$$P\left[\, t_{n-1} > c'''' \,\right] = \alpha.$$

$X \sim Bin(n,p)$

$\therefore P(X=x) = {}^nC_x\, p^x (1-p)^{n-x}, \; x = 0,1,2,\dots$

$P(X = \text{odd})$

$= P(X=1) + P(X=3) + \dots$

$= {}^nC_1\, p(1-p)^{n-1} + {}^nC_3\, p^3 (1-p)^{n-3} + \dots$

$= p(1-p)^{n-1}\left[{}^nC_1 + {}^nC_3 (1-p)^2 + {}^nC_5 (1-p)^4 + \dots\right]$

$X \sim Bin(n,p)$, $P(X=x) = {}^nC_x\, p^x (1-p)^{n-x}$

let $q = (1-p)$

$(q+p)^n = {}^nC_0\, q^n p^0 + {}^nC_1\, q^{n-1} p^1 + \dots + {}^nC_n\, q^0 p^n$

$= P(X=0) + P(X=1) + \dots + P(X=n)$
$= \boxed{P(X=\text{even}) + P(X=\text{odd})}$

Similarly,

$(q-p)^n = {}^nC_0\, q^n p^0 - {}^nC_1\, p^1 q^{n-1} + {}^nC_2\, p^2 q^{n-2} \dots$

$= \boxed{P(X=\text{even}) - P(X=\text{odd})}$

$\therefore (q+p)^n - (q-p)^n = 2 \times P(X=\text{odd})$

$\Rightarrow P(X=\text{odd}) = \dfrac{1-(q-p)^n}{2} \quad \left[\because p+q=1\right]$

Non-parametric

1. Gibbons (book).

Rank

Suppose X_α be the α^{th} observation for a set of n obs. $\alpha = 1,2,\dots,n$, all are i.i.d. from a continuous distribution $F_X(x)$.

Rank of X_α : # of observations $\le X_\alpha$

Rank vector is an ordered permutation

$R = (R_1, R_2, \dots, R_n)$

R_1: Rank of 1^{st} obs, R_2: Rank of 2^{nd} obs, R_n: Rank of n^{th} obs

$P(\underset{\sim}{R} = \underset{\sim}{r}) = \dfrac{1}{n!}$ (Joint), $P(R_\alpha = r_\alpha) = \dfrac{1}{n}$, $n = 1,2,\dots$ (Marginal)

Marginal distribution of rank is discrete uniform.

$P(R_\alpha = r_\alpha \cap R_\beta = r_\beta) = \dfrac{1}{n(n-1)} = \dfrac{1}{{}^nP_2}$

$E(R_\alpha) = \dfrac{n+1}{2}$

$\cancel{Cov(R_\alpha, R_\beta)}$ $Var(R_\alpha) = \dfrac{n^2-1}{12}$

$cov(R_\alpha, R_\beta) = -\dfrac{(n+1)}{2}$

$\therefore R_\alpha$ and R_β are not independent

$corr(R_\alpha, R_\beta) = \dfrac{-\frac{(n+1)}{2}}{\frac{(n+1)(n-1)}{12}} = \dfrac{-6}{(n-1)}$

$-1 \le \dfrac{-6}{(n-1)} \le 1$, $n \ge 7$

Linear Rank Statistic

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

Let $\underset{\sim}{R} = (R_1, R_2, \cdots, R_n)$ be the random permutation of $\{1, 2, \cdots, n\}$. The linear Rank Statistic is

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

a_α's are known as <u>regression constants</u> and b_{R_α}'s are <u>scores</u> ~~coefficients~~

* Joint distribution of ranks is independent of the distribution function from which observations are coming.

Therefore the distribution of T is independent of any $F(\cdot)$.

Hence, T can be used to provide distribution free test (non-parametric)

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

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

$$\Rightarrow E(b_{R_\alpha}) = \frac{1}{n} \cdot b_1 + \frac{1}{n} \cdot b_2 + \cdots + \frac{1}{n} b_n$$
$$= \frac{1}{n} \sum_{i=1}^{n} b_i = \bar{b}$$

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

$$V(T) = Var\left(\sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha}\right)$$
$$= \sum_{\alpha=1}^{n} a_\alpha^2 \, var(b_{R_\alpha}) + \sum\sum_{\alpha \neq \beta} a_\alpha a_\beta \, cov(b_{R_\alpha}, b_{R_\beta})$$

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

$\boxed{V(X) = V(X - a)}$

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

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

$Var(b_{R_\alpha}) = Var(b_{R_1})$ [W.L.G fix $\alpha = 1$]

$$Var(b_{R_1}) = E(b_{R_1}^2) - [E(b_{R_1})]^2$$
$$= \frac{1}{n} \sum_{i=1}^{n} b_i^2 - (\bar{b})^2 = \frac{1}{n} \sum_{i=1}^{n} (b_i - \bar{b})^2$$

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

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

$$= E\left[(b_{R_\alpha}-\bar{b})(b_{R_\beta}-\bar{b})\right]$$

$$= \sum_{R_\alpha \neq R_\beta}\sum (b_{R_\alpha}-\bar{b})(b_{R_\beta}-\bar{b})\, P(R_\alpha=\alpha, R_\beta=\beta)$$

$$= \frac{1}{n(n-1)} \sum_{R_\alpha \neq R_\beta}\sum (b_\alpha-\bar{b})(b_\beta-\bar{b})$$

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

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

$$= \frac{1}{n(n-1)}\left[\underbrace{\left\{\sum_{R_\alpha=1}^{n}(b_\alpha-\bar{b})\right\}^2}_{\text{mean deviation about mean}} - \sum_{R_\alpha=1}^{n}(b_\alpha-\bar{b})^2\right]$$

$$= -\frac{Var(b_{R_\alpha})}{(n-1)}$$

$\therefore Var(T)$ irrespective of index, same $\forall \alpha=1(1)n$

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

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

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

$$= Var(b_{R_\alpha})\cdot\left(\frac{n}{n-1}\right)\sum_{\alpha=1}^{n}(a_\alpha-\bar{a})^2$$

