Nonparametric Inference

1) Gibbons and Chakraborty.

Rank: Suppose X_α be the α^{th} obs^n for a set of n i.i.d obs^{n}s, $\alpha = 1, 2, \ldots, n$ from a continuous dist^n $F_X(x)$.

R_α: rank of X_α or α^{th} smallest obs^n.

: # of obs^{n}s $\leq X_\alpha$

Due to continuity ranks are distinct with prob 1.

Rank is an ordered permutation.

Rank vector $\underset{\sim}{R} = \underset{\sim}{r}$ is a ordered permutation.

$$\underset{\sim}{R} = (R_1 \; R_2 \; R_3 \ldots R_n)$$

R_1: Rank of 1st obs; R_n: Rank of n^{th} obs

$\underset{\sim}{R}$ is the random vector random permutation.

Remmark: $Pr\{\underset{\sim}{R} = \underset{\sim}{r}\} = \frac{1}{n!}$

$Pr\{R_\alpha = r_\alpha\} = \frac{1}{n}$; $\alpha = 1, 2, \ldots, n$

Marginal dist^n of rank is a discreet uniform dist^n.

Remark: $Pr\{R_\alpha = r_\alpha \cap R_\beta = r_\beta\} = \frac{1}{n(n-1)}$ $\alpha \neq \beta$

Remark:

$$\left[\begin{array}{l} E(R_\alpha) = \frac{n+1}{2} \\ V(R_\alpha) = \frac{n^2-1}{12} \\ Cov(R_\alpha, R_\beta) = -\frac{n+1}{12} \end{array}\right]$$

⟨R_α and R_β are not ind⟩

⟨Also they are negatively correlated⟩

$$\left\langle Corr^n(R_\alpha, R_\beta) = -\frac{\frac{n+1}{12}}{\frac{n^2-1}{12}} = -\frac{1}{n-1} \right\rangle ; \boxed{n>1}$$

Linear rank statistic

Let $\underset{\sim}{a} = (a_1, a_2, \ldots, a_n)$ and $\underset{\sim}{b} = (b_1, b_2, \ldots, b_n)$ be two sets of coeff. (constants) based on n natural number.

Let, $\underset{\sim}{R} = (R_1, R_2, \ldots, R_n)$ be the random permutation of $\{1, 2, \ldots, n\}$. Then linear rank statistic is,

$$\left\langle T = \sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha} = a_1 b_{R_1} + a_2 b_{R_2} + \cdots + a_n b_{R_n} \right\rangle$$

a_α's are known as regression constants and b_{R_α}'s are scores. ~~constants~~

Note that, joint dist^n of rank is independent of dist^n fu^n F from which obs^n's come, the dist^n of T is ind. of any F. Hence T can be used to provide dist^n free (non-parametric test)

Also $(R_1, R_2, \ldots, R_n) \overset{\mathcal{D}}{=} (n-R_1+1, n-R_2+1, \ldots, n-R_n+1)$

■ Mean and variance of T

$$E(T) = E\left[\sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha}\right]$$

$$= \sum_{\alpha=1}^{n} a_\alpha E(b_{R_\alpha})$$

$$\therefore E(b_{R_\alpha}) = \frac{1}{n}\left[b_1 + b_2 + \cdots + b_n\right] \quad \left[\text{As the value of } R_\alpha \text{ can be } 1, 2, \ldots n.\right]$$

$$= \bar{b}$$

$$\therefore E(T) = \sum_{\alpha=1}^{n} a_\alpha \bar{b}$$

$$= n\bar{a}\bar{b} \quad \left[\text{where } \bar{a} = \frac{1}{n}\sum_{\alpha=1}^{n} a_\alpha\right]$$

$$V(T) = V\left(\sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha}\right)$$

$$= V\left(\sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha} - n\bar{a}\bar{b}\right)$$

$$= V\left(\sum_{\alpha=1}^{n} (a_\alpha - \bar{a}) b_{R_\alpha}\right)$$

$$= \sum_{\alpha=1}^{n} (a_\alpha - \bar{a})^2 V(b_{R_\alpha}) + \sum\sum_{\alpha \neq \beta} (a_\alpha - \bar{a})(a_\beta - \bar{a}) \operatorname{cov}(b_{R_\alpha}, b_{R_\beta})$$

$$= V(b_{R_\alpha}) \sum (a_\alpha - \bar{a})^2 + \operatorname{cov}(b_{R_\alpha}, b_{R_\beta}) \sum\sum_{\alpha \neq \beta} (a_\alpha - \bar{a})(a_\beta - \bar{a})$$

WLG fix $\alpha = 1$,

$$V(b_{R_1}) = E(b_{R_1}^2) - \left[E(b_{R_1})\right]^2$$

$$= \frac{1}{n}\left[b_1^2 + b_2^2 + \cdots + b_n^2\right] - \bar{b}^2$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n} b_\alpha^2 - \bar{b}^2$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n} (b_\alpha - \bar{b})^2$$

$$\therefore V(b_{R_\alpha}) = \frac{1}{n}\sum_{\alpha=1}^{n}(b_\alpha - \bar{b})^2$$

$\text{cov}(b_{R_\alpha}, b_{R_\beta})$

WLG $\alpha = 1, \beta = 2$

$$\text{cov}(b_{R_1}, b_{R_2}) = E\left[(b_{R_1} - \bar{b})(b_{R_2} - \bar{b})\right]$$

$$= \sum_{\substack{R_1=1 \\ (\alpha)}}^{n} \sum_{\substack{R_2=1 \\ R_1 \neq R_2 (\beta)}}^{n} P[R_1 = \alpha \cap R_2 = \beta] \cdot (b_\alpha - \bar{b}) \cdot (b_\beta - \bar{b})$$

$$= \frac{1}{n(n-1)} \sum_{\substack{R_1 \\ (\alpha)}} \sum_{\substack{\neq R_2 \\ (\beta)}} (b_\alpha - \bar{b})(b_\beta - \bar{b})$$

$$= \frac{1}{n(n-1)}\left[\sum_{R_1(\alpha)} (b_\alpha - \bar{b})\left\{\sum_{R_2(\beta)} (b_\beta - \bar{b}) - (b_\alpha - \bar{b})\right\}\right]$$

$$= \frac{1}{n(n-1)}\left[\left\{\sum_{\alpha} (b_\alpha - \bar{b})\right\}^2 - \sum_{\alpha}(b_\alpha - \bar{b})^2\right]$$

$$= -\frac{1}{n(n-1)}\sum_{\alpha=1}^{n}(b_\alpha - \bar{b})^2 = \boxed{-\frac{V(b_{R_1})}{n-1}}$$

