_i$,where Q_i = rank of X_i in the combined sample

$$\begin{aligned} Q_i &= \text{Rank of } X_i\text{in the combined sample} \\ &= \text{Number of } Y_j's < X_i + \text{rank of } X_i \text{ in } X's \\ &= \sum_{j=1}^{n} D_{ij} + \text{rank of } X_i \text{ in } X's \end{aligned}$$

$$W = \sum_{i=1}^{m} Q_i = \sum_{i=1}^{m}\sum_{j=1}^{n} D_{ij} + \frac{n(n+1)}{2} = U + \frac{n(n+1)}{2}$$

K. The Kruskal-Wallis One-Way ANOVA Test

Here, the interest is centered on the relative locations (median) of three or more populations.

Data : The data consist of $N = \sum_{j=1}^{k} n_j$ observations, with n_j observations from the j^{th} population/treatment, $j = 1, 2, \ldots, k$.

Treatments			
1	2	$\cdots$	k
X_{11}	X_{12}	$\cdots$	X_{1k}
X_{21}	X_{22}	$\cdots$	X_{2k}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$X_{n_1 1}$	$X_{n_2 2}$	$\cdots$	$X_{n_k k}$

Assumptions :

1. The N random variables $\{X_{1j}, X_{2j}, \ldots, X_{n_j j}\}$, $j = 1, 2, \ldots, k$ are mutually independent.
2. $\forall j \in \{1, 2, \ldots, k\}$, the n_j variables $\{X_{1j}, X_{2j}, \ldots, X_{n_j j}\}$ are a random sample from a continuous distribution with distribution function F_j.
3. The distribution functions $F_1, F_2, \ldots, F_k$ are connected through the relationship $F_j(t) = F(t - \theta_j)$, $-\infty < t < \infty$ for $j = 1, 2, \ldots, k$, where F is a distribution function for a continuous distribution with unknown median θ and θ_j is the unknown treatment effect for the j^{th} population.

$$H_0 : \theta_1 = \theta_2 = \ldots = \theta_k$$
Or, equivalently, $H_0 : F_1(x) = F_2(x) = \ldots = F_k(x)\ \forall x$

Vs. $H_1 : \theta_i \neq \theta_j$ for atleast one $i \neq j$
Or, equivalently, $H_1 : F_i(x) \neq F_j(x)$ for some $x \in \mathbb{R}$ and for atleast one $i \neq j$.

Remark : In classical statistics, the usual test for this problem is the ANOVA test for a one-way classification.

Method : Since under the H_0, we have essentially a single sample of size N from the common population, combine the N observations into a single observed sequence from smallest to largest, keeping track of which observation is from which sample and assign the ranks $1, 2, ..., N$ to the sequence.

Under H_0, the total sum of the ranks $\sum_{j=1}^{n} j = \frac{N(N+1)}{2}$ would be divided proportionally according to sample size among the k samples. For the sample, which contains n_j observations, the expected sum of ranks would be $\frac{n_j}{N}\frac{N(N+1)}{2} = \frac{n_j(N+1)}{2}$.

Denote the actual sum of ranks arranged to the elements in the j^{th} sample by R_j.

$$\text{i.e. } R_j = \sum_{i=1}^{n_j} r_{ij} \; j = 1, 2, \ldots, k$$

where r_{ij} denotes the rank of X_{ij} in the joint ranking.

$$\text{Also, let } R_{.j} = \frac{R_j}{n_j} \; j = 1, 2, \ldots, k$$

A reasonable test statistic is -

$$S = \sum_{j=1}^{k} \left(R_j - \frac{n_j(N+1)}{2} \right)^2$$

H_0 is rejected for large values of S.

► **Null distribution of S : (no ties case)**

Under H_0, all $\frac{N!}{\left(\prod_{j=1}^{k}(n_j)!\right)}$ asignments of of n_1 ranks to treatment 1 observations, n_2 ranks to treatment 2 observations,..., n_k ranks to the treatment k observations are equally likely.