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

Asymptotic distribution of linear rank Statistic

✓ Wald Wolforitz condition

✓ Noether condition

✓ Hoeffding condition

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

Inverse permutation

Every permutation(ρ) has its mirror image(ρ')

Such that $\boxed{\rho \underbrace{\circ}_{\text{operation}} \rho' = I} = \{1, 2, \ldots, n\}$

Inverse permutation in Ranking theory is called anti-rank $\underset{\sim}{D} = (D_1\ D_2 \cdots D_n)$

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

Number and Number of the place the obs. occupies is exchanged.

Ex:

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

No. 3	No. 1	No. 4	No. 2
Pl. 1	Pl. 2	Pl. 3	Pl. 4
1	2	3	4
3	1	4	2

$\therefore D = (2\ \ 4\ \ 1\ \ 3)$

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

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

We know ranks are not independent. But the values in inverse permutations are independent.

Moreover joint distribution of any permutation remains the same.

Therefore joint distribution of ranking and anti-ranking are same $= \frac{1}{n!}$

Hence for finding distribution of T, anti-ranking process is more convenient.

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

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

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

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

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

$= \frac{1}{n(n-1)}\left[\frac{3n^2(n+1)^2 - 2n(n+1)(2n+1)}{12}\right]$

$= \frac{n(n+1)}{n(n-1)}\left[\frac{3n(n+1) - 2(2n+1)}{12}\right] - D$

$= \frac{n+1}{n-1}\left[\frac{3n^2 + 3n - 4n - 2}{12}\right] - D$

$= \frac{n+1}{n-1}\left[\frac{3n^2 - 3n + 2n - 2}{12}\right] - D$

$= \frac{n+1}{n-1}\left[\frac{(3n+2)(n-1)}{12}\right] - D$

$cov(R_\alpha, R_\beta) = \frac{(n+1)(3n+2)}{12} - \frac{3(n+1)^2}{12}$

$= \frac{n+1}{12}\left[(3n+2) - 3n - 3\right]$

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

$\therefore corr(R_\alpha, R_\beta) = \boxed{-\frac{1}{n-1}}$

$T = \sum_{\alpha=1}^{3} a_{\alpha} b_{R_\alpha}$ $\quad \underset{\sim}{D}$ = anti-ranks

$T' = \sum_{\alpha=1}^{3} a_{D_\alpha} b_\alpha$

⊛ 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 distribution of T is symmetric about its mean.

Proof ⇒ We know that $(R_1, R_2, \dots, R_n)$

$\overset{D}{\equiv} (n-R_1+1, n-R_2+1, \dots, n-R_n+1)$

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

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

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

$$\overset{D}{\equiv} \sum_{\alpha=1}^{n} a_{D_\alpha} b_{R_{(n-\alpha+1)}}$$

$$\overset{D}{\equiv} \sum_{\alpha=1}^{n} a_{D_\alpha} (k - b_\alpha)$$

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

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

$\therefore T \overset{D}{\equiv} (k n \bar{a} - T)$, it is true for all k.

Choose $k = \bar{b}$

$\therefore T \overset{D}{\equiv} 2n \bar{a} \bar{b} - T$, Hence the proof.

Sign test

Let $X_1, X_2, \dots, X_n$ be a random sample from a continuous distribution $F_X(x)$.
Let μ be the median of the distribution.

We are to test

$H_0: \mu = \mu_0$ (a fixed constant)

$H_1: \mu > \mu_0$

Let S = # of observations greater than μ_0

Let us propose a linear rank statistic for testing the above.

Define, $a_\alpha = \begin{cases} 1 & \text{if } X_\alpha > \mu_0 \\ 0 & \text{o.w.} \end{cases}$

and $b_{R_\alpha} = 1 \quad \forall \alpha = 1(1)n$

$$\therefore T = \sum_{\alpha=1}^{n} a_\alpha b_{R_\alpha} = \sum_{\substack{\alpha=1 \\ X_\alpha > \mu_0}}^{n} a_\alpha = S \text{ (equivalent)}$$

$$\Rightarrow E(T) = \sum_{\alpha=1}^{n} E(a_\alpha)$$

Under H_0

$$= \sum_{\alpha=1}^{n} 1 \cdot P(X_\alpha > \mu_0) = P(X_1 > \mu_0) + P(X_2 > \mu_0) + \dots + P(X_n > \mu_0)$$

$$= \boxed{n/2} \quad (\tfrac{1}{2}, n \text{ times})$$

Under H_0

$$Var(T) = Var_{H_0}\left(\sum_{\alpha=1}^{n} a_\alpha\right)$$

$$= \sum_{\alpha=1}^{n} Var(a_\alpha) + \sum_{\alpha\neq\beta}^{n}\sum^{n} cov(a_\alpha, a_\beta)$$

$$= \sum_{\alpha=1}^{n} E(a_\alpha^2) - [E(a_\alpha)]^2 + \sum_{\alpha\neq\beta}\sum cov(a_\alpha, a_\beta)$$

$$= \sum_{\alpha=1}^{n} E(a_\alpha) - (E(a_\alpha))^2 + \sum_{\alpha\neq\beta}\sum cov(a_\alpha, a_\beta)$$

$\downarrow$ Binary

$$= \frac{n}{4} + \sum_{\alpha\neq\beta}\sum E(a_\alpha a_\beta) - E(a_\alpha)E(a_\beta)$$

$$= \frac{n}{4} + \sum_{\alpha\neq\beta}^{n}\sum^{n} P(X_\alpha > M_0 \cap X_\beta > M_0) - \left(\frac{1}{2}\cdot\frac{1}{2}\right)$$

$$= \frac{n}{4} + \sum_{\alpha\neq\beta}^{n}\sum^{n} P(X_\alpha > M_0)\cdot P(X_\beta > M_0) - \frac{1}{4}$$

Since the sample is random

$$= \frac{n}{4} + \sum_{\alpha\neq\beta}^{n}\sum^{n}\left(\frac{1}{4}-\frac{1}{4}\right) = \boxed{\frac{n}{4}}$$

Therefore S is a particular case of linear rank statistic.

