Presidency University

Non-Parametric Methods

Paper Code : **STAT05C11**

Group 1

Name	:	Spandan Ghoshal Ritwick Mondal Kalpesh Chatterjee Niranjan Dey
Roll No.	:	STAT01 STAT10 STAT26 STAT29
Reg. No.	:	18214110001 18214110010 18214110026 18214110029

❄ *Lecture Notes:~*

DAY	TOPIC
1	*p-value*
2	*Binomial Test*
3	*Wilcoxon-signed rank Test*
4	*Confidence Interval of Sign Test & Wilcoxon signed rank Test*
5	*Runs Test of Randomness*
6	*Wald Wolfowitz Run Test*
7	*One sample Kolmogorov-Smirnov Test*
8	*Two sample Kolmogorov-Smirnov Test*
9	*Median Test*
10	*Linear Rank Statistic and General Two-Sample Problem*
11	*Mann-Whitney-Wilcoxon Test*
12	*Mann-Whitney Confidence Interval for θ*
13	*Kruskal-Wallis Test*

Day 1 **:**

Topic *: P-value:∼*

Thinks and Thoughts *:*

- *In ordinary level of significance test, our decision after performing the test based on a sample is dichotomous in nature i.e. we can either reject H_0 or failed to reject H_0.*

- *It may happen, we have enough evidence against the Null hypothesis in the sample but we failed to reject H_0 and we can observe that the sample values are far from Hypothesized value.*

- *Also, the choice of a specific value of level of signicance α (maximum Probability of Type-I error) is determined by non-statistical considerations.*

- *So, we can't measure the extremity of the chosen sample in ordinary level of significance test.*

- *The* P-VALUE *was first formally introduced by* KARL PEARSON*, in his Pearson's chi-squared test and the use of the p-value in statistics was popularized by Ronald Fisher.*

- *The p-value associated with a test is the probability that the test statistic T takes the observed value t and more extreme value in the direction given by alternative under H_0.*

- *For a chosen sample, on changing α our decision after performing ordinary level of significance test may change but p-value associated with that test is fixed for a given sample.*

- *Additionally, p-value provides an information on the extremity of the chosen sample.*

Definition *:∼*

P-VALUE IS THE PROBABILITY OF OBTAINING A SAMPLE THAT IS AS EXTREME OR MORE EXTREME (IN THE DIRECTION GIVEN BY ALTERNATIVE H_1) THAN THE ONE WE OBSERVED UNDER THE ASSUMPTION THAT THE NULL HYPOTHESIS (H_0) IS TRUE.

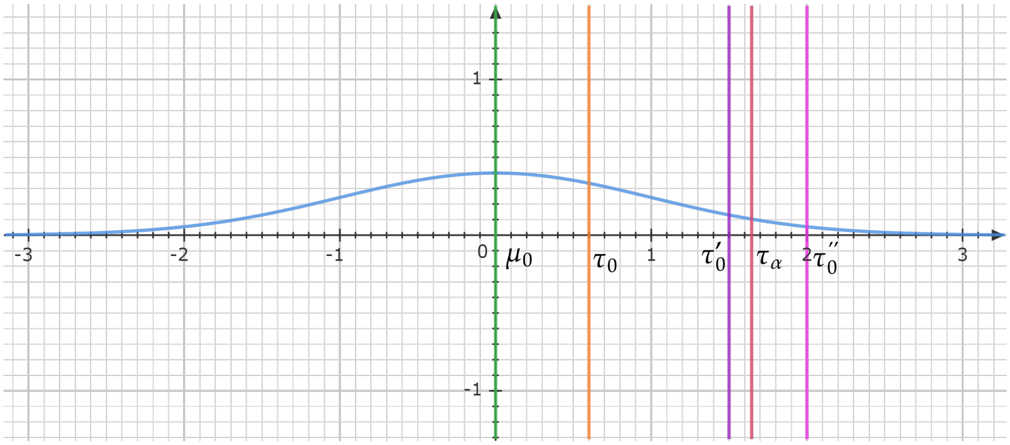
In other words, the p-value associated with a test is the probability that the test statistic T takes the observed value t and more extreme value in the direction given by alternative under H_0.

Examples and Understandings :

For example let $X_i \sim N(\mu, 1),\ i = 1(1)n$

- To test, the Null hypothesis, $H_0 : \mu = \mu_0 = 0$
- Test statistic : $\tau = \sqrt{n}(\bar{x} - \mu_0)$

Right tailed test $(H_1 : \mu > \mu_0)$:

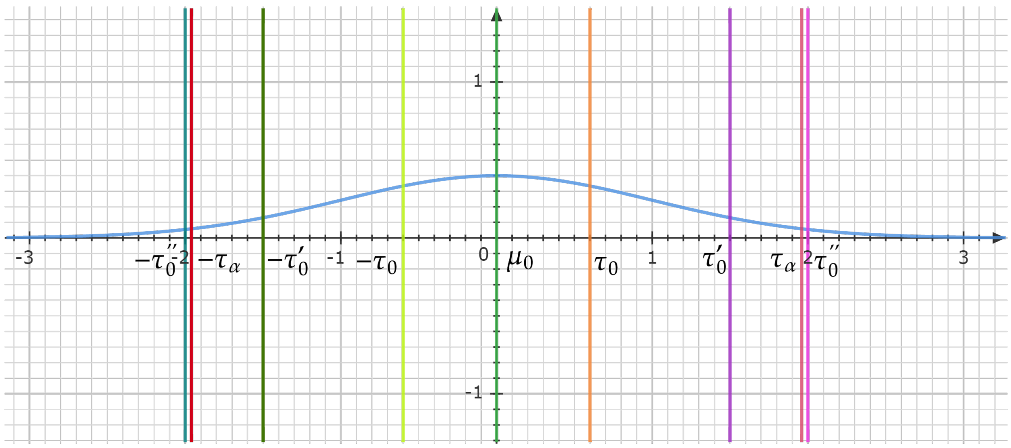


- For right tailed test at 5% level of significance, $\tau_\alpha = \tau_{0.05} = 1.644854$
- Observed values of test statistics :$\tau_0 = 0.6$, $\tau_0^{'} = 1.5$, $\tau_0^{''} = 2$
- We accept Null hypothesis for both $\tau_0 = 0.6$, $\tau_0^{'} = 1.5$.
- We reject Null hypothesis for $\tau_0^{''} = 2$.
- Clearly, the p value of τ_0 is $p = P_{H_0}(\tau > \tau_0) = 0.2742531$ is more than that of $\tau_0^{'}$ is $p^{'} = P_{H_0}(\tau > \tau_0^{'}) = 0.0668072$ is more than that of $\tau_0^{''}$ is $p^{''} = P_{H_0}(\tau > \tau_0^{''}) = 0.02275013 < 0.05($level of significance$)$

Left tailed test $(H_1 : \mu < \mu_0)$:

- For right tailed test at 5% level of significance, $\tau_\alpha = -\tau_{0.05} = -1.644854$
- Observed values of test statistics :$\tau_0 = -0.6$, $\tau_0^{'} = -1.5$, $\tau_0^{''} = -2$
- We accept Null hypothesis for both $\tau_0 = -0.6$, $\tau_0^{'} = -1.5$.
- We reject Null hypothesis for $\tau_0^{''} = -2$.
- Clearly, the p value of τ_0 is $p = P_{H_0}(\tau < \tau_0) = 0.2742531$ is more than that of $\tau_0^{'}$ is $p^{'} = P_{H_0}(\tau < \tau_0^{'}) = 0.0668072$ is more than that of $\tau_0^{''}$ is $p^{''} = P_{H_0}(\tau < \tau_0^{''}) = 0.02275013 < 0.05($level of significance$)$.

Both tailed test** ($H_1 : \mu \neq \mu_0$) **:



- *For both tailed test at* 5% *level of significance,* $\tau_\alpha = \tau_{0.025} = 1.959964$
- *Observed values of test statistics :*$\tau_0 = 0.6$*,* $\tau_0^{'} = 1.5$*,* $\tau_0^{''} = 2$
- *We accept Null hypothesis for both* $\tau_0 = 0.6$*,* $\tau_0^{'} = 1.5$*. We reject Null hypothesis for* $\tau_0^{''} = 2$*.*
- *Clearly, the p value of* τ_0 *is*

  $p = 2 \times min\{P_{H_0}(\tau > \tau_0), P_{H_0}(\tau < \tau_0)\} = 2 \times min\{0.2742531, 0.7257469\} = 0.5485062$

  which is more than that of $\tau_0^{'}$ *that is*

  $p^{'} = 2 \times min\{P_{H_0}(\tau > \tau_0^{'}), P_{H_0}(\tau < \tau_0^{'})\} = 2 \times min\{0.0668072, 0.9331928\} = 0.1336144$

  which is more than that of $\tau_0^{''}$ *that is*

  $p^{''} = 2 \times min\{P_{H_0}(\tau > \tau_0^{''}), P_{H_0}(\tau < \tau_0^{''})\} = 2 \times min\{0.02275013, 0.9772499\} = 0.04550026 < 0.05($*level of significance*$)$

Conclusion :∼

- <u>THE SMALLER THE P-VALUE, THE MORE THE EXTREME OUTCOME (OBSERVED VALUE) AND "STRONGER THE EVIDENCE AGAINST H_0" AND IF P-VALUE$\leq \alpha$ WE REJECT H_0 IN FAVOUR OF H_1</u>.
- *From plots and p-values, it is clear that in all of the above cases,* $\tau_0^{'}$ *has stronger evidence against* H_0 *than* τ_0 *(clearly indicated by their p-values) but we fail to reject* H_0 *in both of the cases.*
- *The additional information of the extremity of the chosen sample is clearly indicated by p-values.*

Day 2 : ∼

Topic : Binomial Test & Applications:∼

Thinks and Thoughts :∼

- *The hypothesis testing procedures we have investigated so far depend on some assumption about the underlying distribution of the data.*
- *In this sceanario, the distribution i.e. the C.D.F. of the study variable itself is unknown to us. The C.D.F. is our parameter.*
- *So, we need to focus on Distribution-free methods.*
- *If the PROBABILITY DISTRIBUTION OF A STATISTIC BECOMES INDEPENDENT OF THE PARENT POPULATION DISTRIBUTION, THEN IT IS CALLED DISTRIBUTION-FREE.*
- *The binomial test is used when an experiment has two possible outcomes (i.e. success/failure) and we have an idea about what the probability of success is.*
- A BINOMIAL TEST IS RUN TO SEE IF OBSERVED TEST RESULTS DIFFER FROM WHAT WAS EXPECTED.

Binomial test criterion (exact case) :∼

- *We observe the outcomes of n i.i.d bernoulli trials possible outcome is dichotomous in nature (i.e. success/failure).*
- *We consider our probability of success be denoted by $p\epsilon(0,1)$ (unknown) .*

Assumptions :∼

1. *The outcomes of each trial can be classified as a success or a failure.*
2. *The probability of a success, denoted by 'p', remains constant from trial to trial.*
3. *The n trials are independent.*

Examples and Understandings :∼

For example let $d_i \sim Ber(p),\ i = 1(1)n$.

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

- *To test, the Null hypothesis, $H_0 : p = p_0 = 0.40$.*
- *Test Statistic :* $B = \sum_{i=1}^{n} d_i \sim Bin(n; p)$.

Right tailed test ($H_1 : p > p_0$) :

```
x=0:8
y=1-pbinom(x,8,0.4)
y                       #upper tail probabilities

[1] 0.98320384 0.89362432 0.68460544 0.40591360 0.17367040 0.04980736 0.00851968
[8] 0.00065536 0.00000000

x[min(which(y<=0.05))]

[1] 5
```

- *For right tailed test at* 5% *level of significance,* $b_\alpha = 5$ *where, the size condition is* $P[B > b_\alpha | H_0] \leq 0.05 (= \alpha)$
- *Clearly, the p-value for the observed* b_0 *is* $P_v = P_{H_0}[B > b_0]$.
- *Test Criterion based on p-value, if* $P_v < \alpha$ *we Reject* H_0 *o/w Accept* H_0.

Left tailed test ($H_1 : p < p_0$) :

```
x=0:8
y=pbinom(x,8,0.4)
y                       #lower tail probabilities

[1] 0.01679616 0.10637568 0.31539456 0.59408640 0.82632960 0.95019264 0.99148032
[8] 0.99934464 1.00000000

x[max(which(y<=0.05))+1]

[1] 1
```

- *For right tailed test at* 5% *level of significance,* $b_\alpha = 1$ *where, the size condition is* $P[B < b_\alpha | H_0] \leq 0.05 (= \alpha)$
- *Clearly, the p-value for the observed* b_0 *is* $P_v = P_{H_0}[B < b_0]$.
- *Test Criterion based on p-value, if* $P_v < \alpha$ *we Reject* H_0 *o/w Accept* H_0.

Both tailed test ($H_1 : p \neq p_0$) ***:***

```
x=0:8
y=pbinom(x,8,0.4)
y                      #lower tail probabilities

[1] 0.01679616 0.10637568 0.31539456 0.59408640 0.82632960 0.95019264 0.99148032
[8] 0.99934464 1.00000000

x[max(which(y<=0.025))+1]

[1] 1

x=0:8
y=1-pbinom(x,8,0.4)
y                      #upper tail probabilities

[1] 0.98320384 0.89362432 0.68460544 0.40591360 0.17367040 0.04980736 0.00851968
[8] 0.00065536 0.00000000

x[min(which(y<=0.025))]

[1] 6
```

- ✳ *For both tailed test at* 5% *level of significance,* $b_{\frac{\alpha}{2}} = 6$, $b'_{\frac{\alpha}{2}} = 1$*where, the size condition is* $P[B < b_{\frac{\alpha}{2}}|H_0] \leq 0.025 (= \frac{\alpha}{2})$, $P[B > b'_{\frac{\alpha}{2}}|H_0] \leq 0.025 (= \frac{\alpha}{2})$
- ✳ *Clearly, the p-value for the observed* b_0 *is* $P_v = 2 \times min\{P_{H_0}(B > b_0), P_{H_0}(B < b_0)\}$.
- ✳ *Test Criterion based on p-value, if* $P_v < \alpha$ *we Reject* H_0 *o/w Accept* H_0.