$$\therefore V(T) = V(b_{R_1})\sum(a_\alpha - \bar{a})^2 - \frac{V(b_{R_1})}{(n-1)}\sum_{\alpha \neq \beta}\sum (a_\alpha - \bar{a})(a_\beta - \bar{a})$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n}(b_\alpha - \bar{b})^2 \sum_{\alpha=1}^{n}(a_\alpha - \bar{a})^2 - \frac{\sum(b_\alpha - \bar{b})^2}{n(n-1)}\sum_{\alpha \neq \beta}\sum(a_\alpha - \bar{a})(a_\beta - \bar{a})$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n}(b_\alpha - \bar{b})^2 \sum_{\alpha=1}^{n}(a_\alpha - \bar{a})^2 - \frac{\sum(b_\alpha - \bar{b})^2}{n(n-1)}\left[\sum_{\alpha}(a_\alpha - \bar{a})\left\{\sum_{\beta}(a_\beta - \bar{a}) - (a_\alpha - \bar{a})\right\}\right]$$

$$= \frac{1}{n}\sum_{\alpha=1}^{n}(b_\alpha - \bar{b})^2 \sum_{\alpha=1}^{n}(a_\alpha - \bar{a})^2 + \frac{\sum(b_\alpha - \bar{b})^2}{n(n-1)}\sum(a_\alpha - \bar{a})^2$$

$$= \left[\frac{1}{n} + \frac{1}{(n-1)n}\right]\sum(a_\alpha - \bar{a})^2 \sum(b_\alpha - \bar{b})^2$$

$$= \frac{1}{n-1}\sum_{\alpha=1}^{n}(a_\alpha - \bar{a})^2 \sum_{\alpha=1}^{n}(b_\alpha - \bar{b})^2$$

Asymptotic distⁿ of linear rank statistic :-

i) Wald Wolfowitz condⁿ
ii) Noether condⁿ
iii) Hoeffding condⁿ

$$\left\langle \frac{T - E(T)}{\sqrt{V(T)}} \xrightarrow{L} N(0,1) \right\rangle$$

Inverse permutation

Every permutation P has its mirror image P' such that, $\langle P \circ P' = I^+ \rangle = \{1, 2, \ldots, n\}$

($\circ$ ↑ operation)

That mirror image is called inverse permutation

- Inverse permutation in Ranking theory is called antirank $\underset{\sim}{D} = (D_1 \; D_2 \ldots D_n)$

$$\langle \underset{\sim}{R} \circ \underset{\sim}{D} = I \rangle$$

Antirank

In antiranking, number and number of the place the obsⁿ occupies (Rank) is exchanged.

$R \; (3 \; 1 \; 4 \; 2)$

$D \; (2 \; 4 \; 1 \; 3)$

Number	Place
3 ⟷ 1	
1 ⟷ 3	
1 ⟷ 2	
2 ⟷ 1	
4 ⟷ 3	
3 ⟷ 4	

$R \; (2 \; 3 \; 4 \; 5 \; 1)$

$D \; (5 \; 1 \; 2 \; 3 \; 4)$

Remark:

We know ranks are not independent, but the values in inverse permutation (antiranks) are ind. Moreover the joint distⁿ of any permutation remains the same. Therefore the joint distⁿ of ranking and joint distⁿ of antiranking are same.

$$\left\langle \begin{array}{l} P(\underset{\sim}{R} = \underset{\sim}{r}) = \frac{1}{n!} \\ P(\underset{\sim}{D} = \underset{\sim}{d}) = \frac{1}{n!} \end{array} \right\rangle$$

[Hence, for finding distⁿ of linear rank statistic antiranking process is more convenient.]

$Cov(R_\alpha, R_\beta)$

$= E(R_\alpha . R_\beta) - E(R_\alpha).E(R_\beta)$

$= \sum_{\alpha \neq \beta}^{n} \sum^{n} \alpha\beta . \frac{1}{n(n-1)} - \left(\frac{n+1}{2}\right)^2$

$= \frac{1}{n(n-1)} \left[(\Sigma\alpha)^2 - \Sigma\alpha^2\right] - \left(\frac{n+1}{2}\right)^2$

$= \frac{1}{n(n-1)} \left[\frac{n^2(n+1)^2}{4} - \frac{n(n+1)(2n+1)}{6}\right] - \left(\frac{n+1}{2}\right)^2$

$= -\frac{n+1}{12}.$

<u>Remark</u>

$T = \sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha} \overset{\mathcal{D}}{\equiv} \sum_{\alpha=1}^{n} a_{\mathcal{D}_\alpha} b_\alpha$

$\mathcal{D}_\alpha \hookrightarrow$ Antirank of the α^{th} obsn.

$\underset{\sim}{R} = (\underset{R_1}{3}\ \underset{R_2}{1}\ \underset{R_3}{2})$. $T = \sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha} = a_1 b_3 + a_2 b_1 + a_3 b_2$

$\underset{\sim}{\mathcal{D}} = (\underset{\mathcal{D}_1}{2}\ \underset{\mathcal{D}_2}{3}\ \underset{\mathcal{D}_3}{1})$ $T' = \sum_{\alpha=1}^{3} a_{\mathcal{D}_\alpha} b_\alpha = a_2 b_1 + a_3 b_2 + a_1 b_3$

<u>Result</u>:

For a linear rank statistic T, if either $a_\alpha + a_{n-\alpha+1}$ or $b_{R_\alpha} + b_{R_{(n-\alpha+1)}}$ is a constant $\forall\alpha$, the distn of T is symmetric ~~about~~ about its mean.

<u>Pf</u>: We know that $(R_1, R_2, \ldots, R_n) \overset{\mathcal{D}}{\equiv} (n-R_1+1, n-R_2+1, \ldots, n-R_n+1)$

Assume $b_{R_\alpha} + b_{R_{n-\alpha+1}} = k\ \forall\alpha$ where k is a constant.

$\bar{b} = \frac{k}{2}$

$\therefore T \overset{\mathcal{D}}{\equiv} \sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha} \overset{\mathcal{D}}{\equiv} \sum_{\alpha=1}^{n} a_\alpha b_{n-R_\alpha+1}$

$\overset{\mathcal{D}}{\equiv} \sum_{\alpha=1}^{n} a_{\mathcal{D}_\alpha} b_{n-\alpha+1}$

$\overset{\mathcal{D}}{\equiv} \sum_{\alpha=1}^{n} a_{\mathcal{D}_\alpha}(k - b_\alpha)$ $[\because b_\alpha + b_{n-\alpha+1} = k]$

$\overset{\mathcal{D}}{\equiv} \sum_{\alpha=1}^{n} a_{\mathcal{D}_\alpha} k - \sum_{\alpha=1}^{n} a_{\mathcal{D}_\alpha} b_\alpha$

$T \overset{\mathcal{D}}{\equiv} kn\bar{a} - \sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha}$

$\equiv k.n\bar{a} - T \quad \forall k$

Choose, $k = 2\bar{b}$ $\quad T \equiv 2n\bar{a}\bar{b} - T$

$\Rightarrow T - n\bar{a}\bar{b} \overset{\mathcal{D}}{\equiv} n\bar{a}\bar{b} - T$. <Hence the proof>

Sign Test

Let $x_1, x_2, \dots, x_n$ be a random sample from a cont. distribution $F_X(x)$. Let μ be the median of the distn.