Also, it is a symmetric linear rank statistic around mean $\frac{n}{2}$.

~~Respectively~~ Critical function of sign test

Let us propose a test function

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

c and γ are determined from the size condition.

Remark

For sign test, if zero difference occurs, for any $X_\alpha = M_0$, for continuous distribution assumptions, this does not create any problem as $Pr(X_\alpha = M_0) = 0$.

But in practical, zero difference can be avoided by ignoring and dropping them, simultaneously. Hence, n is minimized.

Result

(sign test is UMP)

Propose a test function

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

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

$$f(x) = F(0)f^-(x) + (1-F(0))f^+(x) \quad \text{---(1)}$$

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

against $H_1: \mu > \mu_0$.

$F_X(\mu_0) = 1/2$

Let us make a transformation $X_\alpha = \mu_0$ such that $\mu_0 = 0$ and the corresponding test will be

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

$H_1: \mu' > 0$ (where $F_X(0) = 1/2$)

$\Downarrow$

$F_X(0) < \frac{1}{2}$

In (1) $f^- = \dfrac{f(x)}{F(0)}$ if $X \le 0$

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
$$\boxed{f_0(x) = \frac{1}{2} f_0^-(x) + \frac{1}{2} f_0^+(x)}$$

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

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

$$\phi(x) = \begin{cases} 1 & \text{if } \prod_{\alpha=1}^{n} f_1(x_i) > k \prod_{\alpha=1}^{n} f_0(x_i) \\ \gamma & \prod_{\alpha=1}^{n} f_1(x_i) = k \prod_{\alpha=1}^{n} f_0(x_i) \\ 0 & \text{o.w.} \end{cases}$$

$\Downarrow$

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

$$\Rightarrow \phi(x) = \begin{cases} 1 & \text{if } \prod_{\alpha \ne (\alpha_1, \alpha_2 \ldots \alpha_s)} \dfrac{f_1(x_\alpha)}{F_1(0)} \prod_{\alpha = (\alpha_1, \ldots, \alpha_s)} \dfrac{f_1(x_\alpha)}{1 - F_1(0)} > k \prod_{\alpha} \dfrac{f_0(x_\alpha)}{1/2} \prod_{\alpha} \dfrac{f_0(x_\alpha)}{1/2} \\ \gamma & \text{"} \quad = \quad \text{"} \\ 0 & \text{"} \quad < \quad \text{"} \end{cases}$$

$$\Rightarrow \phi(x) = \begin{cases} 1 & \dfrac{\prod_{\alpha \ne (\alpha_1, \ldots, \alpha_s)} f_1(x_\alpha) \prod_{\alpha = (\alpha_1, \ldots, \alpha_s)} f(x_\alpha)}{(F_1(0))^{n-s} (1 - F_1(0))^s} > 2^n \cdot k \prod_{\alpha \ne (\alpha_1 - \alpha_s)} f_0(x_\alpha) \prod_{\alpha = (\alpha_1 - \alpha_s)} f_0(x_\alpha) \\ \gamma & \text{"} \quad = \quad \text{"} \\ 0 & \text{"} \quad < \quad \text{"} \end{cases}$$

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

$$\boxed{k^* = \frac{2^n k}{(F_1(0))^n}}$$

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

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

monotonically increasing f^n of s

$\therefore$ $\phi(x)$ can be written in terms of S, # of positive observations.

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

As the above satisfies N-P test construction, it is a UMP test.

Homework (Practical)

(Problems on Non-parametric inferences)

- Suppose that each of 13 randomly chosen female registered voters was asked to indicate if she was going to vote for candidate A or B in an upcoming election. The result shows that 9 of the subjects preferred A. Is this sufficient evidence to conclude that candidate A is preferred to B by female voters?

Draw the power curve taking at least 8 points.

Solution

We have a population of 13 female registered voters.

In an upcoming election, they have ~~an election~~ option to vote for candidate A or B.

Let us consider p as the probability to vote for candidate A.

We test,

$H_0: p = \frac{1}{2}$ vs $H_1: p > \frac{1}{2}$

We have test statistic $S = 9$

As test is constructed as

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

Under H_0, $S_{H_0} \sim Bin(13, \frac{1}{2})$

γ and k_α are to be determined from size condition,

$$E_{H_0}(\phi(S)) = 0.05 \ (\alpha)$$

$$\Rightarrow P_{H_0}(S > k_\alpha) + \gamma P_{H_0}(S = k_\alpha) = 0.05$$

$$\Rightarrow \gamma = \frac{0.05 - P(S > k_\alpha)}{P(S = k_\alpha)}$$

$$= \frac{P_{H_0}(S \leq k_\alpha) - 0.95}{P_{H_0}(S = k_\alpha)}$$

We construct the table as follows

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

k_α	$P(S=k_\alpha)$	$P(S \le k_\alpha)$
0	0.00012	0.00012
1	0.00159	0.00171
2	0.00952	0.01123
3	0.03491	0.04614
4	0.08728	0.13342
5	0.1571	0.29052
6	0.20947	0.49999
7	0.20947	0.70946
8	0.1571	0.86656
9	0.08728	0.95384 → The point to be randomized
10	0.03491	0.98875

We take $k_\alpha = 9$ for randomization and we see

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

$$= \frac{P(S \le 9) - 0.95}{P(S=9)}$$

$$= \frac{0.95384 - 0.95}{0.08728}$$

$$= 0.044$$

∴ The test is constructed as,-

$$\phi(s) = \begin{cases} 1 & s > 9 \\ 0.044 & s = 9 \\ 0 & \text{o.w} \end{cases}$$

Since, $S = 9$, we reject the null hypothesis with rejection probability $\gamma = 0.044$.