The Asymptotic Distribution of B:$\sim$

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

Application : Single Sample Sign Test for Median :∼

- *The population median, η. So, we should have $P[Observation \leq \eta] = 1/2$.*

For a single sample of size n, to test the Null hypothesis $H_0 : \eta = \eta_0$ for some specified value η_0 we use the Sign Test.. The test statistic S depends on the alternative hypothesis, H_1.

Exact Case :∼

Right tailed test ($H_1 : \eta > \eta_0$) :

- *Test statistic : S = Number of observations greater than η_0.*
- *If H_0 is true, $S \sim Binomial(n; 1/2)$.*
- *Clearly, the p-value for the observed s_0 is $P_v = P_{H_0}[S > s_0]$.*
- *Test Criterion based on p-value, if $P_v < \alpha$ we Reject H_0 o/w Accept H_0.*

Left tailed test ($H_1 : \eta < \eta_0$) :

- *Test statistic : S = Number of observations less than η_0.*
- *If H_0 is true, $S \sim Binomial(n; 1/2)$.*
- *Clearly, the p-value for the observed s_0 is $P_v = P_{H_0}[S < s_0]$.*
- *Test Criterion based on p-value, if $P_v < \alpha$ we Reject H_0 o/w Accept H_0.*

Both tailed test ($H_1 : \eta \neq \eta_0$) :

- *Test statistic : $S = max\{S_1, S_2\}$ where,S_1 = Number of observations less than η_0 and S_2 = Number of observations greater than η_0.*
- *If H_0 is true, $S \sim Binomial(n; 1/2)$.*
- *Clearly, the p-value for the observed s_0 is $P_v = 2 \times min\{P_{H_0}[S > s_0], P_{H_0}[S < s_0]\}$.*
- *Test Criterion based on p-value, if $P_v < \alpha$ we Reject H_0 o/w Accept H_0.*

Large Sample Approximation :∼

- *If n is large (say n >30), and $S \sim Binomial(n; \frac{1}{2})$, then it can be shown that,*

$$S \overset{a}{\sim} N(np, np(1-p))$$

- *Thus for the sign test, where $p = 1/2$, we can use the test statistic,*

$$Z = \frac{S - \frac{n}{2}}{\sqrt{n \times \frac{1}{2} \times \frac{1}{2}}} = \frac{2S - n}{\sqrt{n}}$$

and note that if H_0 is true,

$$Z \overset{a}{\sim} N(0, 1)$$

✳ The test at α = 0:05 uses the following critical values if,

$$\begin{aligned} H_1 : \eta > \eta_0, then \quad & C_r = 1.644854 \\ H_1 : \eta < \eta_0, then \quad & C_r = -1.644854 \\ H_1 : \eta \neq \eta_0, then \quad & C_r = \pm 1.959964 \end{aligned}$$

✳ For the large sample approximation, it is common to make a continuity correction, where we replace S by $S - \frac{1}{2}$ in the definition of Z.

$$Z = \frac{2S - (n + 1)}{\sqrt{n}}$$

✳ ***Using R we can construct a function of sign test of median for large sample:***

```r
median.test <- function(data,med = 0,alt,alpha = 0.05)
{
  data <- data -  med
  n <- length(data)
  if(alt == "less")
  {
    s <- sum(data < 0)
    tau = (2*s - (n+1))/sqrt(n)
    if(tau < qnorm(alpha,lower.tail = T))
      return(list(tau,qnorm(alpha,lower.tail = T),"H1 is true"))
    else
      return(list(tau,qnorm(alpha,lower.tail = T),"H0 is true"))
  }
  else if(alt == "greater")
  {
    s <- sum(data > 0)
    tau = (2*s - (n+1))/sqrt(n)
    if(tau > qnorm(alpha,lower.tail = F))
      return(list(tau,qnorm(alpha,lower.tail = F),"H1 is true"))
    else
      return(list(tau,qnorm(alpha,lower.tail = F),"H0 is true"))
  }
  else
  {
    s1 <- sum(data < 0)
    s2 <- sum(data > 0)
    s <- max(s1,s2)
    tau = (2*s - (n+1))/sqrt(n)
    if(abs(tau) < qnorm(alpha/2,lower.tail = F))
      return(list(tau,qnorm(alpha/2,lower.tail = F),"H0 is true"))
    else
      return(list(tau,qnorm(alpha/2,lower.tail = F),"H1 is true"))
  }
}
```

- *We are generating 100 random data points from Cauchy with location parameter = 0 and 0.5 and performing both sided test.*
- The output is Value of the test statistic, Critical value and Test Decision respectively.

```
data <- rcauchy(100,location = 0)
median.test(data = data,med = 0,alt = "both")

[[1]]
[1] 0.5

[[2]]
[1] 1.959964

[[3]]
[1] "H0 is true"

data <- rcauchy(100,location = 0.5)
median.test(data = data,med = 0,alt = "both")

[[1]]
[1] 3.3

[[2]]
[1] 1.959964

[[3]]
[1] "H1 is true"
```

Comment :

- *The sign test can also test if the median of a collection of numbers is significantly greater than or less than a specified value.*
- The sign test is considered a weaker test, because it tests the pair value below or above the median and it does not measure the pair difference.

Day 3:∼

Topic : Wilcoxon signed rank Test :∼

Thinks & Thoughts :∼

- *Sign test does not make use of the* measure the pair difference *between the observed and the assumed value of the quantile. It tests the pair value below or above the quantile.*

- *Wilcoxon signed rank test provides an alternative test of location by taking into account the magnitude of the difference as well as their sign.*

- *Developed in 1945 by the statistician* Frank Wilcoxon*, the signed rank test was one of the first* "nonparametric" *procedures developed.*

- *The Wilcoxon signed rank test should be used if the differences between pairs of data are* non-normally distributed.

- *The Wilcoxon signed rank test compares your sample median against a hypothetical median. The null hypothesis for this test is that the medians of two samples are equal. One of the samples is from a theoritical distribution and another is our observed sample.*

- *The one-sample t-test is used to test whether the mean of a population is greater than, less than or not equals to a specific value because the t-distribution is used to calculate critical values for the test. The t-test assumes that the population standard deviation is unknown and will be estimated by the data.*

- *The* non parametric analog *of the one-sample or paired t-test is the* wilcoxon signed Rank test.

- Signed Rank test *also can be used for ordered (ranked)* categorical variables *without a numerical scale.*

Assumptions of the Wilcoxon signed-rank test :∼

- *Data are paired and come from the same population and each pair is chosen randomly and independently.*

- *The* DIFFERENCES ARE CONTINUOUS.

- *The* DISTRIBUTION OF EACH DIFFERENCE IS SYMMETRIC.

- *The differences are mutually independent.*

- *The* DIFFERENCES ALL HAVE SAME MEDIAN.

- *The measurement scale is at least interval.*

- *If the* DATA IS DISCRETE WITH AT LEAST FIVE UNIQUE VALUES, WE CAN IGNORE THE CONTINUOUS VARIABLE ASSUMPTION.

- *The greatest restriction is that our data come from a random sample of the population. If we don't have a random sample, our significnce levels will probably be incorrect.*

Test Procedure of one sample (exact test):~

Let $Y_i = X_i - \theta$,i=1(1)m where we have $2n$ observations,two observations ($x_i y_i$) on each of m subjects.

- *First we have find the absolute values $|Y_1|, |Y_2|, \ldots\ldots, |Y_m|$.*
- *Exclude pairs with $|Y_i| = 0$. Let $n(\leq m)$ be the reduced sample size.*
- *Arrange them in the ascending order.*
- *We define R_i= rank of $|Y_i|$*
- *Define ,*

$$d_i = \begin{cases} 1 & Y_i > 0 \\ 0 & Y_i < 0 \end{cases}$$

- *The wilcoxon signed rank statistic T^+ is the sum of the positive signed rank , Denoted by*

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

and T^- is the sum of negative ranks. We also note that $T^+ + T^- = \frac{n(n+1)}{2}$

- *Let $r_1 < r_2 < \ldots\ldots. < r_B$ denote the ordered ranks of the absolute values of z_i. So,The null distribution of T^+ can be obtained directly from*

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

- *Under H_0,the distribution of z_i are all symmetric about $\mu = 0$.*

Test Procedure of two sample (exact test):~

Let $z_i = y_i - x_i$,i=1(1)m where we have $2n$ observations,two observations ($x_i y_i$) on each of m subjects.

- *First we have find the absolute values $|z_1|, |z_2|, \ldots\ldots, |z_m|$.*
- *Exclude pairs with $|z_i| = 0$. Let $n(\leq m)$ be the reduced sample size.*
- *Arrange them in the ascending order.*
- *We define R_i= rank of $|z_i|$*

- ❋ *Define ,*

$$d_i = \begin{cases} 1 & z_i > 0 \\ 0 & z_i < 0 \end{cases}$$

- ❋ *The wilcoxon signed rank statistic T^+ is the sum of the positive signed rank , Denoted by*

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

and T^- is the sum of negative ranks. We also note that T^++T^-=$\frac{n(n+1)}{2}$

- ❋ *Let r_1<r_2<.......< r_B denote the ordered ranks of the absolute values of z_i. So,The null distribution of T^+ can be obtained directly from*

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

- ❋ *Under H_0,the distribution of z_i are all symmetric about $\mu = 0$.*

One sided test (Right tail test) :

- ❋ *Here our test $H_0 : \mu = 0$against $H_1 : \mu > 0$*
- ❋ *we shall reject H_0 at α level of significance if T^+≥t_α*
- ❋ *So we need to find a t_α such that $P[T^+ \geq t_\alpha] \leq \alpha$.*

One sided test (Left tail test):

- ❋ *Here our test H_0:μ=0 against $H_1 : \mu < 0$*
- ❋ *we shall reject H_0 at α level of significance if T^+≤$\frac{n(n+1)}{2}$-t_α*
- ❋ *So we need to find a t_α such that $P[T^+ \leq \frac{n(n+1)}{2} - t_\alpha] \leq \alpha$.*

Both sided test:

- ❋ *Here our test H_0:μ=0 against $H_1 : \mu \neq 0$*
- ❋ *we shall reject H_0 at α level of significance if T^+≥$t_{\alpha/2}$or T^+≤$\frac{n(n+1)}{2}$-$t_{\alpha/2}$*
- ❋ *So we need to find a $t_{\alpha/2}$such that $P[T^+ \geq t_{\alpha/2}] + P[T^+ \leq \frac{n(n+1)}{2} - t_{\alpha/2}] \leq \alpha$.*

Large sample test (n>30):∼

❋ *We can also show that under* H_0, $E_{H_0}(T^+) = \sum_{i=1}^{n}(i \times \frac{1}{2} + 0 \times \frac{1}{2}) = \frac{n(n+1)}{4}$ *and* $var_{H_0}(T^+) = \frac{n(n+1)(2n+1)}{24}$

❋ *We can compute the z-value using* $z_{T^+} = \frac{T^+ - E_{H_0}(T^+)}{\sqrt{var_{H_0}(T^+)}} \overset{a}{\sim} N(0,1)$ *asymptotically.*

❋ *The significance of the test statistic is determined by computing the p-value using the standard normal distribution.*

Note:∼

❋ *1. The test procedures based on* T^+ *are called* DISTRIBUTION FREE PROCEDURE.

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

❋ *3.*$T^+ \overset{d}{=} T^-$ *under* H_0.

Day 4 : ∼

Topic : Confidence Interval of Sign Test & Wilcoxon signed rank Test :-

Basic Notions of Confidence Interval :-

❄ *In statistics, a Confidence Interval (CI) is a type of estimate computed from the statistics of the observed data. This proposes a range of plausible value using the sample observations for unknown population parameter with a certain Confidence Level (CL).*

❄ *More strictly speaking,* THE CONFIDENCE INTERVAL REPRESENTS THE PROPORTION OF POSSIBLE CONFIDENCE INTERVALS THAT CONTAIN THE TRUE VALUE OF THE UNKNOWN POPULATION PARAMETER.

❄ *For example, based on random sample observations $(x_1, x_2, ..., x_n)$,an interval $[a(\underset{\sim}{X}), b(\underset{\sim}{X})]$ is called a 95% confidence interval of a parameter θ if on an average (in long run) 95 times out of 100 times ,the value of the interval is capable to contain the true value of the parameter θ, $\theta \in \Omega$.*

Non-parametric Confidence interval for quantile:

Definition:

❄ *First we need to know what is a quantile function (κ_p).*

- A QUATILE FUNCTION OF A DISTRIBUTION IS THAT VALUE OF THE RANDOM VARIABLE SUCH THAT THE PROBABILITY OF THE VARIABLE BEING LESS THAN OR EQUAL TO THAT VALUE EQUALS THE GIVEN PROBABILITY.THIS IS ALSO CALLED THE INVERSE CUMULATIVE DISTRIBUTION FUNCTION.
- *For example, the p^{th} quantile is the value of the random variable X, denoted by κ_p such that $100p\%$ of the value of X is the population less than or equal to it, for a positive fraction p, $(0 < p < 1)$.*

$$P(X \leq \kappa_p) = p \implies F_X(\kappa_p) = p$$

- *If F_X is strictly increasing then p^{th} quantile is the unique solution to the equation $\kappa_p = F_X^{-}(p)$.*

Basic Idea:~

- *Suppose* $F(x)$ *is continuous and a random sample* $(x_1, x_2, ..., x_n)$ *is drawn from it.*κ_p *is the* p^{th} *order quantile.we define* $X_{(r)}$ *and* $X_{(s)}$ *as the rth and sth order statistics,* $r < s$ *.Then* $(X_{(r)}, X_{(s)})$ *is said to be* $100(1{-}\alpha)\%$ *confidence interval for* κ_p *if ,*

$$P[X_{(r)} \leq \kappa_p \leq X_{(s)}] = 1 - \alpha$$

Now, $P[X_{(r)} \leq \kappa_p \leq X_{(s)}]$

$$\begin{aligned}
&= P[\kappa_p \leq X_{(s)}] - P[\kappa_p \leq X_{(r)}] \\
&= P[\kappa_p > X_{(r)}] - P[\kappa_p > X_{(s)}] \\
&= P[\textit{at least r of the observations} < \kappa_p] - P[\textit{at least s of the observations} < \kappa_p] \\
&= \sum_{x=r}^{s-1} \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{s-1} \binom{n}{x} p^x (1-p)^{n-x} - \sum_{x=0}^{r-1} \binom{n}{x} p^x (1-p)^{n-x} \\
&= 1 - I_p(s, n-s+1) - 1 + I_p(r, n-r+1) \\
&= I_p(r, n-r+1) - I_p(s, n-s+1)
\end{aligned}$$