Each of the possibilities must be enumerated and the value of S calculated for each.

If $t(S)$ denotes the number of assignments with particular value "s" calculated from the equation $S = \sum_{j=1}^{k} (R_j - \frac{n_j(N+1)}{2})^2$, then

$$Pr_{H_0}(S = s) = \frac{t(S)}{\frac{N!}{\left(\prod_{j=1}^{k}(n_j)!\right)}} = t(S) \prod_{j=1}^{k} \frac{n_j!}{N!}$$

► **The Kruskall-Wallis test statistic :**

Kruskall and **Wallis (1952)** proposed a test statistic, which is a weighted sum of squares of deviations with the reciprocals of the respective sample sizes used as weights. And the test statistic is -

$$\begin{aligned} H &= \frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{1}{n_j} (R_j - \frac{n_j(N+1)}{2})^2 \\ &= \frac{12}{N(N+1)} \sum_{j=1}^{k} n_j (R_{.j} - \frac{(N+1)}{2})^2 \\ &= \left[\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{{R_j}^2}{n_j} \right] - 3(N+1) \end{aligned}$$

Remarks :

1. H and S are equivalent test criteria only for all n_i equal.
2. Reject H_0 if $H \geq h_\alpha$; otherwise do not reject H_0 where h_α is chosen such that $Pr_{H_0}(H \geq h_\alpha) \leq \alpha$

Some Moments :

$$\begin{aligned} E_{H_0}(R_{.j}) &= E_{H_0}\left(\frac{R_j}{n_j}\right) \\ &= \frac{1}{n_j} E_{H_0}(R_j) \\ &= \frac{1}{n_j} E_{H_0}\left(\sum_{i=1}^{n_j} r_{ij}\right) \\ &= \frac{1}{n_j} \sum_{i=1}^{n_j} E_{H_0}(r_{ij}) \\ &= \frac{1}{n_j}.n_j.\frac{(N+1)}{2} \\ &= \frac{(N+1)}{2} \end{aligned}$$

Similarly , one can show that ,

$$Var_{H_0}(R_{ij}) = \frac{(N+1)(N-n_j)}{12n_j}$$
$$Cov(R_{.i}, R_{.j}) = -\frac{N+1}{12}$$

► **Assymptotic Distribution :**

If $n_j$is large, the CLT allows us to approx. the distribution of

$$z_j = \frac{R_{.j} - \frac{(N+1)}{2}}{\sqrt{(N+1)(N-n_j)/12n_j}}$$

, by the standard normal.

Consequnetly, $z_j^2 \overset{approx.}{\sim} \chi^2_{(1)}$, $j = 1, 2, ..., k$

But z_j are not independent random variables since $\sum_{j=1}^{k} n_j R_{.j} = \frac{N(N+1)}{2}$

Kruskal (1952) showed that under H_0, if no n_j is very small, the random variable

$$\sum_{j=1}^{k} \frac{N-n_j}{N} z_j^2 = H \overset{approx.}{\sim} \chi^2_{(k-1)}$$

The approx size α rejection is $H \geq \chi^2_{\alpha,k-1}$

► **The problem of ties :**

When two or more observations are tied within a sample , the value of H is the same regardless of the method used to resolve the ties since the rank sum is not affected.

When ties occurs across samples , the mid rank method is geenrally used. Alternatively, the ties can be broken in the way that is least conductive to rejection of $H_0$for a corrective test.

Here,

$$E_{H_0}(R_{.j}) = \frac{N+1}{2}$$
$$\text{and } Var_{H_0}(R.j) = \frac{\sigma^2(N-n_j)}{n_j(N-1)}$$

where , $\sigma^2 = \frac{N^2-1}{12} - \frac{\sum t(t^2-1)}{12}$, where the sum is over all sets of ties in the population.

† * * * * THE END * * * * †