Hence, the evidence is not so conclusive.

~~Power function~~

* Why sign test is a non-parametric test?

Sign test is considered as a non-parametric test because

⇒ (H.W) (1) No pre-assumption of specific distribution
(2) Handles ordinal data and deals with median

Wilcoxon sign rank test

Let $X_1, X_2, \ldots, X_n$ be the random sample from a continuous c.d.f $F(\cdot)$ with median μ

We are to test

$H_0: \mu = \mu_0$

Consider the difference $D_\alpha = X_\alpha - \mu_0$

Clearly, the differences are distributed symmetrically, under H_0.

$$F_D(-c) = P(D_\alpha \le -c) = P(D_\alpha > c)$$
$$= 1 - F_D(c)$$

With the assumption of a continuous population, zero or tied difference can be avoided by dropping them.

Next we order absolute D_α's

i.e. $|D_\alpha|$'s increasingly.

$$|D_1|, |D_2|, \ldots, |D_n|$$

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

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

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

~~From the linear model, solving the normal equation we get~~

~~$H = (X^T X)^{-1} X^T X$~~

~~$\therefore H^T =$~~

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 $Z_\alpha = \begin{cases} 1 & \text{if } D_\alpha > 0 \equiv X_\alpha > M_0 \\ 0 & \text{o.w} \end{cases}$ (check if L. Rank Statistic or not)

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

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

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

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

$\boxed{\text{Difference}}$

Sign test considers only the directions while W-sign rank test considers not only directions but also the magnitude of observation

Under H_0, $Z_1, Z_2, \ldots, Z_n$ are iid random variables with $P(Z_\alpha = 1) = \frac{1}{2}$ because $P(Z_\alpha = 1) = P(X_\alpha > M_0) = 1/2$

• $(Z_1, Z_2, \ldots, Z_n)$ are independent of $(R_1^+, R_2^+, \ldots, R_n^+)$

<u>Proof:</u> $P(Z_\alpha = 1 \cap |D_\alpha| \leq x)$

$= P(D_\alpha > 0 \cap -x \leq D_\alpha \leq x)$

$= P(0 < D_\alpha \leq x)$

$= F_D(x) - F_D(0)$

$= F_D(x) - \frac{1}{2}$ $\left[\text{under } H_0 \; P(D_\alpha < 0) = \frac{1}{2}\right]$

$= \frac{1}{2}[2F_D(x) - 1]$

$= P(Z_\alpha = 1) \, P(-x < D_\alpha < x)$

$= P(Z_\alpha = 1) \, P(|D_\alpha| \leq x)$

Z_α's and $|D|$'s are independent

Now, R_α^+'s are the ranks of $|D_\alpha|$

$\therefore$ Therefore R_α^+ are the fn of $|D_\alpha|$. $\Rightarrow$ Any function of Z_α's and $|D_\alpha|$'s are ind.

$\therefore$ Z_α's and R_α^+'s are ind.

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

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

$$= \sum E(Z_\alpha) E(R_\alpha^+)$$

$$= \sum_{\alpha=1}^{n} \frac{1}{2} \cdot \frac{(n+1)}{2} = \frac{n(n+1)}{4}$$

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

• Determination of rejection region by T^+

To determine the rejection region, the probability distribution of T^+ has to be determined under H_0.

$H_0 : M = M_0 \equiv H_0 : P(X_\alpha > M_0) = \frac{1}{2}$

(left) $M < M_0 \equiv \pi > \frac{1}{2}$

(right) $M > M_0 \equiv \pi < \frac{1}{2}$

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. $\left[2^n \text{ possibility}_{(2\times2\times\ldots 2)}\right]$.

Each of this distinguishable arrangements is equally likely under H_0. Then the null distribution T^+ is $P(T^+=t) = \frac{u(t)}{2^n}$

where, $u(t)$: # of ways to assign '+' and '-' sign on the n integers $(1, 2, \ldots, n)$ such that the sum of ranks of positive obs. is t.

Every assignment has a conjugate assignment interchanging + sign to − sign (vice-versa).

Ex: $n=3$

T^+	Ranks associated to t		Value of $P(T^+=t)$
0	−	− − −	$P(T^+=0)=1/8$
1	1 (+)	+ − −	$P(T^+=1)=1/8$
2	2 (+)	− + −	$P(T^+=2)=1/8$
3	3, 1	1, 2	− − + + + −	$P(T^+=3)=2/8$
4	1, 3	(+ − +)	$P(T^+=4)=1/8$
5	2, 3	− + +	$P(T^+=5)=1/8$
6	1, 2, 3	+ + +	$P(T^+=6)=1/8$

- Distribution of T^+ is symmetric.
- Conjugate pair (proves symmetricity) always exists.

$n=4$

T^+	Ranks associated to t	$P(T^+=t)$
0	− − − −	$\frac{u(t)}{2^n} = 1/16$
1	+ − − −	1/16
2	− + − −	1/16
3	+ + − −, − − + −	2/16
4	+ − + −, − − − +	2/16
5	− + + −, + − − +	2/16
6	+ − − +, + + + −	2/16
7	− − + +, + + − +	2/16
8	+ − + +	1/16
9	− + + +	1/16
10	+ + + +	1/16

H.W: Using any arbitrary point, show that T^+ is symmetric $(n=4)$ around its mean, 5.