- *From Size Condition,* $P[X_{(r)} \leq \kappa_p \leq X_{(s)}] = 1 - \alpha \implies I_p(r, n-r+1) - I_p(s, n-s+1)$*, we can find the values from Table of Incomplete Beta.*

- *Given* α *and* n *,the selection of r and s satisfying is not unique.We select the pair of r and s for which (s-r) is minimum.*

- *If in this equation,the exact probability (1-*α*) is not attained then we choose that pair of r and s such that,*

$$P[X_{(r)} \leq \kappa_p \leq X_{(s)}] \geq 1 - \alpha \quad \textit{i.e.} \quad I_p(r, n-r+1) - I_p(s, n-s+1) \geq 1 - \alpha$$

Note:

- *For symmetrically placed order statistics* $X_{(r)}$ *and* $X_{(s)}$*,we select the pair of* (r, s) *such that* $r + s = n + 1 \implies s = n - r + 1$

- *We can find* r *from* $\sum_{x=r}^{n-r} \binom{n}{x} p^x (1-p)^{n-x} = 1 - \alpha$ *and hence s=n+1-r.*

Confidence interval for population median:∼

Sign test for median :-

- *We draw a random sample* $(X_{(1)}, X_{(2)}, ..., X_{(n)})$ *of size n from a continous population with population median* θ *which is unknown.*

- *To test,*H_0:θ=θ_0 *vs* H_1:$\theta\neq\theta_0$.*We assume* d_i=$\begin{cases}1 & x_i > \theta_0\\ 0 & x_i < \theta_0\end{cases}$

- *Then* $K = \sum_{i=1}^{n} d_i$ *where each* d_i *is i.i.d bernoulli trial.*

- *Under* H_0*,* $P(X_i > \theta_0) = 0.5$ *as* θ_0 *is the median .*

- *Hence* $K \sim Bin(n, 0.5)$ *under* H_0.

- *So here our test will be* $H_0 : p = 0.5$ *vs* $H_1 : p \neq 0.5$.

- *We reject* H_0 *at* α *level of significance if* $K \geq T_{\alpha_1}$*and* $K \leq T_{\alpha_2}$ *such that* α_1+α_2=α

- $P_{H_0}(K \geq T_{\alpha_1}) \leq \alpha_1$ *where* T_{α_1}*upper* α_1 *percentile of* $Bin(n, 1/2)$

$$i.e. \sum_{i=T_{\alpha_1}}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n \leq \alpha_1 ....(i)$$

- $P_{H_0}(K \leq T_{\alpha_2}) \leq \alpha_2$ *where* T_{α_2}*lower* α_2 *percentile of* $Bin(n, 1/2)$

$$i.e. \sum_{i=0}^{T_{\alpha_2}} \binom{n}{i} \left(\frac{1}{2}\right)^n \leq \alpha_2 ....(ii)$$

Confidence interval :

- Now for sample median,

$$\sum_{i=0}^{r-1} \binom{n}{i} \left(\frac{1}{2}\right)^n \leq \alpha_2 \text{ and } \sum_{i=s}^{n} \binom{n}{i} \left(\frac{1}{2}\right)^n \leq \alpha_1 ....(iii)$$

- Where r and s are such that $100(1-\alpha)\%$ confidence interval for population median is given by $(X_{(r)}$,$X_{(s)})$, $r < s$.

- Comparing (i) and (ii) with (iii) we get $r - 1 = T_{\alpha_2} \implies r = 1 + T_{\alpha_2}$ and $s = T_{\alpha_1}$.

Notes :

- ❄ WHEN THE CONFDENCE LEVEL SHRINKS TO ZERO, THE INTERVAL SHRINKS TO THE MIDDLE VALUE IN SORTED ORDER *(or the two middle values when n is even). This is the sample median, which by convention is defned to be the middle data value in sorted order when n is odd and the average of the two middle data values in sorted order when n is even.*
- ❄ THUS THE SIGN TEST GIVES US A COMPLETE THEORY ABOUT THE MEDIAN.
  - *The sample median is the natural point estimator of the median.*
  - *The confdence interval dual to the sign test is the natural confdence interval.*
  - *The sign test is the natural way to test hypotheses about the median.*
- ❄ *This triple of procedures is a complete competitor to the triple based on the mean (sample mean and t test and confdence interval).*

Wilcoxon-signed rank test **:-**

FOR **Wilcoxon signed rank test**, TWO METHODS OF CONFIDENCE INTERVAL ESTIMATION ARE AVAILABLE HERE.

❄ *Method 1 :*

- *For any sample size N, we can find the number $t_{\alpha/2}$ such that if the true population median is θ and T is calculated for the derived sample values X_i-θ, then*

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

- *The confidence interval technique is to find those two number, say θ_1 and θ_2 where $\theta_1 < \theta_2$ such that when T is calculated for the two sets of differences X_i-θ_1 and X_i-θ_2 ,T^+ or T^-,whichever is smaller is just short of significance,i.e larger than $t_{\alpha/2}$.Then (θ_1, θ_2) will be $100(1-\alpha)\%$ confiidence interval estimate of θ.*

❄ *Method 2 :*

Another method is simply trial-and-error procedure .We simply choose some suitable values of θ and calculate the resulting values of T^+ and T^- .We stop whenever we get numbers slightly larger than $t_{\alpha/2}$.But this process generally does not lead to a unique interval for median.

- ***Fortunately there is a function in R that does it all for us.***
- *Let, we have an unknown random sample of size 20 (here this is taken from* $Cauchy(0,1)$*) and testing that with the theoritical distribution* $Normal(0,1)$.

```
set.seed(50)
x=rcauchy(20)
sort(x)

 [1] -24.1704251  -2.1098420  -2.0196308  -1.6192593  -1.3771752  -1.2998927
 [7]  -0.8981350  -0.5688658   0.1328127   0.1413721   0.2371690   0.2485999
[13]   0.3513205   0.5868634   0.7265660   1.1325964   1.1880481   2.2066584
[19]   2.7929643   5.0405684

y=rnorm(20)
sort(y)

 [1] -1.58987765 -1.16601736 -0.76525139 -0.58689714 -0.49863554 -0.45554055
 [7] -0.36285547 -0.36212039 -0.34992735 -0.32342568 -0.15681338  0.02867454
[13]  0.19573384  0.29520677  0.35653495  0.55475223  0.56358364  0.56874633
[19]  1.68955955  2.66763339

wilcox.test(x,y, paired=T,mu=0, conf.int = TRUE, alternative = "two.sided")


Wilcoxon signed rank test

data:  x and y
V = 106, p-value = 0.9854
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -1.0938689  0.9616472
sample estimates:
(pseudo)median
    0.01076606
```

Comment:

- We have $p-value = 0.9854$ which clearly indicates that, the unknown random sample (here $Cauchy(0,1)$) does not have much evidence against the Null Hypothesis $H_0 : \mu = \mu_0 = 0$. So, we fail to reject H_0 in this case.
  - Point estimate of median that we obtained is $= 0.01076606$ which is very close to $\mu_0 = 0$.
  - 95% confidence interval that we obtained is $(-1.0938689, 0.9616472)$ which is almost symmetric about $\mu_0 = 0$.

Day-5:~

Topic:Runs Test of Randomness:~

Introduction:~

Run test of randomness is a statistical test that is used to know the randomness in data.It is sometimes also called Geary test. RUN TEST OF RANDOMNESS IS AN ALTERNATIVE TEST TO TEST AUTOCORRELATION IN THE DATA*.In the stock market,run test of randomness is applied to know if the stock price of a particular company is behaving randomly or if there is any pattern.*

Run test of randomness is basically based on the run.Now we need to know what is "Run".

Basic idea about Run:~

In any ordered sequence with two types of symbols,a run is defined as a succession of one or more identical symbols,which are followed and preceded by a different symbol or no symbol at all.

For example,the males and the females in a line can have pattern such as M F M F M F M F and M M M M F F F F

which have 8 and 2 runs,respectively.Both the number of runs and their lengths can be used as a measure of the randomness of the ordered symbol sequence.Too few runs,too many runs,a run of excessive length,are very rare in truly random sequences,therefore they can serve as statistical criteria for the rejection of null hypothesis.

RUN TEST OF RANDOMNESS ASSUMES THAT THE MEAN AND VARIANCE ARE CONSTANT AND THE PROBABILIT IS INDEPENDENT.

Procedure for run test for randomness:

To test the run test for randomness,null hypothesis assumes that the distributions of the two continuous population are the same.

Exact null distribution of run:~

Let we consider an ordered sequence of n elements with n_1 elements of the 1st kind and n_2 elements of the 2nd kind. Let there are r_1 runs of 1st kind and r_2 runs of 2nd kind.

We need to derive a test of randomness based on the total number of runs r=r_1+r_2.
There are two possible cases

Case 1:(r is odd)

Let r=2k+1,then there are either k runs of 1st kind and k+1 runs of 2nd kind or k+1 runs of 1st kind and k runs of 2nd kind.

Then total number of arrangements in the first case $\binom{n_1-1}{k-1}$ $\binom{n_2-1}{k}$ *and in the second case* $\binom{n_1-1}{k}$ $\binom{n_2-1}{k-1}$.

There are n_1 *elements of 1st kind and* n_2 *elements of 2nd kind.So total number of possible arrangements is* $\binom{n_1+n_2}{n_1}$.

So,Pr(R=r)= $\frac{\binom{n_1-1}{k-1}\binom{n_2-1}{k}+\binom{n_1-1}{k}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}$.

Case 2:(r is even)

Let r=2k,there are k runs of both type.

Then total number of arrangements in both the case $\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}$

So,Pr(R=r)=$\frac{2\binom{n_1-1}{k-1}\binom{n_2-1}{k-1}}{\binom{n_1+n_2}{n_1}}$.

For a large sample ,the test statistic can be calculated by using an approximation of the normal distribution:

z=$\frac{r-\mu_r}{\sigma_r}$

Where r= number of runs;

μ_r*=expected number of runs;*

σ_r*=s.d of number of runs.*

The values of μ_r *and* σ_r *are computed as follows:*

μ_r=$\frac{2n_1n_2}{n_1+n_2}$*+1 and* σ_r=$\sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$

$z\sim$*N(0,1) in large sample distribution (basically this is applicable for n*$\geq$*30)*

Two-tailed runs Test:∼

At α *level of significance,Critical value=*$z_{1-\alpha/2}$.

For example,at the 5% level of significance a test statistic z *with an absolute value greater* $z_{0.975}$*=1.96 indicates non randomness (i.e. We reject the null hypotheis).*

One-tailed runs test:∼

Here we need to compare the z-score with upper tail critical value and lower tail critical value.

Critical value(upper tail)=$z_{1-\alpha}$.

For lower-tail critical value we just add negative sign on the critical value.

Note:∼

When n<30 we take the data in the given order and mark with 1 the data greater than the median and with 2 the data less than the median.(numbers equal to the median are omitted).

We compare the number of runs with theoretical numbers of runs which we can find in table.

Benefits of a Runs Test:

- *The runs test model is important in determining whether* AN OUTCOME OF A TRIAL IS TRULY RANDOM.
- *A runs test can be a* VALUABLE TOOL FOR INVESTORS WHO EMPLOY TECHNICAL ANALYSIS TO MAKE THEIR TRADING DECISIONS.*These traders analyze statistical trends such as price movement and volume.*
- *This type of testing is helpful in* TIME-SERIES AND QUALITY CONTROL ANALYSIS.

Day 6:

Topic : Wald Wolfowitz Run Test

Basic Idea:

- *In this section, we consider a general two sample test for testing $H_0 : F(z) = G(z)$ for all z against the alternative hypothesis that the two distributions differ in any manner, in location, in spreadth, in skewness, in kurtosis or in any other respect.*
- *Let X_1, X_2, ... , $X_m$and Y_1,Y_2, ... ,Y_n be independent random samples from absolutely continuous distribution functions F(.) and G(.), if it is a simple test of the hypothesis $H_0 : F(z) = G(z)$ for all z, based upon the notion of runs of values of X and of values of Y.*
- *A run is a sequence of letters of the same kind bounded by letters of another kind except for the first and the last position. If we read from left to right, we should say that, we have a run of one value of y, followed by a run of two values of x, followed by a run of one value of y and so on.*
- *We shall now explain what we mean by runs. For example, if $m = 4$, $n = 5$, one might obtain: $yxxyxyyyx$. In our example, there is a total of 6 runs.*
- *Note that, the number of runs is always one more than the number of unlike adjacent symbols.*
- *Suppose that with $m = 7$, $n = 8$, we have the following ordering: $xxxxxyyxxyyyyy$.*
- *For if, $F(u) = G(u)$, we would anticipate a larger number of runs.*

Testing Procedure:-

- *Let us combine the sample of m values of X and the sample of n values of Y, into one $(m+n)$ ordered values arranged in ascending order of magnitude.*
- *Under $H_0 : F(u) = G(u)$, for all $u\epsilon\mathbb{R}$, two samples are taken from the same population and the values of X_is and Y_is will ordinarily be well mixed, the **number of runs will be large**.*
- *The difference between the disribution functions will tend to REDUCE THE NUMBER OF RUNS.*
- *Let R be the number of runs in the combined sample has been ordered. If $R \leq c$, then we reject $H_0 : F(u) = G(u)$ for all $u\epsilon\mathbb{R}$against, $H_1 : F(u) \neq G(u)$. If r_0 is the observed value of R, then the p−value is $P_{H_0}[R \leq r_0]$.*

Null Distribution of R:~

- *To find, $P_{H_0}[R = r]$.*
- *Here our experiment is that assign the values of m x_is and n y_is into $(m+n)$ places i.e. to arrange m alike objects of 1st kind and n alike objects of 2nd kind. This can be done in $\begin{pmatrix} m+n \\ n \end{pmatrix}$ ways.*
- *Under $H_0 : F(u) = G(u)$, the x_is and y_is are i.i.d. random variables and all possible arrangements are equally likely.*
- *To find, $P_{H_0}[R = r]$, it is necessary now to count all arrangements with exactly r runs.*

Case 1:- $(r = 2k+1 = odd)$

There must be $(k+1)$ *runs in* x *values and* k *runs in* y *values or* k *runs in* x *values and* $(k+1)$ *runs in* y *values.*

To get $(k+1)$ *runs in* m X *values we can form* $(k+1)$ *of those runs by inserting* k *dividers into* $(m-1)$ *spaces between* m *values of* X *and this can be done in* $\begin{pmatrix} m-1 \\ k \end{pmatrix}$ *ways. Similarly,* k *runs in* n Y *values can be done in* $\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}$ *ways. Hence,*

$$P_{H_0}[R=r] = \frac{\begin{pmatrix} m-1 \\ k \end{pmatrix}\begin{pmatrix} n-1 \\ k-1 \end{pmatrix} + \begin{pmatrix} m-1 \\ k-1 \end{pmatrix}\begin{pmatrix} n-1 \\ k \end{pmatrix}}{\begin{pmatrix} m+n \\ m \end{pmatrix}}$$

.