We are to test, $H_0: \mu = \mu_0$ (a fixed constant)

$H_1: \mu > \mu_0$ <median of $F_X(x) > \mu_0$>

Let, S = no. of positive obsn. $(x_i - \mu_0 > 0)$

∴ Let us propose a linear rank statistic for testing the above,

Define, $\left[a_\alpha = \begin{cases} 1 & \text{if } x_\alpha > \mu_0 \\ 0 & \text{o.w} \end{cases} \right]$

and, $b_{R_\alpha} = 1 \ \forall \alpha$

$$\therefore \left\langle T = \sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha} = \sum_{\substack{\alpha=1 \\ x_\alpha > \mu_0}}^{n} a_\alpha = S \right\rangle$$

$$E_{H_0}(T) = E_{H_0}\left(\sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha}\right)$$
$$= E_{H_0}\left(\sum_{\alpha=1}^{n} a_\alpha\right)$$
$$= \sum_{\alpha=1}^{n} E_{H_0}(a_\alpha)$$
$$= \sum_{\alpha=1}^{n} 1.\, P_{H_0}\left[X_\alpha > \mu_0\right]$$
$$= \frac{n}{2}$$

$\left[E_{H_0}[a_\alpha] = \tfrac{1}{2}\right]$

$$V_{H_0}(T) = V_{H_0}\left(\sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha}\right)$$
$$= V_{H_0}\left(\sum_{\alpha=1}^{n} a_\alpha\right)$$
$$= \sum_{\alpha=1}^{n} V(a_\alpha) + \sum\sum_{\alpha \neq \beta} \operatorname{cov}(a_\alpha, a_\beta)$$
$$= \frac{n}{4} \quad [\text{Using (i) and (ii)}]$$

Now,

$V(a_\alpha)$

$= E[a_\alpha^2] - [E(a_\alpha)]^2$

$= 1.\, P[x_\alpha > \mu_0] - \{1.\, P[x_\alpha > \mu_0]\}^2$

$= \tfrac{1}{2} - \tfrac{1}{4} = \tfrac{1}{4}$ ... (i)

$\operatorname{cov}(a_\alpha, a_\beta)$

$= E(a_\alpha a_\beta) - E(a_\alpha)E(a_\beta)$

$= 1.\, P[x_\alpha > \mu_0 \cap x_\beta > \mu_0] - \tfrac{1}{4}$

$= P[x_\alpha > \mu_0].\, P[x_\beta > \mu_0] - \tfrac{1}{4}$

$= \tfrac{1}{2} \cdot \tfrac{1}{2} - \tfrac{1}{4}$ [∵ x_β and x_β are ind]

$= 0$ ... (ii)

∴ <Therefore S is a particular case of linear rank statistic>

∴ <Also it is a symmetric linear rank statistic around mean $\frac{n}{2}$>

Critical fu^n of sign test:

Let us propose a test fu^n; $\phi(s)$

$$\left[\phi(s) = \begin{cases} 1 & \text{if } s - n/2 > c \\ \gamma & s - n/2 = c \\ 0 & \text{o.w} \end{cases}\right]$$

c and γ are determined from size $cond^n$.

Remark: For sign test, if zero difference occurs for any $x_\alpha = \mu_0$, ~~for a~~ under cont. $dist^n$ $assump^n$, this does not create any problem as $P[x_\alpha = \mu_0] = 0$.

But in practical, ~~for~~ zero difference can be avoided by ignoring ~~and~~ them and reducing the sample size simultaneously.

Result:

Sign test is a UMP test.

proof: Propose a test fu^n as,

$$\phi(x) = \begin{cases} 1 & \text{if } \prod_{i=1}^{n} f_1(x_i) > k \prod_{i=1}^{n} f_0(x_i) \\ \gamma & \prod f_1 = k \prod f_0 \\ 0 & \prod f_1 < k \prod f_0 \end{cases}$$

For any c.d.f $F_X(x)$ with p.d.f $f_X(x)$, define

$$f(x) = F(0) f^-(x) + (1 - F(0)) f^+(x) \quad \dots (1)$$

For testing, $H_0: \mu = \mu_0$

against $H_1: \mu > \mu_0$

$F_X(\mu_0) = 1/2$. Let us make a transformation $x_\alpha - \mu$ such that $\mu_0 = 0$ and the corresponding test will be,

$$H_0: \mu' = 0 \equiv F_X(0) = 1/2$$

$$H_1: \mu' > 0 \text{ where, } \langle F_X(0) = 1/2 \rangle$$

$$\equiv F_X(0) < 1/2$$

In (1), $f^-(x) = \begin{cases} \dfrac{f(x)}{F(0)}, & \text{if } x \le 0 \\ 0, & \text{if } x > 0 \end{cases}$

and, $f^+(x) = \begin{cases} 0 & \text{if } x < 0 \\ \dfrac{f(x)}{1 - F(0)} & \text{if } x > 0 \end{cases}$

Under H_0, $f_0(x) = \frac{1}{2} f_0^-(x) + \frac{1}{2} f_0^+(x)$, $\langle f_0(x) =$ p.d.f under $H_0 \rangle$
$[F_X(0) = 1/2]$

Under H_1, $f_1(x) = F_1(0) f_1^-(x) + (1 - F_1(0)) f_1^+(x)$.

Let us reframe the test fun. as per f^+, f^-.

$$\phi(x) = \begin{cases} 1 & \text{if } \prod_{\alpha \neq (\alpha_1, \alpha_2, \dots \alpha_s)} f_1^-(x_\alpha) \prod_{\alpha = (\alpha_1, \alpha_2, \dots \alpha_s)} f_1^+(x_\alpha) > k \prod_{\alpha \neq (\alpha_1, \dots, \alpha_s)} f_0^-(x_\alpha) \prod_{\alpha = (\alpha_1, \dots, \alpha_s)} f_0^+(x_\alpha) \\ \gamma & = \\ 0 & < \end{cases}$$

$$\Rightarrow \phi(x) = \begin{cases} 1 & \text{if } \prod_{\alpha \neq (\alpha_1, \dots \alpha_s)} \frac{f_1(x_\alpha)}{F_1(0)} \prod_{\alpha = (\alpha_1, \dots \alpha_s)} \frac{f_1(x_\alpha)}{1 - F_1(0)} > k \prod_{\alpha \neq} \frac{f_0(x_\alpha)}{1/2} \prod_{\alpha =} \frac{f_0(x_\alpha)}{1/2} \\ \gamma & = \\ 0 & < \end{cases}$$

$$\Rightarrow \phi(x) = \begin{cases} 1 & \frac{\prod f_1(x_\alpha)}{[F_1(0)]^{n-s}} \cdot \frac{\prod f_1(x_\alpha)}{(1 - F_1(0))^{s}} > 2^n k \prod_{\alpha \neq} f_0(x_\alpha) \prod_{\alpha} f_0(x_\alpha) \\ \gamma & = \\ 0 & < \end{cases}$$

$$\Rightarrow \phi(x) = \begin{cases} 1 & \frac{\prod_{\alpha=1}^n f_1(x_\alpha)}{\prod_{\alpha=1}^n f_0(x_\alpha)} > 2^n k (F_1(0))^{n-s} (1 - F(0))^s \\ \gamma & = \\ 0 & < \end{cases}$$

$$\Rightarrow \phi(x) = \begin{cases} 1 & \left(\frac{F_1(0)}{1 - F_1(0)}\right)^s > k^* \frac{\prod_{\alpha=1}^n f_0(x_\alpha)}{\prod_{\alpha=1}^n f_1(x_\alpha)} \\ \gamma & = \\ 0 & < \end{cases}$$