The educational testing service (ETS) reports that the 75^{th} percentile for scores of the GRE examinations is 693 in a ~~year~~ certain year. A random sample of 15 freshmen majoring in statistics report their GRE scores as

690, 750, 680, 700, 660, 710, 720, 730, 650, 670, 740, 730, 660, 750, 690

Are the scores of students majoring in statistics consistent with the 75^{th} percentile value?

$$\left[s - np > K_{\alpha/2} \right. \text{ or } \left. s - np < K'_{\alpha/2} \right]$$

Using any arbitrary points show that, T^+ is symmetric about its mean for $n=4$

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

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

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

$= Pr\left(T^+ < \frac{n(n+1)}{4} - c\right)$ [$\because T^+$ and T^- are linearly related and identically distributed]

$\therefore$ T^+ has its distribution symmetric about mean $\forall n$

Practical-2

Educational Testing Agency reports that 75^{th} percentile scores of GRE exams is 693.

Let p be the probability GRE scores lies in the percentile range.

We test

$H_0: p = \frac{3}{4}$ vs $H_1: p \neq \frac{3}{4}$

We have a sample of GRE scores of size 15.

Under H_0, the no. of observations satisfying $P(X_i < 693) = 3/4$, $S_{H_0} \sim Bin\left(15, \frac{3}{4}\right)$

$\therefore P_{H_0}(S = s) = \binom{15}{s}\left(\frac{3}{4}\right)^s \cdot \left(\frac{1}{4}\right)^{15-s}$

Now,

Obs − 693	sign
690 − 693	−
750 − 693	+
680 − 693	−
700 − 693	+
660 − 693	−
710 − 693	+
720 − 693	+
730 − 693	+
650 − 693	−
670 − 693	−
740 − 693	+
730 − 693	+
660 − 693	−
750 − 693	+
690 − 693	−

$\therefore$ the no. of positive differences = the value of test statistic

$$\boxed{S = 8}$$

Fix α = (level of significance) $= 0.1$

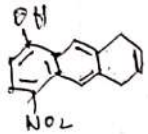
Let $K_{\alpha/2}$ and $K'_{\alpha/2}$ be two constants constructing rejection region

From the size condition,

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

When $k_{\alpha/2} = 7$ we find that

$$P_{H_0}(S \leq 7) \leq 0.05$$



Also,

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

finding $k'_{\alpha/2} = 14$ satisfies the ineanality

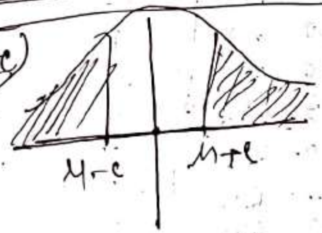
Hence the test is constructed by

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

As we have $s = 8$, we can conclude that, students ~~has~~ are majoring in Statistics has scores consistent with the percentile value (Accepting H_0)

$\cancel{P(X > \mu + c) = P(X < \mu - c)}$

$= P(X - \mu > c)$



Prove that T^{+} is symmetric.

In the construction of T^{+}, every assignment has a conjugate assigrment with plus and minus sign interchange. Since we defined

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

Conjugate variable of Z_α,

$$1 - Z_\alpha = \begin{cases} 1 & \text{if } X_\alpha < \mu_0 \\ 0 & \text{if } X_\alpha > \mu_0 \end{cases}$$

$\therefore$ Therefore the value of T^{+} for those conjugate assigements will be

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

Since every assignment occurs with eanal prob. $\frac{1}{2^n}$ this implies that

$T^{+}_{conj} \to T^{-}$

T^{+} is symmetric.

$\therefore T^{+}$ is symmetric about $\frac{n(n+1)}{4}$.

$$\boxed{T^{+}_{conj} = \frac{n(n+1)}{2} - T^{+}_{org}}$$

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

Scanned with CamScanner

Result

$\left[\begin{array}{l} P(X > \mu + c) \\ = P(X < \mu - c) \end{array}\right]$

T^+ and T^- are identically distributed

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

$$= P\left[\frac{n(n+1)}{4} - T^+ \geq c - \frac{n(n+1)}{4}\right]$$

[$\because T^+$ is symmetric]

$$= P\left[\frac{n(n+1)}{2} - T^+ \geq c\right]$$

$$= P\left[T^- \geq c\right]$$

Result (Rejection region will be on the left-side)

Since it is more convenient to work with the smaller sum, so we use T^+ or T^- accordingly. If t_α is the critical point such that $P(T \leq t_\alpha) = \alpha$

the rejection region for different alternative will be as follows.

H_1	Interpretation
$H_1: \mu > \mu_0$	T^+ higher, T^- lower, to reject, $P(T^- \leq t_\alpha) = \alpha$ (critical point)
$H_1: \mu < \mu_0$	T^+ lower, T^- higher, to reject $P(T^+ \leq t_\alpha) = \alpha$
$H_1: \mu \neq \mu_0$	We reject $T^+ \leq t_{\alpha/2}$ or $T^- \leq t_{\alpha/2}$

* For every choice of n and α, the cut off point may not be found in Wilcoxon-signed rank test.

Therefore

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

(ii) The critical point is not found does not imply ~~that~~ the test is invalid.

~~Practical~~ ● For paired (bivariate) observations, 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 observation.

Practical (H.W)

In a marketing research test, 15 adult males ~~when~~ were asked to shape one side of their face with a brand A ~~raisor~~ razor and the other side of their face with a brand B razor and state their preferied ~~result~~ razor.

12 men preferred brand A. Find the p-value for the alternative for preferring brand A is greater than 0.5.