Case 2:- $(r = 2k = even)$

There must be k *runs in* x *values and* k *runs in* y *values. Then,*

$$P_{H_0}[R=r] = \frac{2\begin{pmatrix} m-1 \\ k-1 \end{pmatrix}\begin{pmatrix} n-1 \\ k-1 \end{pmatrix}}{\begin{pmatrix} m+n \\ m \end{pmatrix}}$$

as we may begin with either a run in X *values or a run in* Y *values.*

Asymptotic Distribution of R under H_0:-

In the combined ordered sample, we have ordered sequence of two types of symbols like $xyxxyyxyy$*.*

We define,

$$I_k = \begin{cases} 1 & \textit{if the symbols in } k^{th} \textit{and } (k-1)^{th} \\ & \textit{places are different} \\ 0 & otherwise \end{cases}$$

Clearly, $R = 1 + I_2 + ... + I_{m+n}$ *.*

Here, I_k *is a Bernoulli random variable with parameter* $p = \frac{2mn}{(m+n)(m+n-1)}$*,*

so that $E(I_k) = E(I_k^2) = p$*, under* H_0*.*

Hence,

$$\begin{aligned} E_{H_0} &= 1 + (m+n-1)\frac{2mn}{(m+n)(m+n-1)} \\ &= 1 + \frac{2mn}{(m+n)} \\ &= \mu_R, (say). \end{aligned}$$

Also,

$$\begin{aligned} Var(R) &= Var(\sum_{k=2}^{m+n} I_k) \\ &= \sum_{k=2}^{(m+n)} Var(I_k) + \sum \sum_{2 \le j \ne k \le m+n} Cov(I_j, I_k) \\ &= \sum_{k=2}^{(m+n)} E(I_k^2) + \sum \sum_{2 \le j \ne k \le m+n} E(I_j I_k) - (m+n-1)^2 E^2(I_k). \end{aligned}$$

For $j = k-1, k+1$,

$$E(I_j I_k) = 1.1.\frac{m(m-1)n + n(n-1)m}{(m+n)(m+n-1)(m+n-2)}$$

For $j \neq k-1, k+1$,

$$E(I_j I_k) = 1.1.\frac{4m(m-1)n(n-1)}{(m+n)(m+n-1)(m+n-2)(m+n-3)}$$

Hence,

$$Var(R) = \frac{(\mu_R - 1)(\mu_R - 2)}{(m+n-1)} = \sigma_R^2$$

For large m, n *it is seen that,*

$$\frac{R-\mu_R}{\sigma_R} \overset{a}{\sim} N(0,1), \textit{asymptotically under } H_o.$$

Asymptotic distribution is usually good enough for practical purposes when both m *and* n *exceeds* 10.

The Problem of Ties:~

- Ideally, no ties should occur because of the assumption of continuous populations.

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

- *For a conservative test, we can break all ties in all possible ways and compute the total number of runs for each resolution of all ties.*

- *The actual R used as the value of the test statistic is the largest computed value, since that is the one least likely to lead to rejection of* H_0.

- *For each group of ties across samples, where there are* s *number of* x*'s and* t *number of* 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 number of values of R to be computed is the product* $\Pi_{i=1}^{k}\binom{s_i + t_i}{s_i}$

Day 7:∼

Topic:Kolmogorov-Smirnov Test:

Introduction:

The Kolmogorov-Smirnov Test is used to decide if a sample comes from a population with specific distribution. THIS TEST CAN BE MODIFED TO SERVE AS A GOODNESS OF FIT TEST AND IS IDEAL WHEN THE SIZE OF THE SAMPLE IS SMALL.*This test is used in situation where a comparison has to be made between an observed sample distribution($F_0(x)$) and theoretical distribution(e.g.* NORMAL DISTRIBUTION*). Hence, we have the following hypothesis :-*

$$\mathcal{H} : F = F_0$$
$$\mathcal{K} : F \neq F_0 \text{ or } F > F_0 \text{ or } F < F_0$$

One-Sample test:∼

- *Let* $(X_1, X_2, ..., X_n)$ *be a random sample from the theoretical distribution with distribution function* $F(x)$ *(abosolutely continious) which is known to us.*
- *Let our observed distribution function is* $F_0(x)$.
- *Suppose* $X_{(i)}$=i^{th}*order statistics and* $R_i = Rank(X_i)$, $\forall = 1(1)n$
- *Now,we shall define empirical distribution function based on Order Statistics* $(X_{(1)}, X_{(2)}, ..., X_{(n)})$.

$$\hat{F(x)} = \begin{cases} 0 & x < X_{(1)} \\ i/n & X_{(i)} \leq x < X_{(i+1)} \ \forall = 1(1)n \\ 1 & x \geq X_{(n)} \end{cases}$$

This is actually called EMPIRICAL C.D.F .

Now what is Empirical C.D.F.?:

- *Empirical c.d.f we denoted it by* $\hat{F(x)}$ *is the Proportion of sample value less than or equal to x.*
- $\hat{F(x)}$*=(no of values*$\leq x$*)/n*
- *Let us,* $T_n(x) = n.\hat{F(x)}$.
- *For any fixed value* x, $T_n(x) \sim Bin(n, F(x))$.

- ❈ *So,* $E(F\hat{(x)}) = F(x)$ *and* $var(F\hat{(x)}) = \frac{F(x).(1-F(x))}{n}$
- ❈ $F\hat{(x)}$ *converges uniformly to* $F(x)$ *with probability 1.*
- ❈ *Which implies,* $P\{\underset{n\to\infty}{lim}\,\underset{x}{sup}|F\hat{(x)} - F(x)|\} = 0 = 1.$

Kolmogorov-Smirnov test statistic:

Kolmogorov-smirnov test statistic is defined as,

- ❈ $D_n^+ = \underset{x}{sup}\{\hat{F}(x) - F_0(x)\}$
- ❈ $D_n^- = \underset{x}{sup}\{F_0(x) - \hat{F}(x)\}$
- ❈ $D_n = max\{D^+, D^-\}$*.*

For alternative hypothesis, $\mathcal{K} : F > F_0$*, a right tailed test based on* D_n^+ *(or a left tailed test based on* D_n^-*) is appropriate.*

For alternative hypothesis, $\mathcal{K} : F < F_0$*, a left tailed test based on* D_n^+ *(or a right tailed test based on* D_n^-*) is appropriate.*

For alternative hypothesis, $\mathcal{K} : F \neq F_0$*, a both tailed test based on* D_n *is appropriate.*

Testing procedure:

Alternate Hypothesis	*Critical Region*	*p-value*
$\mathcal{K} : F > F_0$	$\{D_n^+ > D_{n,\alpha}\}$	$P_H(D_n^+ > obs(D_n^+))$
$\mathcal{K} : F < F_0$	$\{D_n^+ < D_{n,\alpha}\}$	$P_H(D_n^+ < obs(D_n^+))$
$\mathcal{K} : F \neq F_0$	$\left\{D_n > D_{n,\alpha/2}\right\}$	$P_H(D_n > obs(D_n))$

Where, $D_{n,\alpha}$ *is such that* $Pr\left\{D_n^+ < D_{n,\alpha}|\mathcal{H}\right\} = \alpha$*.*

Large sample KS test:

As $n \to \infty$*, it can be shown that,*

$$\lim_{n\to\infty} Pr\left\{\sqrt{n}D^+ \geq x|\mathcal{H}\right\} = e^{-2x^2}, x \geq 0$$
$$\implies Pr\left\{\sqrt{n}D^+ \geq x|\mathcal{H}\right\} \approx e^{-2x^2}, x \geq 0 \text{ for large } n$$

We reject $\mathcal{H}$ *at level* α *iff* $D_n^+ \geq \frac{D_{n,\alpha}}{\sqrt{n}}$ *and accept* $\mathcal{H}$ *o.w for* $\mathcal{K} : F > F_0$ *where* D_α *is such that* $Pr\left\{D^+ \geq \frac{D_{n,\alpha}}{\sqrt{n}}\right\} = \alpha$*.*

- ❈ *Fortunately there is a function in R that does it all for us.*
- ❈ *Let, we have an unknown random sample of size 20 (here this is taken from* $Laplace(0,1)$ *and testing that with the theoritical distribution* $Normal(0,1)$*.*

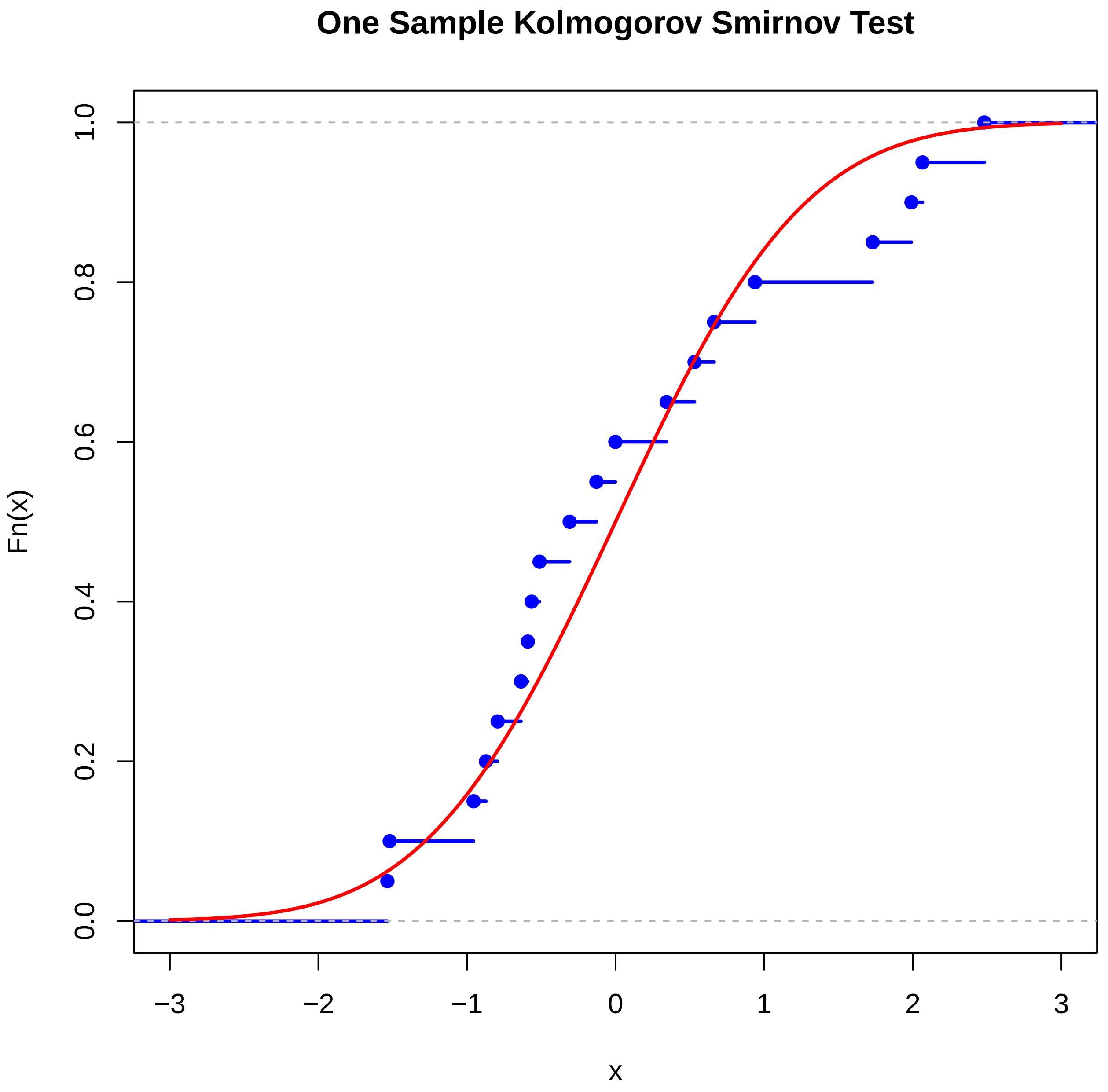
```
set.seed(50)
x1=rexp(20)
x2=rexp(20)
x=x1-x2
sort(x)

 [1] -1.53580931 -1.52024107 -0.95553061 -0.87267757 -0.79373271 -0.63656244
 [7] -0.59076241 -0.56529578 -0.51176530 -0.30905527 -0.12885773 -0.00134054
[13]  0.34370346  0.53146374  0.66287063  0.93854299  1.72989632  1.99096363
[19]  2.06545200  2.48106937

y=rnorm(20)
sort(y)

 [1] -2.11475814 -2.07083822 -1.77015052 -1.54229072 -1.44397435 -1.04392641
 [7] -0.62331406 -0.54093340 -0.07628458  0.21615009  0.37374680  0.40116716
[13]  0.46980188  0.49173751  0.70665103  0.86719825  1.59058620  1.71047138
[19]  1.71446802  2.60427625

ks.test(x, "pnorm", alternative="two.sided")


One-sample Kolmogorov-Smirnov test

data:  x
D = 0.15818, p-value = 0.6424
alternative hypothesis: two-sided
```

Comment:∼

❊ *We have* $p{-}value = 0.6424$ *which clearly indicates that, the unknown random sample (here* $Laplace(0,1)$*) does not have much evidence against the Null Hypothesis* $H_0 : F(Normal) = F_0(Laplace($*given data*$))$*. So, we fail to reject* H_0 *in this case.*

Plot:

```r
plot(ecdf(x),col = "blue",lwd = 2, xlim=c(-3, 3),
main="One Sample Kolmogorov Smirnov Test")
sp <- seq(-3,3,0.001)
lines(sp,pnorm(sp),col = "red",lwd = 2)
```



Result:

Kolmogorov-smirnov test is exactly distribution free.