$\therefore$ For fixed $x_1, x_2, \dots, x_n$, $\prod \frac{f_0}{f_1}$ is a constant.

$$\phi(x) = \begin{cases} 1 & \left(\frac{F_1(0)}{1 - F_1(0)}\right)^s > k^{**} \\ \gamma & = k^{**} \\ 0 & < k^{**} \end{cases}$$

$\therefore \phi(x)$ can be written in terms of S = no. of positive obs^{n}s.

Hence,

$$\phi(s) = \begin{cases} 1 & S > k' \\ \gamma & S = k' \\ 0 & S < k' \end{cases}$$

As the above satisfies N-P test construct^n it is a UMP test.

HW (Practical) < Problems on Non-par. Inference)

Suppose that each of 13 randomly choosen female registered voters was asked to indicate if she is going to vote for candidate A or candidate in the upcoming election. The result shows that 9 of the subjects prefers A. Is this sufficient evidence to conclude that candidate A is prefered to B by female voters. Draw the power curve taking at least 8 points.

$\Rightarrow$ $S \underset{H_0}{\sim} Bin(13, \frac{1}{2})$

$S = 9$

We test,
$H_0: p = \frac{1}{2}$
ag. $H_1: p > \frac{1}{2}$

k_α	$P(S=k_\alpha)$	$P(S \le k_\alpha)$
~~6~~	0.20947	0.709~~8~~
7	0.1571	0.867
8	0.08728	0.954 → It is the point,
9	0.03491	0.989
10		

Test is constructed as,

$$\phi(s) = \begin{cases} 1 & S > k_\alpha \\ \gamma & S = k_\alpha \\ 0 & S < k_\alpha \end{cases}$$

$$\therefore \gamma = \frac{P(S \le 9) - 0.95}{P(S = 9)}$$ Here, $k_\alpha = 9$

$$= \frac{0.954 - 0.95}{0.08728}$$

$$= 0.044$$

The test is constructed as

$$\phi(s) = \begin{cases} 1 & \cancel{k_\alpha}\ S > 9 \\ 0.044 & \cancel{k_\alpha}\ S = 9 \\ 0 & 0.\omega \end{cases}$$

- **Why sign test is a non-parametric test?**

Here $S \underset{H_0}{\sim} Bin(n, \frac{1}{2})$ but x_i's are cont.

$\therefore$ Dist^n of S does not depend on the parent pop^n.

Wilcoxon Signed-rank Test: (Test of loca^n)

Let $x_1, x_2, \ldots, x_n$ be the r.s. from a cont. c.d.f $F(\cdot)$ with median μ.

We are to test, $H_0: \mu = \mu_0$.

First consider the difference $D_\alpha = X_{(\alpha)} - \mu_0$. Clearly, the differences are distributed symmetrically under H_0.

$$F_D(-c) = Pr(D_\alpha \leq -c) = Pr(D_\alpha > c) = 1 - F_D(c)$$

With the assumption of a cont. pop^n, zero or tied differences can be avoided by dropping them.

Next we order absolute D_α's, i.e. $|D_\alpha|$'s inc.ly (from smallest to the largest) $|D_1|, |D_2|, \ldots, |D_n|$

Then we rank $|D_\alpha|$'s.

The test statistic is T^+ = Sum of ranks for positive obs^n $(D_\alpha > 0)$

T^- = Sum of ranks for negative obs^n $(D_\alpha < 0)$

- $T^+ + T^-$ = Sum of all possible ranks $= 1 + 2 + \ldots + n = \frac{n(n+1)}{2}$

$\left\langle T^+ = \frac{n(n+1)}{2} - T^- \right\rangle \to$ It means T^+ and T^- are lin. related

Tests based on T^+ only, T^- only or $T^+ - T^-$ are all equivalent.

Let us define the rank of $|D_\alpha|$, R_α^+. T^+ is a linear rank statistic.

Redefine, $T^+ = \sum_{\alpha=1}^{n} Z_\alpha R_\alpha^+$ where, $Z_\alpha = \begin{cases} 1 & \text{if } D_\alpha > 0 \equiv X_\alpha > \mu_0 \\ 0 & \text{o.w} \end{cases}$

<Check it is a linear rank statistic>

Similarly, $T^- = \Sigma(1 - Z_\alpha) R_\alpha^+$

$$\therefore T^+ - T^- = \sum_{\alpha=1}^{n} Z_\alpha R_\alpha^+ - \sum_{\alpha=1}^{n} (1 - Z_\alpha) R_\alpha^+$$

$$= \sum_{\alpha=1}^{n} Z_\alpha R_\alpha^+ + \sum_{\alpha=1}^{n} Z_\alpha R_\alpha^+ - \sum_{\alpha=1}^{n} R_\alpha^+$$

$$= 2 \sum_{\alpha=1}^{n} Z_\alpha R_\alpha^+ - \frac{n(n+1)}{2}$$

- **Difference b/w sign test and Wilcoxon rank sum test:**

Sign test considers only the directions, while W-sign test considers not only directions but also the magnitude of the obs^{n}s.

- Under H_0, $Z_1, Z_2, \ldots, Z_n$ are iid r.v with $P(Z_\alpha = 1) = \frac{1}{2}$
because $P(Z_\alpha = 1) = Pr(X_\alpha > \mu_0) = \frac{1}{2}$
(X_α's are independent so Z_α's are also independent)
- $(Z_1, Z_2, \ldots, Z_n)$ are ind. of $(R_1^+, R_2^+, \ldots, R_n^+)$

proof

$$P(Z_\alpha = 1 \cap |D_\alpha| \leq x)$$
arbitrary pt.

$$= P(0 < D_\alpha \leq x) = F_D(x) - F_D(0) = F_D(x) - \tfrac{1}{2} \text{ [under } H_0]$$
dist^n fn^n of D
$$= \frac{1}{2}\left[2F_D(x) - 1\right]$$
$$= Pr(Z_\alpha = 1) \cdot Pr(-x < D_\alpha < x)$$
$$= Pr(Z_\alpha = 1) \cdot P(|D_\alpha| \leq x)$$

$\therefore$ Z_α's and $|D_\alpha|$'s are ind.
Now the R_α^+'s are the ranks of $|D_\alpha|$'s.
$\therefore$ R_α^+'s are the fn^n of $|D_\alpha|$
$\therefore$ Z_α's and R_α^+'s are ind.