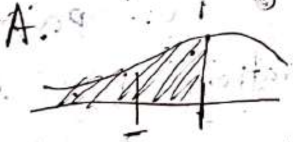
H_0: A and B are equally likely

The preference is more, median shifts.

$H_0: \mu = \mu_0$ (proportion of finding choice of A)

$H_1: \mu > \mu_0$

Let S be sample statistic is no. of adults preferring A.



$S = \underline{12}$.

We use sign test where

$S \underset{H_0}{\sim} Bin(15, \tfrac{1}{2})$

p-value

$P(\text{Reject } H_0 \mid H_0 \text{ is true})$ $\alpha = 0.1$ (say)

$= P(\cancel{S \le 12 \mid \pi = \tfrac{1}{2}})$

$= P(\cancel{S - n\pi \mid \pi = \tfrac{1}{2}})$ → 12 ✓

$= P(S > S_\alpha)$

A study 5 years ago reported that liniant amount of sleep by American adults is 7.5 hours out of 24 hours. A current sample sample of 8 adults reported their average amount of sleep per 24 hrs. as 7.2, 8.3, 5.6, 7.4, 7.8, 5.2, 9.1, 5.8 hours.

Use the most apt test to determine whether American adults sleep less today than 5 years ago. $\alpha = 0.05$

$n = 8$

−0.3 →	0.3	2.5 ✓
$\boxed{0.8}$ →	0.8	$\boxed{4}$
−1.9 →	1.9	7
−0.1 →	0.1	1 ✓
$\boxed{0.3}$ →	0.3	$\boxed{2.5}$ ✓
−2.3 →	2.3	8
$\boxed{1.6}$ →	1.6	$\boxed{5}$
−1.7 →	1.7	6

$T^+ = 11.5$

$T^- = 24.5$

We can't reject

$\frac{8(8+1)}{2} \Rightarrow \boxed{36}$ ✓

$n = 8$, $\alpha = 0.05$

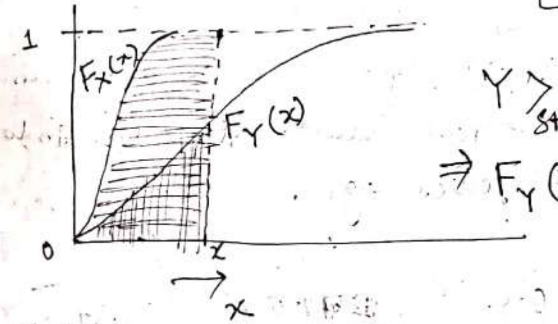
We reject if $T^+ \le t$

Two population Median test

Let $X_1, X_2, \ldots, X_{n_1}$ come from a continuous population with c.d.f $F_X(x)$

And $Y_1, Y_2, \ldots, Y_{n_2}$ come from another independent continuous population with c.d.f $F_Y(x)$.

The r.v Y is called stochastically larger to X if Y takes some probability for higher values while X takes that probability for lower value.



$Y \underset{st}{>} X$

$\Rightarrow F_Y(x) < F_X(x)$

to match the area covered by $F_X(\cdot)$, x has to be increased.

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

$$Y \underset{st}{>} X \Rightarrow \mu_Y > \mu_X \Rightarrow F_Y(x) < F_X(x)$$

We are to test

$H_0: \mu_X = \mu_Y$

$H_1: \mu_X < \mu_Y \equiv Y >_{st} X$

$\mu_X > \mu_Y \equiv Y \underset{st}{<} X$

$\mu_X \neq \mu_Y \quad Y >_{st} X$ or $Y < X$

Mann-Whitney Test

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

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

Therefore, $H_0: \mu_X = \mu_Y$ $\quad H_0: \theta = 0$

against $H_1: \mu_Y > \mu_X \equiv H_1: \theta > 0$

(Analogous to test)

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

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

$\therefore$ U-statistic is

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} D_{ij} \Rightarrow \text{The no. of times } Y \text{ preceeds } X.$$

Clearly small value of U reject H_0. Therefore the test based on U will be a left-tailed test.

$\boxed{E(U),\ Var(U)} \Rightarrow$ Is to be calculated.

Assume $P(Y_j < X_i) = \pi$ (say) $\quad i = 1(1)n_1$, $j = 1(1)n_2$

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

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

$Var(D_{ij})$

$= E(D_{ij})[1 - E(D_{ij})]$

$= \pi(1-\pi)$

D_{ij}'s are independent variable with D_{lk} as $Y_j < X_i$ and $Y_k < X_l$ are two independent events [paired obs. are different].

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

Let $P(Y_j < X_i$ and $Y_k < X_i)$

$= P(\max(Y_j, Y_k) < X_i)$

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

Also $P(Y_j < X_i$ and $Y_j < X_k)$

$= P(\min(X_i, X_k) > Y_j)$

$= \int^{\infty} (1 - F_X(x))^2 dF_Y(x) = \pi_2$

$cov(D_{ij}, D_{ik})$

$= E(D_{ij} D_{ik}) - E(D_{ij})E(D_{ik})$

$= \pi_1 - \pi^2$

Similarly, $cov(D_{ij}, D_{kj})$

$= \pi_2 - \pi^2$

$\therefore Var(U)$ $\quad$ k is a dummy notation there.