Justification:

$$
\begin{aligned}
D_n^+ &= \sup_{-\infty<x<\infty}\left[\widehat{F}_n\left(x\right)-F_0\left(x\right)\right] \\
&= \max\left\{\sup_{-\infty<x\leq x_{(1)}}\left[\widehat{F}_n\left(x\right)-F_0\left(x\right)\right], \sup_{x_{(1)}\leq x<x_{(2)}}\left[\widehat{F}_n\left(x\right)-F_0\left(x\right)\right], \ldots, \sup_{x\geq x_{(n)}}\left[\widehat{F}\left(x\right)-F_0\left(x\right)\right]\right\} \\
&= \max\left\{\left(\frac{1}{n}-0\right), \left(\frac{2}{n}-F_0\left(x_{(1)}\right)\right), \left(\frac{3}{n}-F_0\left(x_{(2)}\right)\right), \ldots, \left(1-F_0\left(x_{(n)}\right)\right)\right\} \\
&= f\left(U_{(1)}, U_{(2)}, \ldots, U_{(n)}\right) \text{ where } U_{(i)}=F_0\left(x_{(i)}\right), i=1,2,\ldots,n
\end{aligned}
$$

Now,$X\sim F_0$ under H.

- ❄ *The distribution of D_n^+depends only one the $U_{(i)}\sim Beta\left(n-i+1,i\right)$ which is free of any specified F_0.*
- ❄ *Hence,Under H the distribution of D_n^+ is independent of F_0.*

Advantages:$\sim$

- ❄ *KS statistic is easy to calculate.*
- ❄ *It can be used as a* GOODNESS OF FIT TEST FOLLOWING REGRESSION ANALYSIS.
- ❄ *There are* NO RESTRICTIONS ON SAMPLE SIZE,SMALL SAMPLES ARE ACCEPTABLE.

Disadvantages:$\sim$

- ❄ *It generally* CAN'T BE USED FOR DISCRETE DISTRIBUTION.
- ❄ Sensitivity is higher at the center of the distribution and lower at the tails.

Day 8:∼

Topic:Two-sample Kolmogorov-Smirnov Test:

Introduction:

The two-sample kolmogorov-smirnov test is used to test whether two samples come from the same distribution.The procedure is very similar to the one kolmogorov-smirnov test.

Detailed discussion:

- Let be $(X_1, X_2, \ldots, X_{n_1})$ a random sample from the distribution with distribution function $F_1(x)$ and $(Y_1, Y_2, \ldots, Y_{n_2})$ be a random sample from the distribution with distribution function $F_2(x)$. Both are independent.
- $F_1(x)$ and $F_2(x)$both are absolutely continuous distribution functions.
- Suppose $X_{(i)} = i^{th}$ order statistics corresponding to X-sample $\forall i = 1, 2, \ldots, n_1$ and $Y_{(i)} = i^{th}$ order statistics corresponding to y-sample $\forall i = 1, 2, \ldots, n_2$.
- Now, we shall define empirical distribution function based on $\left(X_{(1)}, X_{(2)}, \ldots, X_{(n_1)}\right)$ $\left(Y_{(1)}, Y_{(2)}, \ldots, Y_{(n_2)}\right)$.
- $\hat{F_1}(x)$= Empirical distribution function based on $x_i^{'}$

$$\widehat{F}_1(x) = \begin{cases} 0 & x < X_{(1)} \\ \frac{i}{n_1} & X_{(i)} \le x < X_{(i+1)} \ , i = 1(1)n_1 - 1 \\ 1 & x \ge X_{(n_1)} \end{cases}$$

- Similarly,$\widehat{F}_2(x)$ = Empirical distribution function based on $y_i^{'}$

$$\widehat{F}_2(x) = \begin{cases} 0 & x < Y_{(1)} \\ \frac{i}{n_2} & Y_{(i)} \le x < Y_{(i+1)}, i = 1(1)n_2 - 1 \\ 1 & x \ge Y_{(n_2)} \end{cases}$$

This is actually called EMPIRICAL CDF.

Kolmogorov-Smirnov test statistic:∼

Two-sample Kolmogorov-smirnov test statistic is defined as,

- $D^+ = \sup\left\{\widehat{F}_1(x) - \widehat{F}_2(x)\right\}$.
- $D^- = \sup\left\{\widehat{F}_2(x) - \widehat{F}_1(x)\right\}$.
- $D = \max\{D^+, D^-\}$.

Here our null hypothesis $\mathcal{H} : F_1 = F_2$.

For alternative hypothesis, $\mathcal{K} : F_1 > F_2$*,a right tailed test based on* D^+*(or a left tailed test based on* D^-*) is appropriate.*

For alternative hypothesis, $\mathcal{K} : F_1 < F_2$*,a left tailed test based on* D^+*(or a right tailed test based on* D^-*) is appropriate.*

For alternative hypothesis, $\mathcal{K} : F_1 \neq F_2$*,a right tailed test based on* D *is appropriate.*

***Testing procedure*:**

- *For right tailed test,* $\mathcal{H} : F_1 = F_2$ *vs* $\mathcal{K} : F_1 > F_2$ *we reject* $\mathcal{H}$ *at level* α *iff* $D^+ > D_\alpha$ *and accept* $\mathcal{H}$ *ow where* D_α*is such that* $Pr\{D^+ > D_\alpha | \mathcal{H}\} = \alpha$.

- *For left tailed test,* $\mathcal{H} : F_1 = F_2$ *vs* $\mathcal{K} : F_1 < F_2$ *we reject* $\mathcal{H}$ *at level* α *iff* $D^+ < D_\alpha$ *and accept* $\mathcal{H}$ *ow .*

where D_α*is such that* $Pr\{D^+ < D_\alpha | \mathcal{H}\} = \alpha$.

- *For both sided test,* $\mathcal{H} : F_1 = F_2$ *vs* $\mathcal{K} : F_1 \neq F_2$ *we reject* $\mathcal{H}$ *at level* α *iff* $D > D_\alpha$ *and accept* $\mathcal{H}$ *ow .*

where D_α*is such that* $Pr\{D > D_\alpha | \mathcal{H}\} = \alpha$.

- *Fortunately there is a function in R that does it all for us.*

- *Let, we have two unknown random sample of size 20 (here this is taken from*

$Laplace(0,1)$*,*$Normal(0,1)$*)*

- *Now, we want to cheak if both of the unknown samples have same distribution or not.*

```
set.seed(50)
x1=rexp(20)
x2=rexp(20)
x=x1-x2
sort(x)

 [1] -1.53580931 -1.52024107 -0.95553061 -0.87267757 -0.79373271 -0.63656244
 [7] -0.59076241 -0.56529578 -0.51176530 -0.30905527 -0.12885773 -0.00134054
[13]  0.34370346  0.53146374  0.66287063  0.93854299  1.72989632  1.99096363
[19]  2.06545200  2.48106937

y=rnorm(20)
sort(y)

 [1] -2.11475814 -2.07083822 -1.77015052 -1.54229072 -1.44397435 -1.04392641
 [7] -0.62331406 -0.54093340 -0.07628458  0.21615009  0.37374680  0.40116716
[13]  0.46980188  0.49173751  0.70665103  0.86719825  1.59058620  1.71047138
[19]  1.71446802  2.60427625

ks.test(x, y, alternative="two.sided")
```

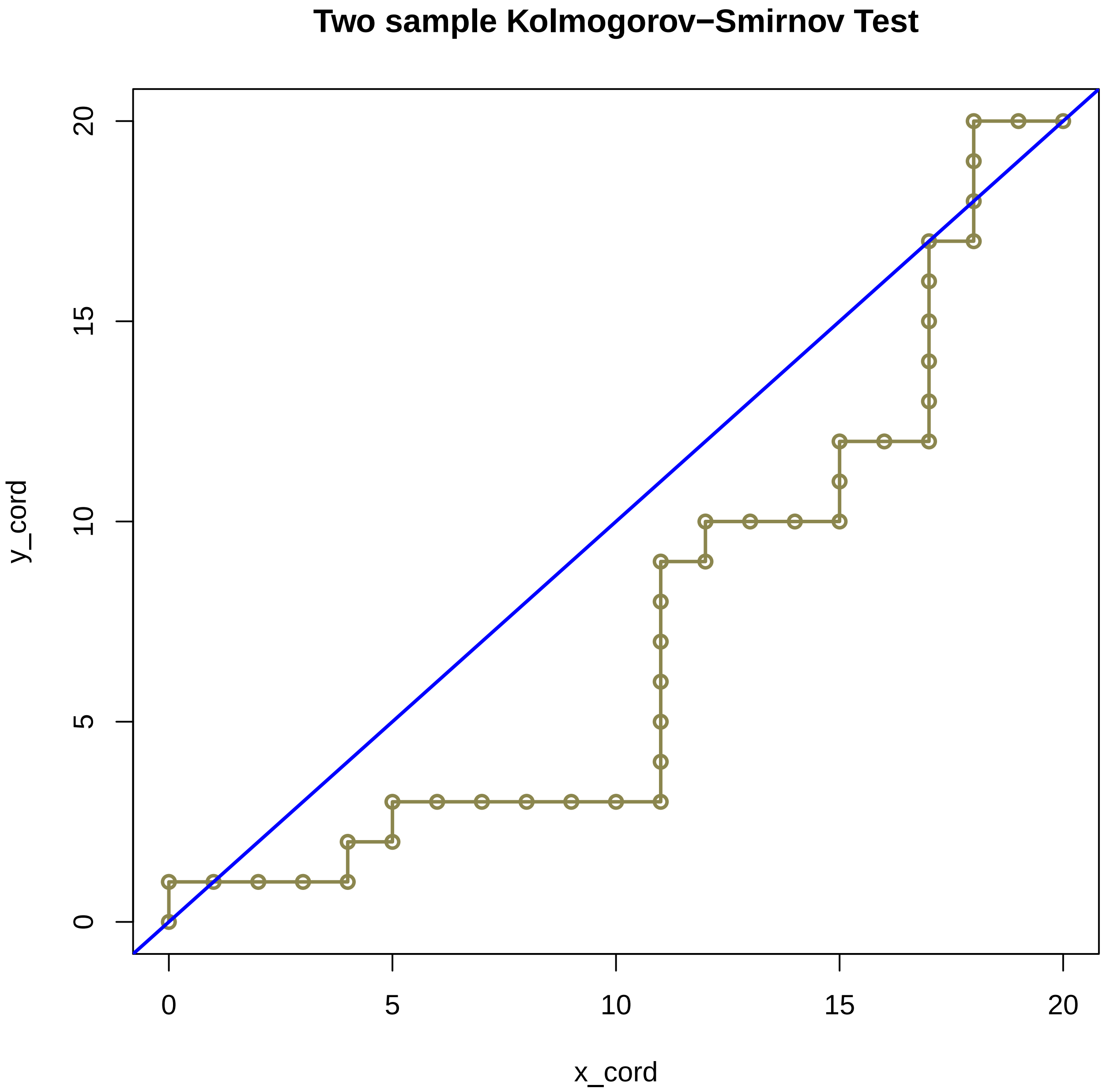
```
Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.2, p-value = 0.832
alternative hypothesis: two-sided
```

Comment:~

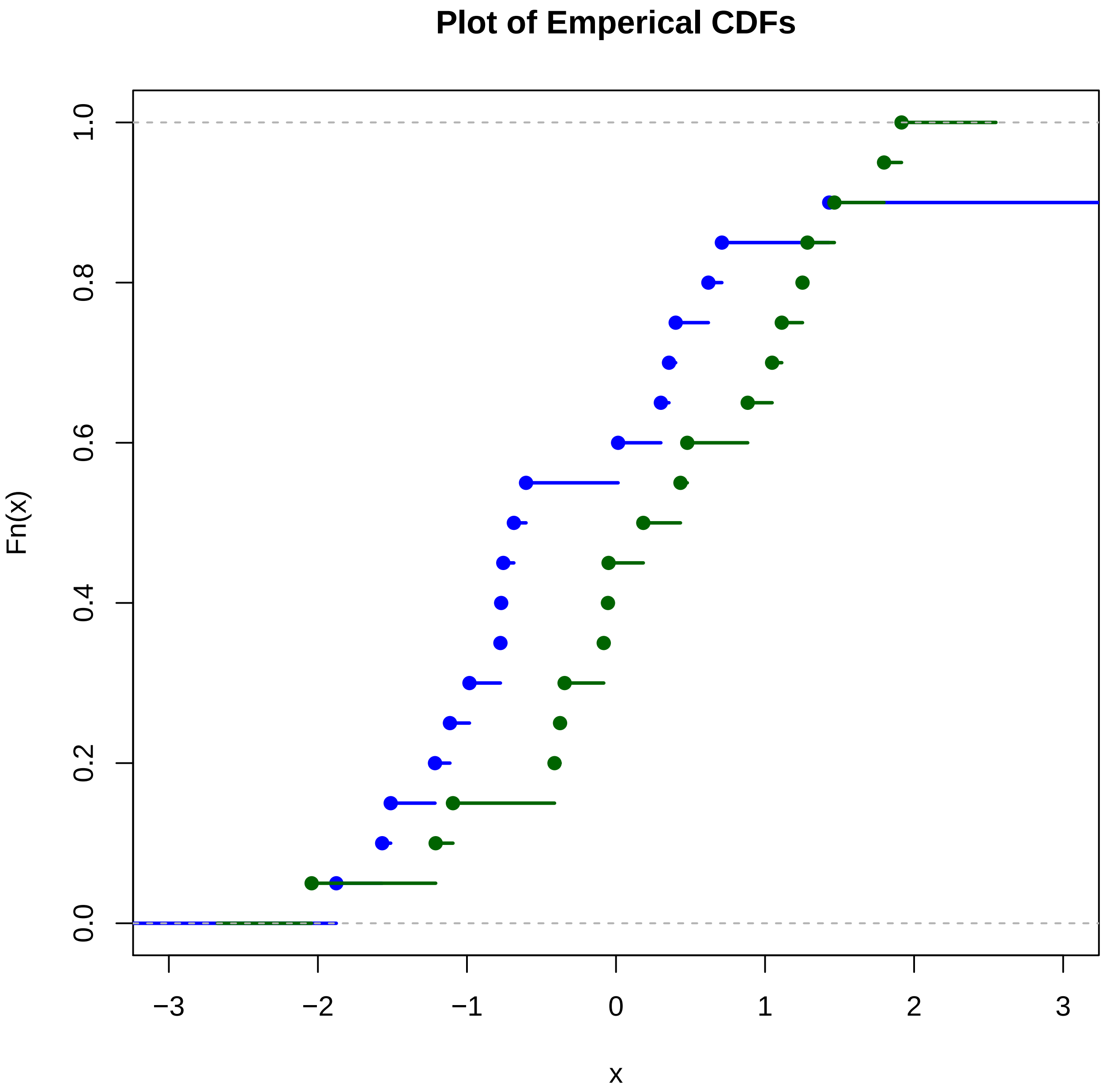
- ❊ *We have* $p-value = 0.832$ *which clearly indicates that, the unknown random sample (here* $Laplace(0,1)$*) does not have much evidence against the Null Hypothesis* $H_0 : F_1(Normal($*unknown data*$)) = F_2(Laplace($*unknown data*$))$*. So, we fail to reject* H_0 *in this case.*