- $E(Z_\alpha) = \frac{1}{2}$, $V(Z_\alpha) = \frac{1}{4}$

$$E(T^+) = E\left(\sum_{\alpha=1}^{n} Z_\alpha R_\alpha^+\right)$$
$$= \sum_{\alpha=1}^{n} E(Z_\alpha) R_\alpha^+$$
$$= \frac{1}{2}\sum_{\alpha=1}^{n} R_\alpha^+ = \frac{1}{2}(1+2+\ldots+n) = \frac{n(n+1)}{4}$$

$$V(T^+) = \frac{n(n+1)(2n+1)}{24}$$

- <u>Determination of rejection region by T^+</u>

To determine the rejection region, the prob. dist^n of T^+ has to be determined under H_0.

$$H_0: \mu = \mu_0 \equiv H_0: \underbrace{Pr(X_\alpha \geq \mu_0)}_{\pi} = 0.5$$
$$H_1: \mu < \mu_0 \equiv \pi > 0.5$$
$$H_1: \mu > \mu_0 \equiv \pi < 0.5$$
$$\mu \neq \mu_0 \equiv \pi \neq 0.5$$

The extreme values of T^+ are zero and $\frac{n(n+1)}{2}$
Since T^+ is completely determined by Z_α's the sample space can be considered to be the set of all possible n-tuples $\{Z_1, Z_2, \ldots, Z_n\}$ with components either 0 or 1 forming 2^n ~~possible~~ possibilities (arrangements)
Each of these distinguishable arrangements is equally

likely under H_0. Then the null distⁿ T^+ is,

$$P(T^+ = t) = \frac{u(t)}{2^n}$$, where $u(t)$ is the number of ways to assign + and − sign on the n-integers $(1, 2, \ldots, n)$ such that the sum of the positive ranks of obs's is t.

Every assignment has a conjugate assignment. Interchanging + sign to − sign and − to +.

Ex $n = 3$, x_1, x_2, x_3 $T^+ = 0, 1, 2, 3, 4, 5, 6$

R 1, 2, 3

T^+	Ranks associated to t	prob. value of $u(t)$
0	—	$P(T^+ = 0) = \frac{1}{8}$
1	1 (+)	$P(T^+ = 1) = 1/8$
2	2 (+)	$P(T^+ = 2) = 1/8$
3	3 (+); 1, 2	$P(T^+ = 3) = 2/8$
4	1, 3	1/8
5	2, 3	1/8
6	1, 2, 3	1/8

- Distⁿ of T^+ is symmetric
- Conjugate pair always exists

Ex $n = 4$, x_1, x_2, x_3, x_4

R 1 2 3 4

T^+	Ranks associated to +	Value of $P(u(t))$
10	1, 2, 3, 4	1/26
9	2, 3, 4	1/16
8	1, 3, 4	1/16
7	4, 3 ; 1, 2, 4	2/16
6	4, 2 ; 1, 2, 3	2/16
5	~~2, 3 ; 1, 4~~ 2, 3 ; 1, 4	2/16
4	1, 3 ; 4	2/16
3	1, 2 ; 3	2/16
2	2	1/16
1	1	1/16
0	0	1/16

H.W ~~Using~~ Using any arbitrary ex (n = 4) show that T^+ is symmetric around its mean 5.

H.W The educational testing service reports that the 75th percentile for scores of the GRE is 693 in a certain year. A r.s of 15 freshmen majoring in Stat. report their GRE scores as 690, 750, 680, 700, 660, 710, 720, 730, 650, 670, 740, 730, 607, 750 and 690, are the scores of students majoring in Stat consistent with the 75th percentile value. $\left(\begin{matrix} H_0 = \pi = 3/4 \\ H_1 : \pi \neq 3/4 \end{matrix}\right)$

$\langle \pi = Pr(x > 693) = \frac{75}{100} = 3/4$

1) Mean $= \frac{n(n+1)}{4}$

$\therefore P\left(T^+ > \frac{n(n+1)}{4}\right) \quad \left\langle T^+ + T^- = \frac{n(n+1)}{2} \right\rangle \ldots (i)$

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

$= P\left(T^- < \frac{n(n+1)}{4}\right)$

$= P\left(T^- < \frac{T^+ + T^-}{2}\right)$

$= P(T^- < T^+)$

$= \cancel{P\left(\frac{n(n+1)}{2} - T^+ < T^+\right)}$

$\boxed{\cancel{P\left(T^+, \frac{n(n+1)}{\cdot}\right)}}$

$P\left(T^+ = \frac{n(n+1)}{4} + t\right)$

$\Rightarrow P(T^+ = T^- + 2t)$

$\Rightarrow$

$\therefore T^+ - \frac{n(n+1)}{4} = T^+ - T^-$

$\Rightarrow T^+ - \frac{n(n+1)}{4} = -\frac{n(n+1)}{2}$

$\Rightarrow T^+ = \frac{n(n+1)}{4}$

$\Rightarrow T^+ - \frac{n(n+1)}{2}$

$= \frac{n(n+1)}{2} - T^+$

$P\left(T^+ = \frac{n(n+1)}{4} + t\right)$

$= P\left(T^+ = \frac{T^+ + T^-}{2} + t\right)$

$= P\left(\frac{T^+}{2} = \frac{T^-}{2} + t\right)$

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

$\Rightarrow P(T^+ = T^- + t)$

$P\left(T^+ > \frac{n(n+1)}{4}\right)$

$= P\left(T^+ > \frac{T^+ + T^-}{2}\right)$

$= P(T^+ > T^-)$

$P\left(T^+ > \frac{n(n+1)}{4}\right) \quad P\left(\frac{n(n+1)}{4} - T^- = t\right)$

$\Rightarrow$

$= P\left(2T^+ > \frac{n(n+1)}{2}\right)$

$= P\left(T^+ - \frac{n(n+1)}{4} > \frac{n(n+1)}{4} - T^+\right)$

$\therefore P\left(T^+ - \frac{n(n+1)}{4} > 0\right) = P\left(T^+ - \frac{n(n+1)}{4} > \frac{n(n+1)}{4} - T^+\right)$

2) $H_0 : p = 3/4$ ag. $H_1 : p \neq 3/4$

$$P(X_i < 693) = 3/4$$

$$S_{H_0} \sim Bin(15, 3/4)$$

$x_i - 693$	Sign
-3	-
57	+
-13	-
7	+
-33	-
17	+
27	+
37	+
-43	-
-23	-
47	+
37	+
-33	-
57	+
+3	-

Number of x_i's greater than 693 is 8.