$= Var\left(\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} D_{ij}\right)$

$= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} Var(D_{ij}) + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{h}\sum_{k} cov(D_{ij}, D_{hk})$ $\quad h \neq k$, $(h,k) \neq (i,j)$ not together

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

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

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

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

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

~~$= n_1 n_2 (\pi + n_2\pi_1 + n_1\pi_2$~~

~~$= n_1 n_2 [\pi - \pi^2 + n_2\pi_1 - n_2\pi^2 - \pi_1 + \pi^2 + n_1\pi_2 - n_1\pi^2 - \pi_2 + \pi^2]$~~

$\therefore Var(U)$

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

where $N = n_1+n_2$

$\therefore Var\left(\frac{U}{n_1 n_2}\right) \longrightarrow 0$ as $n_1 \to \infty$, $n_2 \to \infty$

$\therefore E\left(\frac{U}{n_1 n_2}\right) = \pi$

$\therefore \frac{U}{n_1 n_2}$ is a consistent estimator of π.
$= P(Y_j < X_i)$ $i = 1(1)n_1$, $j = 1(1)n_2$

Show that $\left(1 - \frac{U}{n_1 n_2}\right)$ is consistent for $(1-\pi)$

Discrete Distribution of U.

For n_1 X obs. and n_2 Y obs. there are $\binom{n_1+n_2}{n_1}$ arrangements by X and Y in combined sample. For every particular arrangement Z.

There exist one conjugate arrangement as if Z denotes a set of X and Y written from smallest to largest, then its conjugate arrangement may be proposed from largest to smallest. (conjugate arrangement: how many X follow Y).

If U be the value of an arrangement then the prob. dist. of its conjugate arrangement will be the same and that value is.

$$U' = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} (1 - D_{ij})$$

Ex. $n_1 = 4$, $n_2 = 5$

$\# \binom{4+5}{5} = 126$

The pm.f of U is

$$P(U=u) = \frac{r_u}{\binom{n_1+n_2}{n_1}}$$

r_u: No of distinguishable arrangements for which r.v U takes the value u.

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

For $H_1: Y \geq_{st} X$ or $\mu_Y > \mu_X$

we reject H_0 is $U < U_\alpha$ (tab. value at α^{th} level of significance)

On the basis of discrete distribution and n_1, n_2 the table is available.

For, $H_1: Y \leq_{st} X$ or $\mu_Y < \mu_X$ we reject H_0 if $U' < U_\alpha$

Construct conjugate arrangement and statistic U'

Scanned with CamScanner

⊛ For tied case

$$D_{ij} = \begin{cases} 1 & Y_j < X_i \\ 0.5 & Y_j = X_i \\ 0 & \text{o.w} \end{cases}$$

We might have fraction part in U for this.

⊛ The 2000 census statistics for Alabama district give the % changes in population between 1990 and 2000 for each 67 countys. There are two types of countys — rural and non-rural, according to the population size less than 25000. Below is the data 9 rural and 7 non-rural countyss:

Rural: 1.1, −21.7, −16.3, −11.3, −10.4, −7, −2.0, 1.9, 6.2

Non-rural: −2.4, 9.9, 14.2, 18.4, 20.1, 23.1, 70.4

Use Mann-Whitney U test to test equal population change.

→ Let the population change of rural county come from a continuous distribution with c.d.f F_Y and median M_Y.

In contrast, non-rural come from F_X and median M_X

$H_0: M_X = M_Y$

$H_1: M_X \neq M_Y$

i) Arranged combined sample

−20.7	−16.3	−11.3	−10.4	−7	−2.4	−2.0	1.1	1.9	6.2
Y	Y	Y	Y	Y	X	Y	Y	Y	Y

9.9	14.2	18.4	20.1	23.1	70.4
X	X	X	X	X	X

$U = 5 + 9 + 9 + 9 + 9 + 9 + 9 = 59$

$U' = \cancel{6 + 6 + 6 + 7 + 7 + 7 + 7 + 7}$

$63 - 54 = 4$

(7×9)

$\alpha = 0.02$

$U_{\alpha/2} = 9$

$\boxed{U_{0.01,7,9} = 9}$

$n_1 = 9, \; n_2 = 7$

$U \cancel{<} U_{\alpha/2}$ $\quad U' < U_{\alpha/2}$ ✓

rejection from one side

Pearson's goodness of fit test
(approximately χ^2)

$$\sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2_{n-1}$$

Kolmogorov - Smirnov Goodness of fit test

Empirical cumulative distribution function

Definition : Empirical cdf of a random sample of size n is denoted by $F_n(x)$

where $F_n(x) = \dfrac{\# \le x}{n}$

$$= \frac{\text{no. of observations} \le x}{n}$$

- ecdf is an estimate of $F(x)$ (c.d.f).

- Distribution of $F_n(x)$

Define an indicator function $\delta_j(x) = \begin{cases} 1 & \text{if } x_j \le x \\ 0 & x_j > x \end{cases}$

$$P(\delta_j(x) = 1) = P(X_j \le x) = F_X(x)$$

$\therefore$ $\delta_j(x)$ is a Bernoullian trial with $F_X(x)$

$\therefore nF_n(x) = \sum_j \delta_j(x)$ → sum of n 'ber' trials

$$\therefore E(nF_n(x)) = \sum_j 1 . F_X(x) = n F_X(x)$$

$\therefore nF_n(x) \sim Bin(n, F_X(x))$

$\therefore F_n(x)$ is unbiased for $F_X(x)$.

is ecdf a consistent estimator of $F_X(x)$

(right tailed always for goodness of fit)

K.S test statistic

$H_0 : F_X(x) = F_0(x)$ ← known cdf

$$D = \sup_{-\infty < x < \infty} \left| F_n(x) - F_0(x) \right|$$

- D is distribution free test statistic
  D takes partwise ~~prob~~ probability based on N.
  For n, the cutoff value of D

- for various values n, the cut off value of D is available in table.

$D_{cal} > D_{tab}$, we reject H_0.