Plots:

```
x1=rexp(20)
x2=rexp(20)
x <- x1-x2
y <- rnorm(20)

Fx <- ecdf(x)

Fy <- ecdf(y)

names(x) <- rep("x",length(x))
names(y) <- rep("y",length(y))

xy <- sort(c(x,y))
xyl <- noquote(names(xy))
xyl

 [1] y x x x x y x y x x x x x x y y y y y y x y x x x y y x x y y y y y x y y y
[39] x x

x_cord <- NULL
y_cord <- NULL
x_cord[1] <- y_cord[1] <- 0

for(i in 1:length(xyl))
{
  if(xyl[i] == "x")
  {
    x_cord[i+1] = x_cord[i] + 1
    y_cord[i+1] = y_cord[i] + 0
  }
  else
  {
    x_cord[i+1] = x_cord[i] + 0
    y_cord[i+1] = y_cord[i] + 1
  }
}

plot(x_cord,y_cord,col="khaki4",lwd=2,type = "o",
main="Two sample Kolmogorov-Smirnov Test")
abline(a = 0,b = 1,col="blue",lwd=2)
```



```
plot(ecdf(x),col = "blue",lwd = 2, xlim=c(-3, 3),
main="Plot of Emperical CDFs")
plot(ecdf(y), add=TRUE, col = "darkgreen",lwd = 2)
```



Large sample KS test:

- *As* $\min(n_1, n_2) \to \infty$ *,it can be shown that,*

$$Pr\left\{\sqrt{\frac{n_1 n_2}{n}} D^{+} \geq x|\mathcal{H}\right\} \backsimeq e^{-2x^2}, x \geq 0$$

$$Pr\left\{\sqrt{\frac{n_1 n_2}{n}} D^{-} \geq x|\mathcal{H}\right\} \backsimeq e^{-2x^2}, x \geq 0$$

$$Pr\left\{\sqrt{\frac{n_1 n_2}{n}} D \geq x|\mathcal{H}\right\} \backsimeq e^{-2x^2} = 2\sum_{k=1}^{\infty} (-1)^{k+1} e^{-2k^2x^2}$$

- *For right tailed test,* $\mathcal{H} : F_1 = F_2$ *vs* $\mathcal{K} : F_1 > F_2$ *we reject* H *at level* α *iff* $\sqrt{\frac{n_1 n_2}{n}} D^{+} \geq D_\alpha$.

and accept $\mathcal{H}$ *o.w .*
where D_α*is such that* $Pr\left\{\sqrt{\frac{n_1 n_2}{n}} D^{+} \geq D_\alpha|\mathcal{H}\right\} = \alpha$.

- *For left tailed test,* $\mathcal{H} : F_1 = F_2$ *vs* $\mathcal{K} : F_1 < F_2$ *we reject* H *at level* α *iff* $\sqrt{\frac{n_1 n_2}{n}} D^{-} \geq D_\alpha$.

and accept $\mathcal{H}$ *o.w .*
where D_α*is such that* $Pr\left\{\sqrt{\frac{n_1 n_2}{n}} D^{-} \geq D_\alpha|\mathcal{H}\right\} = \alpha$.

- *For both sided test,* $\mathcal{H} : F_1 = F_2$ *vs* $\mathcal{K} : F_1 \neq F_2$ *we reject* H *at level* α *iff* $\sqrt{\frac{n_1 n_2}{n}} D \geq D_\alpha$.

and accept $\mathcal{H}$ *o.w .*
where D_α*is such that* $Pr\left\{\sqrt{\frac{n_1 n_2}{n}} D \geq D_\alpha|\mathcal{H}\right\} = \alpha$.

Result:

Two sample Kolmogorov-smirnov test is exactly distribution free.

Justification:

Similarly like one sample kolmogorov-smirnov test we can show that the exact null distribution of kolmogorov-smirnov test statistic D^{+} *doesn't depend on the parent distribution.*

Hence the result.

Important facts:

- *The data is non parametric so it doesn't assume that data are sampled from Gaussian distributions or any other defined distributions.*

- *The* results will not change if we transform all the values to logarithms or any transformation*.A transformation will stretch the X-axis of the frequency distribution, but can't change the maximum distance between two frequency distributions.*

- *Since the test doesn't compare any particular parameter(like median),* it doesn't report any confidence interval*.*

Day 9:~

Topic: Median Test:-

This is a test of equality of medians of two independent distribution functions.

Test Procedure:-

- *Combine the two samples into one sample of size $(m+n)$ and order the $(m+n)$ values in ascending order of magnitude.*
- *Let,$Z_1 < ... < Z_{m+n}$ be the combined entire sample. Let,$\widetilde{Z}$ be the median of the combined entire sample. Let, V be the number of X_is, which are less than $\widetilde{Z}$. If the observed value of V is quite large, then one might suspect that $\zeta_{\frac{1}{2}}(x)$ is smaller than $\zeta_{\frac{1}{2}}(y)$. Hence, we reject $H_0 : F(x) = G(x)$, for all $x\epsilon\mathbb{R}$, in favour of $H_1 : F(x) \geq G(x)$ for all x,and $F(x) > G(x)$ for some x. If v_0 is observed value of V, then for the alternative $H_1 : F(x) \geq G(x)$ for all x and $F(x) > G(x)$ for some x, the p-value is$P_{H_0}[V \geq v_0]$.*
- *A test of the sort just described is called a median test. It is important to note that a median test will tend to accept $H_0 : F(x) = G(x)$, for all $x\epsilon\mathbb{R}$, even if the samples of F and G are different, as long as medians are equal.*

Null distribution of V :-

Null distribution of V is given by,

$P_{H_0}[V \geq v_0]$

$=P_{H_0}$*[Exactly v of the X_is are less then the median, in the combined sample].*

Case 1:-

There are exactly $\frac{m+n}{2} = p$ values, which are less than $\widetilde{Z}$ in the combined sample. p values can be assigned under $\widetilde{Z}$ in $\begin{pmatrix} m+n \\ p \end{pmatrix}$ways.

Under $H_0 : F(u) = G(u)$, for all $u\epsilon\mathbb{R}$, the X_is and Y_is are all i.i.d. random variables and all the possible cases are equally likely. For the favourable cases, v values of X_is and (p-v) values of Y_is can be assigned in $\begin{pmatrix} m \\ v \end{pmatrix}\begin{pmatrix} n \\ p-v \end{pmatrix}$ways.

Hence, $P_{H_0}[V = v_0] = \dfrac{\begin{pmatrix} m \\ v \end{pmatrix}\begin{pmatrix} n \\ p-v \end{pmatrix}}{\begin{pmatrix} m+n \\ p \end{pmatrix}}$.

Case 2:-

Here, $m+n=2p+1(odd)$.

$$P_{H_0}[V=v_0]=\frac{\binom{m}{v}\binom{n}{p-v}}{\binom{m+n}{p}}.$$ *Under* H_0, V *follows hypergeometric distribution.*

Asymptotic Distribution of V:-

- V *has a hypergeometric distribution, under* H_0.
- *Hence,* $E(V)=p.\frac{m}{m+n}$.

  $Var(V)=p.\frac{m}{m+n}.(1-\frac{m}{m+n})(\frac{m+n-p}{m+n-1})$, *under* H_0.
- *For large* m,n,

  $E(V)\simeq\frac{m}{2}$ *and* $Var(V)\simeq\frac{mn}{4(m+n)}$, *under* H_0.
- *It can be shown that,*

  $\frac{V-\frac{m}{2}}{\sqrt{\frac{mn}{4(m+n)}}}\overset{a}{\sim}N(0,1)$, *asymptotically, under for* $m,n>10$.
- *The p-value is* $P_{H_0}[V\geq v_0]\simeq\Phi(-\frac{v_0-\frac{m}{2}}{\sqrt{\frac{mn}{4(m+n)}}})$.

Remark:-

- *The counting in the median test can be put into a* 2×2 *contingency table.*

	$<\tilde{Z}$	$\geq\tilde{Z}$	*Total*
Values of X_i*'s*	v	$m-v$	m
Values of Y_j*'s*	u	$n-u$	n
Total	$u+v$	$m+n-u-v$	$m+n$

Under $H_0,\ u+v=p=\frac{m+n}{2}$.
For moderately large m,n, the frequency chi-square $\chi^2\overset{a}{\sim}\chi^2_{1,}asymptotically$.

- *Under* H_0,

$$\chi^2=\frac{(m+n)(nv-mu)^2}{(\frac{m+n}{2})^2mn}=\frac{(V-\frac{m}{2})^2}{\frac{mn}{4(m+n)}}\overset{a}{\sim}\chi^2_1$$

  for moderately large m,n.
- *We reject,* $H_0:F(u)=G(u)$, *against* $H_1:F(u)\neq G(u)$, *at level* α,

  if observed $\frac{(V-\frac{m}{2})^2}{\frac{mn}{4(m+n)}}>\chi^2_{\alpha,1}$.

Hence, for large samples a test based on frequency χ^2*and a test based on the assymptotic distribution,*
Under H_0, $\frac{V-\frac{m}{2}}{\sqrt{\frac{mn}{4(m+n)}}}\overset{a}{\sim}N(0,1)$, *asymptotically for the two sided alternative* $H_1:F(u)\neq G(u)$.

Day 10:∼

Topic : Linear Rank Statistic and General Two-Sample Problem

Introduction:

Suppose we have two independent random samples $X_1, X_2, \ldots, X_m$ and $Y_1, Y_2, \ldots, Y_n$ from populations with continuous CDFs F_X and F_Y respectively. For testing whether the two samples come from the same population, we take the following null hypothesis,

$$\mathscr{H}_0 : F_X(x) = F_Y(x) = F(x) \;\; \textit{for all } x, F \textit{ unspecified}$$

For testing the desired null hypothesis, it is quite natural to consider the rank of each observations in the combined sample. However , it is easier to denote the combined ordered sample by a vector of indicator random variables as follows. Let,

$$\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_N)$$

be the combined sample where $Z_i = \begin{cases} 1 & \textit{if } i^{th} \textit{ ordered sample comes from } X \\ 0 & \textit{if } i^{th} \textit{ ordered sample comes from } Y \end{cases}$ *for* $i = 1, 2, \ldots, N$ *where* $N = m + n$*. In simple words, The rank of the observation for which* Z_i *is an indicator is* i*, and therefore the vector* $\boldsymbol{Z}$ *indicates the rank-order statistics of the combined samples and in addition identifies the sample to which each observation belongs.*

Example:

For example given a X *sample* $(X_1, X_2, X_3, X_4) = (2, 9, 3, 4)$ *and a* Y *sample* $(Y_1, Y_2, Y_3) = (1, 6, 10)$*, the combined orderd sample is* $(1, 2, 3, 4, 6, 9, 10)$ *for which corresponding vector* $\boldsymbol{Z}$ *is* $(0, 1, 1, 1, 0, 1, 0)$*. Since* $Z_6 = 1$*, it implies that an observation from* X *has rank* 6 *in the combined sample.*

Linear Rank Statistic

Many of the statistics based on rank-order statistics which are useful in the two-sample problem can be easily expressed in terms of this notation. An important class of statistics of this type is called a linear rank statistic, defined as a linear function of the indicator variables Z*, as*

$$T_N(Z) = \sum_{i=1}^{N} a_i Z_i$$

where the a_i *are given constants called weights or scores. It should be noted that the statistic* T_N *is* LINEAR IN THE INDICATOR VARIABLES AND NO SIMILAR RESTRICTION IS IMPLIED FOR THE CONSTANTS.

Distribution Of Linear Rank Statistic:

Under the null hypothesis $\mathscr{H}_0 : F_X(x) = F_Y(x) = F(x)$ *for all* x*, we have for* $i = 1, 2, \ldots, N$*,*

$$E(Z_i) = \frac{m}{N} \quad V(Z_i) = \tfrac{mn}{N^2} \quad cov(Z_i, Z_j) = \frac{-mn}{N^2(N-1)}$$

hence,

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

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

$$\textit{where } \overline{a} = \sum_{i=1}^{N} \frac{a_i}{N}$$

Since each Z occurs with probability $1/\binom{N}{m}$, the exact probability distribution under the null hypothesis of any linear rank statistic can always be found by direct enumeration. The values of $T_N(Z)$ are calculated for each Z, and the probability of a particular value k is the number of Z vectors which lead to that number k divided by $\binom{N}{m}$. In other words, we have

$$P\left[T_N(Z)=k\right]=\frac{t(k)}{\binom{N}{m}}$$

where $t(k)$ is the number of arrangements of mX and nY random variables such that $T_N(Z)=k$. Naturally, the tediousness of enumeration increases rapidly as m and n increase.