∴ Here $\langle S = 8 \rangle$

Let, $\langle \alpha = 0.1 \rangle$

We have to find $K_{\alpha/2}$ and $K'_{1-\alpha/2}$ such that it holds,

$$\sum_{s=0}^{K_{\alpha/2}} \binom{15}{s} \left(\frac{3}{4}\right)^s \left(\frac{1}{4}\right)^{n-s} \leq 0.05 \quad \text{---(i)}$$

and,

$$\sum_{s=K'_{1-\alpha/2}}^{15} \binom{15}{s} \left(\frac{3}{4}\right)^s \left(\frac{1}{4}\right)^{n-s} \leq 0.05 \quad \text{---(ii)}$$

respectively.

(i) holds for, $K_{\alpha/2} = 7$ and (ii) holds for, $K'_{1-\alpha/2} = 14$

∴ The test construction will be,

$$\phi(s) = \begin{cases} 1 & s \leq 7 \text{ and } s \geq 14 \\ 0 & \text{o.w} \end{cases}$$

∴ H_0 is accepted, i.e scores of students majoring in Stats consistent with the percentile value.

3) Prove that T^+ is symmetric:

In the construcn. of T^+, every assignment has a conjugate assignment with plus and minus sign interchanged. Since we defined,

$$Z_\alpha = \begin{cases} 1 & x_\alpha > \mu_0 \\ 0 & x_\alpha < \mu_0 \end{cases}$$

Conjugate variable of Z_α will be, $(1 - Z_\alpha)$

∴ The value of T_{conj} for those conjugate assignments will be

$T^+_{comj} = \sum_{\alpha=1}^{n} R^+_\alpha (1-Z_\alpha) = \frac{n(n+1)}{2} - \sum_{\alpha=1}^{n} R^+_\alpha Z_\alpha = \frac{n(n+1)}{2} - T^+_{org}$.

Since every assignment occurs with equal prob. $\frac{1}{2^n}$, it implies that T^+ is symmetric around its mean $\frac{n(n+1)}{4}$

$$\Rightarrow T^+_{comj} - \frac{n(n+1)}{4} = \frac{n(n+1)}{4} - T^+_{org}.$$

Result: T^+ and T^- are identically distributed.

$$\Rightarrow P[T^+ \geq c]$$
$$= P\left[T^+ - \frac{n(n+1)}{4} \geq c - \frac{n(n+1)}{4}\right]$$
$$= P\left[\frac{n(n+1)}{4} - T^+ \geq c - \frac{n(n+1)}{4}\right] \quad [\because T^+ \text{ is symmetric}]$$
$$= P\left[\frac{n(n+1)}{2} - T^+ \geq c\right]$$
$$= P[T^- \geq c]$$

X is sym.
$P(X > \mu + c)$
$= P(X \leq \mu - c)$
$\Rightarrow P(X - \mu > c)$
$= P(X - \mu < -c)$
$= P(\mu - X > c)$

T^+ and T^- follow the identical prob. dist^n.

Remark: Since it is more convenient to work with the smaller rank sum, so we use T^+ or T^- accordingly. If t_α is the critical pt. such that $P(T \leq t_\alpha) = \alpha$, the rejection region for different alt. will be as follows. (rejection region will be on the left hand side)

H_1	Interpretation
$H_1: \mu > \mu_0$	under H_1 T^+ will be higher, T^- will be smaller $P(T^- \leq t_\alpha) = \alpha$ critical point we reject H_0 $T^- \leq t_\alpha$
$H_1: \mu < \mu_0$	~~under~~ $P(T^+ \leq t_\alpha) = \alpha$ we reject H_0 if $T^+ \leq t_\alpha$
$H_1: \mu \neq \mu_0$	$T^+ \leq t_{\alpha/2}$ or $T^- \leq t_{\alpha/2}$

Result $n = 3$

T^+	
0	1/8
1	1/8
2	1/8
3	1/4
4	1/8
5	1/8
6	1/8

$H_0: \mu = \mu_0$
$H_1: \mu < \mu_0$
$P(T^+ \leq t_\alpha) = 0.05$

For every choice of n and α the cut off pt. may not be found in Wilcoxon const^n signed rank test.

Therefore (i) choice of α is essential before constructing the test.

(ii) The critical pt. is not found does not imply the test is invalid.

mark: For paired obs^{n}s both sign test and Wilcoxon signed rank test can be applied by constructing the test on the differences, $D_\alpha^* = X_\alpha - Y_\alpha$ as the univariate obs^n.

roblem 1: In a marketing research test 15 adult males were asked to shave one side of their face with a brand A razor and the other side " " " " " " B " and state their preferred razor. 12 male ~~prefered~~ preferred brand A find the p-value for the alt. that the prob. of preferring the brand A is greater than ~~0.05~~ 0.5.

H_0: A and B are equally preferable

$\equiv H_0 : \pi = 1/2$

ag. H_1 : A is more preferable

$\equiv H_1 : \pi > 1/2$

Let S be the ~~Sm~~ Sample statistic is no. of adults who preferred brand A.

$S = 12$

We use sign test where under H_0,

$S \sim Bin(15, 1/2)$

$$p\text{-value} = Pr[S \leq 12 / H_0]$$

H_0: A and B are equally preferable

That means if μ_0 is the median of the pop^n. prefers brand A

$\Rightarrow F_{\mu_0}(X) = 1/2$, under H_0, $H_0 : \mu = \mu_0$

When brand preference of A is more than 50% median should be shifted to the right of μ_0.

Then ~~H~~ alt. hypo. is $H_1 : \mu > \mu_0$

We reject H_0 is $S - n/2 > S_\alpha$

where, $P_{H_0}\left[S - n/2 \geq s_\alpha\right] \leq 0.1$

$$\Rightarrow \sum_{s'_\alpha}^{n} \binom{n}{s} \left(\frac{1}{2}\right)^n \leq 0.1$$

Problem (H.W)

A study of 5 years ago reported that median amount of sleep by American adults is 7.5 hours out of 24 hours. A current sample of 8 adults reported their avg. amount of sleep per 24 hours as. 7.2, 8.3, 5.6, 7.4, 7.8, 5.2, 9.1 and 5.8 hours. Use the most appropriate stat test to determine whether American adults sleeps less today than they did 5 years ago.

$\Rightarrow$

$D(x_\alpha - \mu_0)$	$\lvert D \rvert$	R_α^+
-0.3	0.3	2.5
0.8	0.8	4
-1.9	1.9	7
-0.1	0.1	1
0.3	0.3	2.5
-2.3	2.3	8
1.6	1.6	5
-1.7	1.7	6

$T^+_{cal} = 4 + 2.5 + 5 = 11.5$

$T^- = 18.5$

We reject H_0 if, $\langle T^+ < T_\alpha \rangle$

$\langle 11.5 \rangle$; 2

Accept H_0 that american adults equally today as they did ~~for~~ 5 years ago.