Cramer-Von-Mises

$$D = \int_{-\infty} (F_n(x) - F_0(x))^2 \, dF_0(x)$$

Two sample - KS Test

$X : x_1 \ x_2 \ x_3 \ \dots \ x_{n_1}$

$Y : y_1 \ y_2 \ y_3 \ \dots \ y_{n_2}$

are they coming from same dist?

$$D_{n_1, n_2} = \sup \left| F_{n_1}(x) - F_{n_2}(x) \right|$$

Practical

1) 10 students take a test and their scores are as follows (out of hundred)

95, 80, 40, 52, 60, 80, 82, 58, 65, 50

Test the null hypothesis that the e.d.f of the proportion of right answer a student gets on the test is

$$F_0(x) = \begin{cases} x^2(3-2x) & 0 \le x < 1 \\ 1 & x \ge 1 \end{cases}$$

x	P(x)	Ordered P(x)	$F_n(x)$	$F_0(x)$	$\lvert F_n(x) - F_0(x) \rvert$
95	0.95	0.4	0.1	0.352	0.252
80	0.8	0.5	0.2	0.5	0.3
40	0.4	0.52	0.3	0.53	0.23
52	0.52	0.58	0.4	0.619	0.219
60	0.6	0.6	0.5	0.648	0.148
80	0.8	0.65	0.6	0.718	0.118
82	0.82	0.8	0.8	0.896	0.096
58	0.58	0.8	0.8	0.896	0.096
65	0.65	0.82	0.9	0.914	0.014
50	0.50	0.95	1	0.993	0.007

$D_{cal} = \sup \lvert F_n(x) - F_0(x) \rvert$

$= 0.3$

$\alpha = 0.05$

$D_\alpha = 0.409$

$D_{cal} < D_{tab}$

$\Rightarrow$ Failed to reject H_0

$\therefore$ The proportion of right answer is coming from $F_0(x)$

2) A random sample of 12 persons are interviewed to estimate median annual gross income in a certain economically depressed town. Use the most apt. test for the null hypothesis that income data are standard normally distributed.

9800, 10200, 9300, 8700, 15200, 6800, 8600, 9600, 12200, 15500, 11600, 7200.

$$\boxed{\begin{array}{l} H_0 : F(x) = \phi(x) \\ H_1 : F(x) \neq \phi(x) \end{array}}$$

$\rightarrow \dfrac{X - \bar{X}}{s_x}$ $\quad s_x^2 = \dfrac{1}{n-1} \Sigma (x_i - \bar{x})^2$

③ two mutually ind. r.s each of size 8, were generated, one from the standard normal distribution and another from the chi-square dist. with df 18.

The resulting data are as follows

<u>Norm</u>: −1.91 −1.22 −0.96 −0.72 0.14 0.82 1.45 1.86

χ^2: 4.90 7.25 8.04 14.10 18.30 21.21 23.10 28.12

Do you believe they are coming from the same dist?

convert all to standard form

$$\frac{(\chi^2)-18}{\sqrt{36}}$$

Norm: −1.91 −1.22 −0.96 −0.72 0.14 0.82 1.45 1.86

χ^2: −2.183 −1.79 −1.66 −0.65 0.05 0.535 0.85 1.687

We would test whether two sets of obs. are coming from the same dist.

So, $H_0 : F_1(x) = F_2(x)$ F_1 is standard normal

$H_1 : F_1(x) \neq F_2(x)$ F_2 is standard χ^2 dF

First we combine both the samples increasingly

Norm samp (1^{st})	Chi sq sample (2^{nd})	$F_{n_1}(x)$	$F_{n_2}(x)$	$\lvert F_{n_1}(x) - F_{n_2}(x)\rvert$
	−2.183	0	1/8	1/8
−1.91	−1.79	1/8	2/8	1/8
−1.22	−1.66	2/8	3/8	
−0.96	−0.65	3/8	4/8	
−0.72	0.05	4/8	5/8	
0.14	0.535	5/8	6/8	
0.82	0.85	6/8	7/8	
1.45	1.687	7/8	8/8	
1.86		8/8	8/8	0

	$F_{n_1}(x)$	$F_{n_2}(x)$	$\lvert F_1(x) - F_2(x)\rvert$
−2.183 (2^{nd})	0	1/8	1/8
−1.91 (1^{st})	1/8	1/8	0
−1.79 (2^{nd})	1/8	2/8	1/8
−1.66 (2^{nd})	1/8	3/8	$\boxed{1/4}$
−1.22 (1^{st})	2/8	3/8	1/8
−0.96 (1^{st})	3/8	3/8	0
−0.72 (1^{st})	4/8	3/8	1/8
−0.65 (2^{nd})	4/8	4/8	0
0.05 (2^{nd})	4/8	5/8	1/8
0.14 (1^{st})	5/8	5/8	0
0.535 (2^{nd})	5/8	6/8	1/8
0.82 (1^{st})	6/8	6/8	0
0.85 (2^{nd})	6/8	7/8	1/8
1.45 (1^{st})	7/8 7/8	7/8	0
1.687 (2^{nd})	8/8	8/8	1/8
(1^{st})		8/8	0

$\max |F_{n_1}(x) - F_{n_2}(x)| = 1/4$

$n_1 n_2 D = 8 \times 8 \times \frac{1}{4} = \boxed{16}$

From table for

$n_1 = 8$, $n_2 = 8$

if $n_1 n_2 D$ $\longrightarrow$ p-value is increasing

$= 32$ has p-value $= \boxed{0.283}$

if we fix $\alpha = 0.05$,

$\therefore$ for $n_1 n_2 D = 16$

p-value > 0.283

~~0.05~~

$\Rightarrow$ We fail to reject H_0.

From the table $n_1 n_2 D = 32$ yields the p-value as 0.283. Therefore for $n_1 n_2 D_{cal} = 16$ the p-value will be greater than 0.283. So if we fix $\alpha = 0.05$, we fail to reject H_0.

$\therefore$ This two obs. are coming from the same distribution.