✻ **Remark :** *The null distribution of $T_N(Z)$ is symmetric about its mean $\mu=m\sum_{i=1}^{N}\frac{a_i}{N}$ whenever the weights satisfy the relation,*

$$a_i+a_{N-i+1}=c \textit{ for some constant } c \textit{ for } i=1,2,\ldots,N$$

The Wilcoxon Rank-Sum Test and Confidence Interval

*The ranks of the X's in the combined ordered arrangement of the two samples will generally be larger than the ranks of the Y's if the MEDIAN OF THE X POPULATION EXCEEDS THE MEDIAN OF THE Y POPULATION. Therefore, **Wilcoxon (1945)** proposed a test where we accept the one-sided location alternative $\mathscr{H}_1$: $\theta<0\,(X>Y)$, if the sum of the ranks of the X's is TOO LARGE, or $\mathscr{H}_1:\theta>0\,(X<Y)$ if the sum of the ranks of the X's is TOO SMALL, and the two-sided location alternative $\mathscr{H}_1:\theta\neq 0$ if the sum of the ranks of the X's is EITHER TOO LARGE OR TOO SMALL. This function of the ranks expressed as a linear rank statistic has the simple weights $a_i=i, i=1,2,\ldots,N$. The Wilcoxon Rank-Sum test statistic is*

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

If there are no ties then, $E(W_N)=\frac{m(N+1)}{2}, V(W_N)=\frac{mn(N+1)}{12}$. The minimum value, W_N can obtain is $\sum\limits_{i=1}^{m} i=\frac{m(m+1)}{2}$ and the maximum it can attain is, $\sum\limits_{i=N-m+1}^{N} i=\frac{m(2N-m+1)}{2}$. Further, the statistic is symmetric about its mean under $\mathscr{H}_0:\theta=0$. Since,

$$a_i+a_{N-i+1}=N+1 \textit{ for } i=1,2,\ldots,N$$

If we denote by $r_{m,n}(k)$, the number of arrangements of mX and nY random variables such that the sum of the X ranks is equal to k, it is evident that

$$r_{m,n}(k)=r_{m-1,n}(k-N)+r_{m,n-1}(k)$$

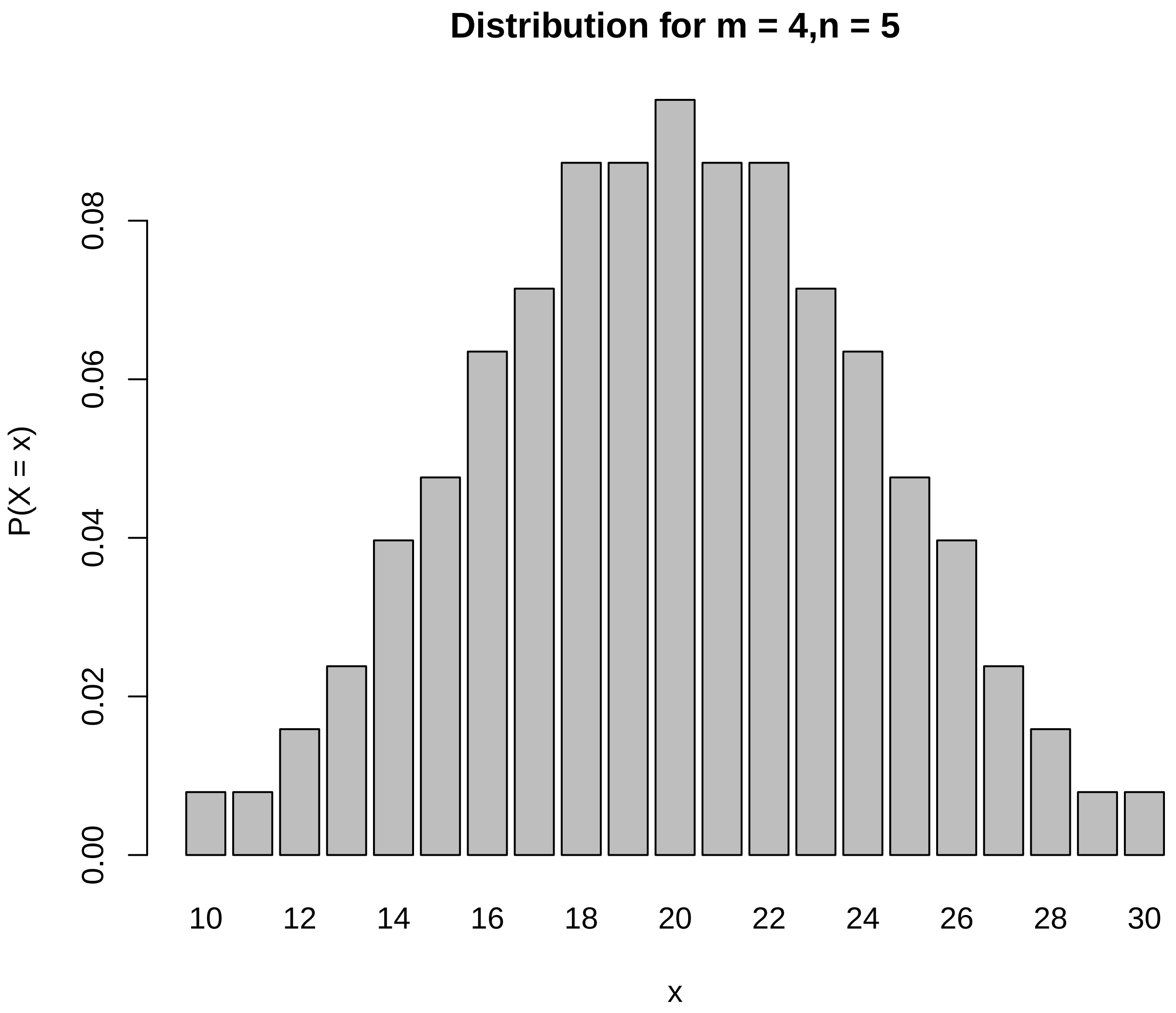
Null Distribution of W_N

We calculate the null distribution for $m=4, n=5$ using R.

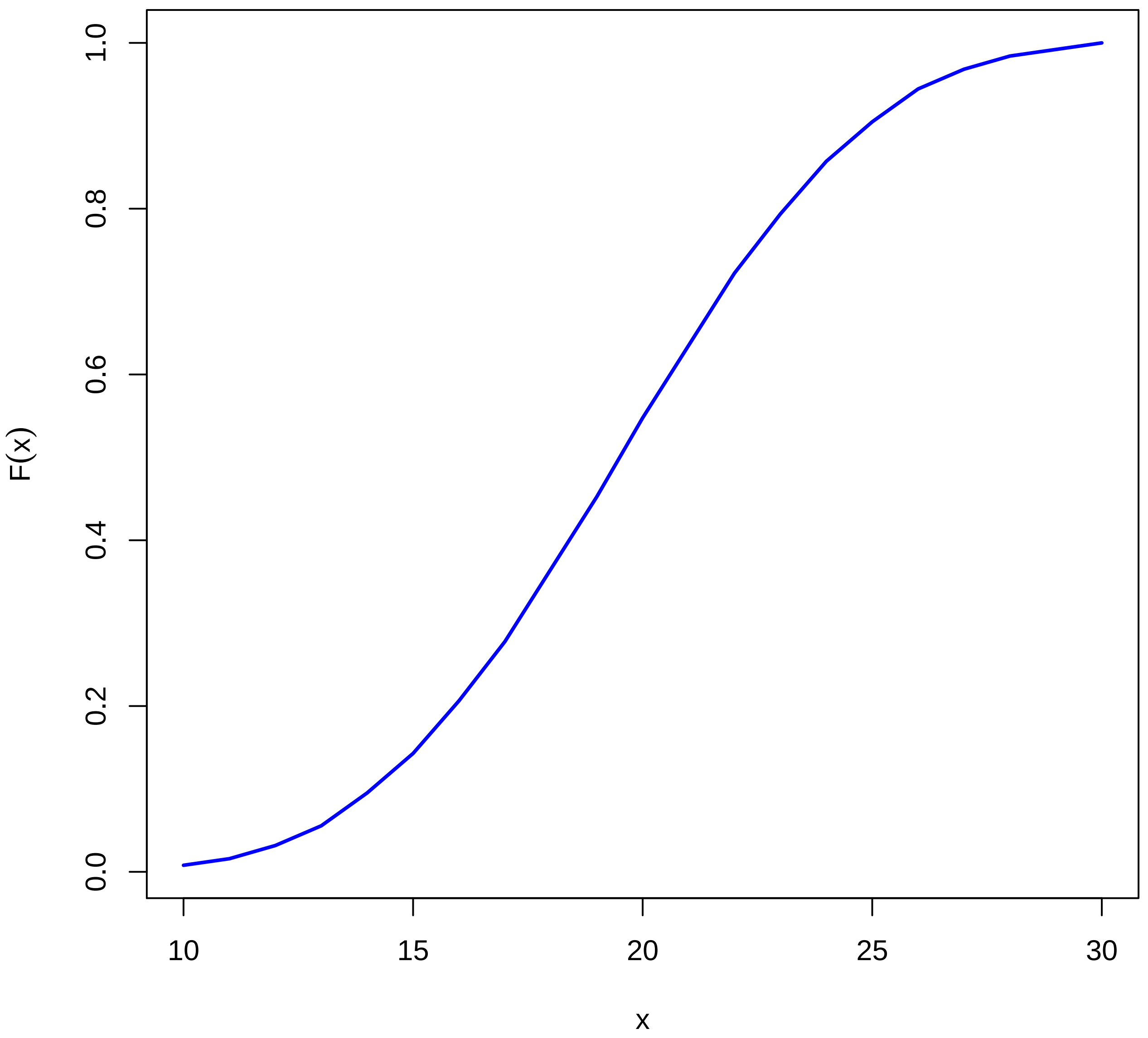
k	$r_{m,n}(k)$	$p_{m,n}(k)$	k	$r_{m,n}(k)$	$p_{m,n}(k)$
10	1	0.007936508	21	11	0.087301587
11	1	0.007936508	22	11	0.087301587
12	2	0.015873016	23	9	0.071428571
13	3	0.023809524	24	8	0.063492063
14	5	0.039682540	25	6	0.047619048
15	6	0.047619048	26	5	0.039682540
16	8	0.063492063	27	3	0.023809524
17	9	0.071428571	28	2	0.015873016
18	11	0.087301587	29	1	0.007936508
19	11	0.087301587	30	1	0.007936508
20	12	0.095238095			

❄ *The pmf of* W_N *for* $m = 4, n = 5$ *is :-*

```
Attaching package:  'combinat'
The following object is masked from 'package:utils':
        combn
```



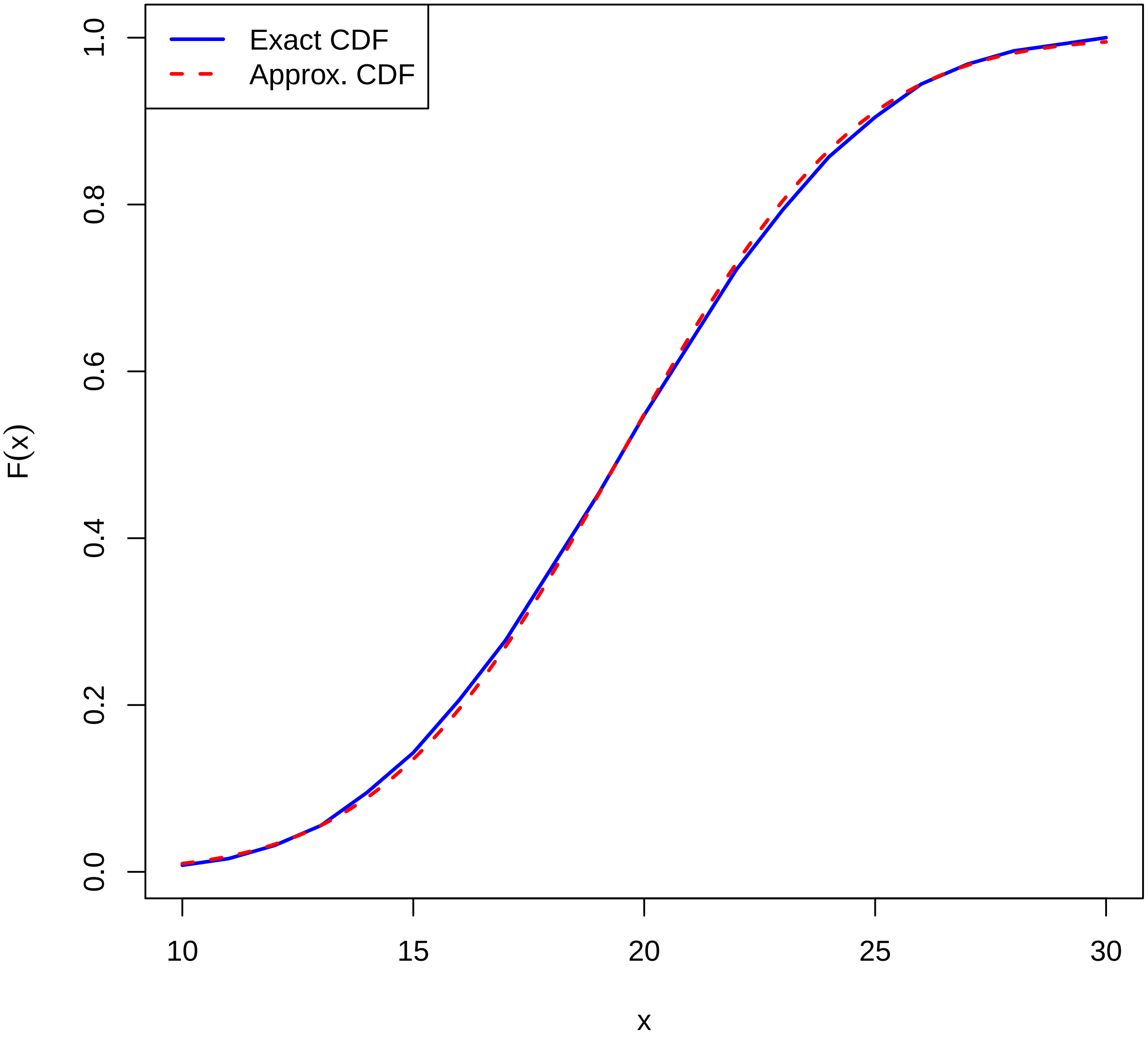
❇ *We also draw the exact CDF of W_N as the following :-*



❊ *Now from Large Sample Theory, we can approximate the distribution of W_N as,*

$$P\left(W_N \leq w\right) \approx \Phi\left(\frac{w_0 + 0.5 - {}^{m(N+1)}/_{2}}{\sqrt{{}^{mn(N+1)}/_{12}}}\right)$$

❊ *putting $m = 4, n = 5$, we again plot the approximate CDF of W_N :-*



Day 11:∼

Topic:Mann-Whitney-Wilcoxon-Test:

Introduction:

- The Mann-Whitney test can be used when the aim is to show a DIFFERENCE BETWEEN TWO GROUPS IN THE VALUE OF AN ORDINAL, INTERVAL OR RATIO VARIABLE.