Application of Wilcoxon signed rank test in paired obs^n:

Given the r.s of n-pairs from any specific individual $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ come from a cont. dist^n.

Suppose (X_i, Y_i); $i = 1(1)n$ come from a cont. dist^n having the jt prob. dist^n fu^n $F_{X,Y}(x, y) = P_n(X \leq x \cap Y \leq y)$

Further suppose the differences b/w X and Y be $D = X - Y$. We are to test whether median of X and median of Y differed by μ_0 or not.

In order to test, $H_0: \mu_D = \mu_0$ where μ_D = median$(X - Y)$

We have to form n differences $D_i = X_i - Y_i - \mu_0$. Remember by differencing X and Y, we convert a bivariate random variable to a univariate one. $\langle \mu_D = \mu_X - \mu_Y \rangle$

First we rank absolute value of D_i and thereafter construct the Wilcoxon signed rank test statistic, based on the positive obs or negative obs.

$$T_{\mathcal{D}}^{+} = \sum_{\mathcal{D}_i > \mu_0} R_{\mathcal{D},i}^{+} \quad \langle R_{\mathcal{D},i}^{+} = \text{Rank of } |\mathcal{D}_i| \rangle$$

Rejection criterion remains the same as before.

Practical

1) A large company was ~~distributed~~ disturbed about the number of person-hours lost per month due to plant accidents and instituted an extensive industrial safety program. The data below show the nuber of person-hours lost in a month at each of 8 diff. plant befor and after the safety program was implemented. Has the safety program being effective in reducing time lost from accidents.

Plant:	1	2	3	4	5	6	7	8
(X) Before:	51.2	46.5	24.1	10.2	65.3	92.1	30.3	49.2
(Y) After:	45.8	41.3	15.8	11.1	58.5	70.3	31.6	35.4

⇒ Suppose person hour loss before and after safety program is denoted a bivariate random variable (X, Y)

Assume (X, Y) coming from a cont. dist^n for $F_{X,Y}(x,y)$.

Let us assume that (μ_X, μ_Y) is the bivariate median of $F_{X,Y}(x,y)$.

We are to test, $H_0: \mu_X = \mu_Y$

$H_1: \mu_X > \mu_Y$ --- (1)

Transform $\mathcal{D} = X - Y$. Assume $\mu_{\mathcal{D}}$ be the median of the dist^n of $\mathcal{D}$.

(1) can be rewritten as,

$H_0: \mu_{\mathcal{D}} = 0$

$H_1: \mu_{\mathcal{D}} > 0$

Under alt. hypothesis rank of positive obs^n will be higher, resulting T^+ larger and T^- smaller simultaneously.

No.	1	2	3	4	5	6	7	8
$\|X-Y\|$	5.4	5.2	8.3	$\bar{0.9}$	6.8	21.8	$\bar{1.3}$	13.8
R^+	4	3	6	1	5	8	2	7

$\langle T^- = 3 \rangle$ ∴ We reject H_0:

$\langle T^+ = 33 \rangle$ if $T^- < t_\alpha$ where T^- being

$t_\alpha = 2$ at $\alpha = 0.01$. the tabular value.

$3 > 2$, we fail to reject H_0.

Safety program is not effective in reducing time loss from accident.

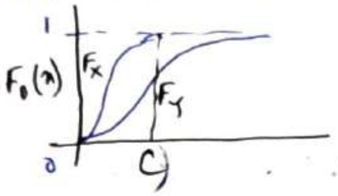
2) Reducing High BP by diet requires reduc^n of Na intake. Listed below are the avg. Na ~~contains~~ contents of 5 ordinary foods in processed form and natural for for equivalent quantities, do you see any difference b/w the median of processed and natural food.

Natural food:	Corn of the cob (2)	Chicken (63)	Ground surloin (60)	Beans (3)	Fresh tuna (40)
Processed food	Canned Corn (251)	Fried Chicken (1220)	All beef biscoti (461)	Canned beans (300)	Canned tuna (409)

Two popⁿ median test

Let $X_1, X_2, \ldots, X_{n_1}$ come from a cont. popⁿ $F_X(x)$.
and $Y_1, Y_2, \ldots, Y_{n_2}$ " " another ind. cont. popⁿ $F_Y(x)$

The r.v Y is called stochastically larger to X if Y takes same ~~prob~~ prob. for higher values while X takes that prob. for lower values.



$$Y >_{st} X \Rightarrow F_Y(c) < F_X(c)$$

Remark Two popⁿ non-parametric location test is based on the idea of equality of two medians (μ_X and μ_Y)

$$\boxed{Y >_{st} X \equiv \mu_X > \mu_Y}$$

We are to test, $H_0: \mu_X = \mu_Y$ $H_1:$ $\mu_X < \mu_Y \equiv Y >_{st} X$
$\mu_X > \mu_Y \equiv Y <_{st} X$
$\mu_X \neq \mu_Y$

Mann-Whitney Test

M-W U test is a special choice of testing the above, where it is assumed that two popⁿs are differed by a locaⁿ parameter θ, i.e

$$F_X(x) = F_Y(x+\theta)$$

$H_0: \mu_X = \mu_Y$ is analogous of writting
eg. $H_1: \mu_Y > \mu_X$
$Y >_{st} X$

$$H_0: \theta = 0$$
$$H_1: \theta > 0$$

For testing the above, we check how ~~many~~ many of Y sample obsⁿs are less than X obsⁿs in the combined sample.

Define, $D_{ij} = \begin{cases} 1 & \text{if } Y_j < X_i \\ 0 & \text{o.w} \end{cases}$ $\begin{cases} i = 1(1)n_1 \\ j = 1(1)n_2 \end{cases}$

$\therefore$ U-statistics is $U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} D_{ij}$

= Number of times Y precedes X

Clearly small value of U reject H_0
~~Therefore~~ Therefore the test based on U will be a left tail test

E(U) and V(U)

Assume $P[Y < X] = \pi$ (say)

$E(D_{ij}) = \pi$

$E(U) = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \pi = n_1 n_2 \pi$

D_{ij}'s are ind variable with D_{jk} as $P[Y_j < X_i \cap Y_k < X_l]$ is just the product of $P[Y_j < X_i] \cdot P[Y_k < X_l]$

But D_{ij}'s are not independent for common subscript.

let, $\left[\begin{array}{l} P[Y_j < X_i \text{ and } Y_k < X_i] \\ = \int_{-\infty}^{\infty} [F_Y(x)]^2 \, dF_X(x) = \pi_1 \text{ (say)} \end{array}\right]$

and, $\left[\begin{array}{l} P[Y_j < X_i \text{ and } Y_j < X_k] \\ = P[X_i \text{ and } X_k > Y_j] \\ = \int_{-\infty}^{\infty} (1 - F_X(x))^2 dF_Y(x) = \pi_2 \text{ (say)} \end{array}\right]$

$\therefore \left[\begin{array}{l} cov(D_{ij}, D_{ik}) = E(D_{ij} \cdot D_{ik}) - E(D_{ij}) \cdot E(D_{ik}) \\ \qquad = \pi_1 - \pi^2 \end{array}\right]$

$\therefore \langle cov(D_{ij}, D_{kj}) = \pi_2 - \pi^2 \rangle \langle cov(D_{ij}, D_{lk}) = 0 \rangle$

$$V(U) = V\left(\sum_i^{n_1} \sum_j^{n_2} D_{ij}\right)$$

$$= \sum_i \sum_j V(D_{ij}) + \sum_{\substack{i \\ i \neq l, j \neq k}}^{n_1} \sum_j^{n_2} \sum_l^{n_1} \sum_k^{n_2} cov(D_{ij}, D_{lk})$$

$$+ \sum_{i=1}^{n_1} \sum \sum_{1 \le j \neq k \le n_2}^{n_2} cov(D_{ij}, D_{ik})$$

$$+ \sum \sum_{1 \le i \neq k \le n_1} \sum_{j=1}^{n_2} cov(D_{ij}, D_{kj})$$

$$= n_1 n_2 \pi(1-\pi) + 0 + n_1(n_2-1)n_2(\pi_1 - \pi^2) + n_2(n_1-1)(\pi_2 - \pi^2)n_1$$

$$= n_1 n_2 [\pi(1-\pi) + (n_2-1)(\pi_1 - \pi^2) + (n_1-1)(\pi_2 - \pi^2)]$$

$$= n_1 n_2 [\pi - \pi^2 - (n_2-1)\pi^2 + (n_2-1)\pi_1 + (n_1-1)\pi_2 - (n_1-1)\pi^2]$$

$$= n_1 n_2 [\pi - \pi^2(1 + (n_2-1) + (n_1-1)) + (n_2-1)\pi_1 + (n_1-1)\pi_2]$$

$$= n_1 n_2 [\pi - (N-1)\pi^2 + (n_2-1)\pi_1 + (n_1-1)\pi_2]$$

$\langle V\left(\frac{U}{n_1 n_2}\right) = 0$ as $n_1 \to \infty, n_2 \to \infty \rangle$ $\quad [N = n_1 + n_2]$

$\langle E\left(\frac{U}{n_1 n_2}\right) = \pi \rangle \quad \therefore \langle \frac{U}{n_1 n_2}$ is a consistent estimator $P[Y < X] = \pi \rangle$

<u>Discrete dist</u>

__Discrete distⁿ of U__ For n_1 X obs and n_2 Y obs. There are $\binom{n_1+n_2}{n_1}$ arrangements of X and Y in combined sample. For every particular arrangement Z, $\exists$ one conjugate arrangement as if Z denotes a set of X and Y written from smallest to largest. Then its conjugate arrangement z' may be proposed from largest to smallest value (conjugate arrangement : how many X follow Y). If U be the value of an arrangement then the prob. dist^n of its conjugate arrangement will be the same, and that value is $U' = \sum_i \sum_j (1-\mathcal{D}_{ij})$

Ex: $n_1 = 4, n_2 = 5$ $\quad \binom{4+5}{5} = 126$
$\quad X \quad\quad Y$

	u	$P(U=u)$
X X Y X Y Y Y Y Y	0	1/126
X X Y X X Y Y Y Y	2	2/126
X X X Y Y X Y Y Y	2	
X X X Y X Y Y Y Y	1	1/126

∴ The pmf of U is $P[U=u] = \dfrac{r_u}{\binom{n_1+n_2}{n}}$

where, r_u is the number of distinguishable arrangements, for which r.v U takes the value u.

HW Find out E(U) and V(U) under H_0

Remark. i) For alt. hypothesis, $H_1 : Y \underset{st}{\geq} X \equiv H_1 : \mu_Y > \mu_X$

we reject H_0 if $\boxed{U < u_\alpha}$

where u_α be the tabular value at α level of significance.

For, $H_1 : Y \underset{st}{\leq} X \equiv H_1 : \mu_Y < \mu_X$ $\quad \left[U' = \sum_i \sum_j (1-\mathcal{D}_{ij})\right]$

we reject H_0 if $\boxed{U' < u_\alpha}$

For, $H_1 : \mu_X \neq \mu_Y$

we reject H_0 if $\boxed{U < u_{\alpha/2}}$ or $\boxed{U' < u_{\alpha/2}}$

Remarks ii) For tied case, $\mathcal{D}_{ij} = \begin{cases} 1 & \text{if } Y_j < X_i \\ 0.5 & Y_j = X_i \\ 0 & Y_j > X_i \end{cases}$

Practical The 2000 census statistics for Alabamma district if the % changes in pop^n b/w 1990 and 2000 for each of the 67 counties. There are 2 types of counties - rural and non rural acc^n to the pop^n size < 25000. Below is the data of 9 rural and 7 non-rural counties on % of pop^n change

~~Row~~ Rural : 1.1, -21.7, -16.3, -11.3, -10.4, -7.0, -2.0, 1.9, 6.2

Non " : -2.4, 9.9, 14.2, 18.4, 20.1, 23.1, 70.4 the null hypo.

Use Mann-Whitney U test for testing equal pop^n change.

⇒ Let the pop^n change of rural county comes from a cont. dist with c.d.f F_Y and ~~whose~~ median μ_Y. In contrast the pop^n change of non " " " " " " " " " " " F_X and median μ_X. We are to test $H_0 : \mu_X = \mu_Y$ - $H_1 : \mu_X \neq \mu_Y$

Arrange the combined sample,

Y Y Y YY X Y Y Y Y X X X X X X

$U = 59, \quad U' = 4$

$\alpha = 0.05 \langle U_{tabular} = 9 \rangle$

$U' = 4 < U_{tab} = 9$

$\therefore$ We reject the null hyp i.e H_0

Under H_0 we know that, $\pi = 1/2$

$$\therefore E(U) = n_1 n_2 / 2$$

Under H_0, $$\pi_1 = \int_{-\infty}^{\infty} [F_X(x)]^2 \, dF_X(x)$$

$$= \frac{[F_X(x)]^3}{3} \Big|_{-\infty}^{\infty}$$

$$= 1/3$$

$$\pi_2 = \int_{-\infty}^{\infty} [1 - F_X(x)]^2 \, dF_X(x)$$

$$= -\frac{[1 - F_X(x)]^3}{3} \Big|_{-\infty}^{\infty}$$

$$= 1/3$$

$\therefore$ Putting the value of π_1 and π_2 in the expression of $V(U)$ we will get that,

$$V(U) = n_1 n_2 \left[1/2 - (N-1) 1/4 + \frac{(n_2-1)}{3} + \frac{(n_2-1)}{3} \right]$$

$$= n_1 n_2 \left[\frac{6 - 3N + 3 + 4n_2 - 4 + 4n_1 - 4}{12} \right]$$

$$= n_1 n_2 \left[\frac{n_1 + n_2 + 1}{12} \right]$$