- It is the NON-PARAMETRIC VERSION OF THE T-TEST, which can be used for interval, ratio or continuous data unless there are large departures from the parametric assumptions.

- When presenting the results of Mann-Whitney tests, the median of each group should be presented along with a description of the skewness of each sample (e.g. with a box plot).

- The MANN-WHITNEY TEST ALSO ASSUMES THAT THE TWO GROUPS ARE INDEPENDENT. Where the measurements are paired (i.e. are two measurements from the same individual) the Wilcoxon matched-pairs test should be used instead.

Basic-Idea:∼

- Let $X_1,X_2,.....,X_m$ and $Y_1,Y_2,....,Y_n$ be independent random samples from abs. continous distribution function $F_X(.)$ and $F_Y(.)$ respectively.

- To test,H_0:$F_X(u) = F_Y(u)\ \forall u$

Now we define,

$$D_{ij}=\begin{cases}1 & X_i > Y_j\\ 0 & X_i < Y_j\end{cases},\ \text{i=1(1)m and j=1(1)n}$$

Since both the populations are continous then $Pr(X_i = Y_j) = 0$.

Then,U_i=$\sum_{j=1}^{n}D_{ij}$,we count the number of y_j which are less than x_i, for each i=1(1)m.

Now we define,

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

- Also define, $p = Pr(Y < X) = \int_{-\infty}^{\infty}\int_{-\infty}^{x} f_X(x).f_Y(y)dydx = \int_{-\infty}^{\infty} F_Y(x)dF_X(x)$

The statistic U is known as MANN-WHITNEY STATISTIC.

Alternative form of Mann-Whitney test:

Let ,Q_i=The rank of x_i in the combined (m+n) values

={The number of y_j which are less than x_i}+{The rank of x_i among x_1,x_2,....,x_m}

=U_i+R_i

Then,$W=\sum_{i=1}^{m}Q_i=\sum_{i=1}^{m}U_i+\sum_{i=1}^{m}R_i$

Now,$U=\sum_{i=1}^{m}U_i$ *and* $\sum_{i=1}^{m}R_i=\frac{m(m+1)}{2}$.

so,$W=U+\frac{m(m+1)}{2}$

So we see that U AND W ARE LINEARLY RELATED RANK STATISTIC AND U AND W ARE EQUIVALENT STATISTIC .

Hence,Mann-Whitney Wilcoxon test.

Testing criteria:

Clearly,$0\leq U\leq mn$

Now U=0 =>ALL THE x_i ARE LESS THAN ALL THE y_j.

and $U=mn$=>ALL THE x_i ARE GREATER THAN ALL THE y_j.

If U *is large ,the value of* x *tends to be larger than the value of y.*

and this supports the alternative, H_1: $F_X(u)\leq F_Y(u)$ *i.e.*$X\underset{st}{\geq}Y$*i.e.*$p>0.5$

and on the other hand a small value of U *supports the alternative,* H_2:$F_X(u)\geq F_Y(u)$*i.e.*$X\underset{st}{\leq}Y$ *i.e.*$p<0.5$

and general alternative, H_3:$F_X(u)\neq F_Y(u)$ *i.e.*$p\neq 0.5$, $U'=\sum_{i=1}^{m}\sum_{j=1}^{n}(1-D_{ij})=mn-U$.

Alternate Hypothesis	*Critical Region*	*p-value*
$X\underset{st}{\geq}Y$ i.e. $p>0.5$	$\{U'\leq C_\alpha\}$i.e.$\{U\geq mn-C_\alpha\}$	$P_{H_0}(U'\leq u_0)$
$X\underset{st}{\leq}Y$ i.e.$p<0.5$	$\{U\leq C_\alpha\}$	$P_{H_0}(U\leq u_0)$
$F_X(u)\neq F_Y(u)$ i.e. $p\neq 0.5$	$\left\{U\leq C_{\alpha/2}\right\}$ or $\left\{U\geq C_{\alpha/2}\right\}$	$2\cdot min\{P_{H_0}(U'\leq u_0),P_{H_0}(U\leq u_0)\}$

Asymptotic distribution of Mann-Whitney statistic:

Under H_0,$E(D_{ij})=1.P(X_i>Y_j)=1/2$

As x_i *and* y_j *are iid continuous R.V.*

so,$E(U)$=$E(\sum_{i=1}^{m}\sum_{j=1}^{n}D_{ij})$

$=\sum_{i=1}^{m}\sum_{j=1}^{n}E(D_{ij})$

$=\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}$

$=\frac{mn}{2}$..................*(i)*

$E(U^2)$=$E(\sum_{i=1}^{m}\sum_{j=1}^{n}D_{ij})^2$

=$\sum_{i=1}^{m}\sum_{j=1}^{n}E(D_{ij}^2)$+$\sum_k\sum_j\sum_i E(D_{ij}D_{ik})$+$\sum_h\sum_j\sum_i E(D_{ij}D_{hj})$+$\sum_h\sum_k\sum_j\sum_i E(D_{ij}D_{hk})$.

Under H_0,
$E(D_{ij}^2)$=1/2
$E(D_{ij}D_{ik})$=$P(Y_j$<X_i,Y_k<$X_i)$=$\frac{2!}{3!}$=1/3,$j \neq k$.
$E(D_{ij}D_{hj})$=$P(Y_j$<X_i,Y_j<$X_h)$=$\frac{2!}{3!}$=1/3,i≠h.
$E(D_{ij}D_{hk})$=$P(Y_j$<X_i,Y_k<$X_h)$=$\frac{1}{2}.\frac{1}{2}$=1/4,j≠k and i≠h
so under H_0,
$Var(U) = E(U^2) - E^2(U)$
$= \frac{mn}{2} + \frac{m(m-1)n}{3} + \frac{n(n-1)m}{3} + \frac{m(m-1)n(n-1)}{4} - (\frac{mn}{2})^2$
$= \frac{mn(m+n+1)}{12}$......................(ii)

It can be shown that under H_0,
$\frac{U-\frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \sim$ N(0,1) when m,n are large.
This asymptotic distribution is applicable when both m,n>8.

Null distribution of Mann-whitney statistic:

❉ Let,$P_{H_0}(U = u) = p_{m,n}(u)$

Under H_0, each of the $\binom{m+n}{m}$ arrangements of the random variable occours with equal probability

$f_U(u) = Pr(U = u) = \frac{r_{m,n}(u)}{\binom{m+n}{m}}$,where $r_{m,n}(u)$ is the no. of distinguishable arrangements of m no. of x 's and n no. of y 's.

❉ Recursive relation : If the alternatives are arranged in increasing order of magnatude ,then the largest value can be either "a y-valued (B)" or "a y-valued (B^c)"

$$p_{m,n}(u) = P_{H_0}[U = u|B] + P_{H_0}[U = u|B^c] = p_{m-1,n}(u).\frac{n}{m+n} + p_{m,n-1}(u).\frac{m}{m+n}.$$

Notes:∼

Let, z be the perticular arrangement of m no. of x 's and n no. of y 's and z' as the sequence z written backward.

Under H_0, every y that preceeds an x in z that y follows that x in z'. So, if u is the value for z, then (mn-u) is the value for z'.

$Pr(U - \frac{mn}{2} = u)$
$= Pr(U = \frac{mn}{2} + u)$
$= Pr(U = mn - (\frac{mn}{2} - u))$
$= Pr(U = \frac{mn}{2} - u)$ [under H_0,$r_{m,n}(u) = r_{m,n}(mn - u)$]

Hence,U is symmetric about the $\frac{mn}{2}$ under H_0.

So,The Mann-Whitney wilcoxon test is symmetric about the mean under H_0.

The problem of Ties:

It is possible that two or more observations may be same. If this is the case we can still calculate U by allocating half the tie to the x value and half the tie to the y value.

In case of Ties the formula for the standard deviation is more complicated.

σ_{tie}=$\sqrt{\frac{mn}{2}[(N+1)-\sum_{i=1}^{k}\frac{t_i^3-t_i}{n(n-1)}]}$

Where N=m+n, t_i is the number of subjects sharing rank i and k is the number of ranks.

if the number of ties is small ties can be ignored when doing calculations by hand.

- ❈ *The conservative approach tells us to break all ties such that U (or U') has largest value in the alternatives H_2(or H_1) i.e. break all the ties in favour of the Null.*

Day 12:~

Topic: Mann-Whitney Confidence Interval for θ

Thinks & Thoughts :

- *If the populations from which the X and Y samples are drawn are identical in every respect except location, say $F_Y(x) = F_X(x-\theta)$ for all x and some θ, we say that the Y population is the same as the X population but shifted by an amount θ, which may be either positive or negative, the sign indicating the direction of the shift.*

- *We wish to use the Mann-Whitney test procedure to find a confidence interval for θ, the amount of shift. Under the assumption that $F_Y(x) = F_X(x-\theta)$ for all x and some θ, the sample observations $X_1, X_2, ..., X_m$ and $Y_1-\theta, Y_2-\theta, ..., Y_n-\theta$ come from identical populations By a confidence interval for θ with confidence coefficient $1-\alpha$ we mean the range of values of θ for which the null hypothesis of identical populations will be accepted at significance level α.*

- *To apply the Mann-Whitney procedure to this problem, the random variable U now denotes the number of times a $Y-\theta$ precedes an X, that is, the number of pairs $(X_i, Y_j-\theta)$, $i = 1, 2, ..., m$ and $j = 1, 2, ..., n$, 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 of $U \leq k$, (say) $U \geq mn-k$, we reject H_o when no more than k and no less than $(mn-k)$ differences Y_j-X_i are less than the value θ, and accept H_o otherwise.*

- *The total number of differences Y_j-X_i is mn. If these differences are ordered from smallest to largest according to actual (not absolute) magnitude, denoted by $D_{(1)}, D_{(2)}, ..., D_{(mn)}$.*

Hunting for confidence limits :

Target 1 : The lower confidence limit

- ***Case 1 :*** *If $\theta \in (D_{(k-1)}, D_{(k)}]$*
  - *Then exactly $(k-1)$ differences of the paires (Y_j, X_i) are less than the value θ.*
  - *$Y_j - X_i < \theta$ for only $(k-1)$ pairs.*
  - *$U = k-1$*
  - *So, our left rejection region is $U \leq k \implies$ we reject H_0*

- ***Case 2 :*** *If $\theta \in (D_{(k)}, D_{(k+1)}]$*
  - *Then exactly k differences of the paires (Y_j, X_i) are less than the value θ.*
  - *$Y_j - X_i < \theta$ for only k pairs.*

- ➥ $U = k - 1$
- ➥ *So, our left rejection region is* $U \leq k \implies$ *we reject* H_0

❀ ***Case 3 :*** *If* $\theta \in (D_{(k+1)}, D_{(k+2)}]$

- ➥ *Then exactly* $k{+}1$ *differences of the paires* (Y_j, X_i) *are less than the value* θ*.*
- ➥ $Y_j - X_i < \theta$ *for only* $k + 1$ *pairs.*
- ➥ $U = k + 1$
- ➥ *So, our left rejection region is* $U \leq k \implies$ *we accept* H_0 *henceforth upto upper limit.*

Target 2 : *The upper confidence limit*

❀ ***Case 1 :*** *If* $\theta \in (D_{(mn-(k-1))}, D_{(mn-(k-2))}]$

- ➥ *Then exactly* $mn - (k - 1)$ *differences of the paires* (Y_j, X_i) *are less than the value* θ*.*
- ➥ $Y_j - X_i < \theta$ *for only* $mn - (k - 1)$ *pairs.*
- ➥ $U = mn - (k - 1) = mn - k + 1$
- ➥ *So, our right rejection region is* $U \geq mn - k \implies$ *we reject* H_0

❀ ***Case 2 :*** *If* $\theta \in (D_{(mn-k)}, D_{(mn-(k-1))}]$

- ➥ *Then exactly* $mn - k$ *differences of the paires* (Y_j, X_i) *are less than the value* θ*.*
- ➥ $Y_j - X_i < \theta$ *for only* $mn - k$ *pairs.*
- ➥ $U = mn - k$
- ➥ *So, our right rejection region is* $U \geq mn - k \implies$ *we reject* H_0

❀ ***Case 3 :*** *If* $\theta \in (D_{(mn-(k+1))}, D_{(mn-k)}]$

- ➥ *Then exactly* $k{+}1$ *differences of the paires* (Y_j, X_i) *are less than the value* θ*.*
- ➥ $Y_j - X_i < \theta$ *for only* $k + 1$ *pairs.*
- ➥ $U = k + 1$
- ➥ *So, our right rejection region is* $U \geq mn{-}k \implies$ *we accept* H_0 *for less values of* θ *upto lower limit.*

Hence, we can conclude that the required $100(1 - \alpha)\%$ *Confidence Interval for* θ *is :~*

$$(D_{(k+1)}, D_{(mn-k)}]$$

Day 13:∼

Topic: Kruskal-Wallis Test

Introduction:

- *Kruskal wallis is a non parametric test that can be used to determine whether two or more independent samples were selected from populations having the same distribution.*
- *Also this way it can be thought as an extension of the Mann-Whitney U test, which can be used for only two groups.*
- *It is also known as Kruskal-Wallis H test since* WILLIAM KRUSKAL AND W.ALLEN WALLIS FIRST PUBLISHED THIS METHOD IN THE YEAR 1952.

Important points:∼

- *Kruskal wallis test is equivalent to the* ONE-WAY ANOVA.
- *An* EXTENSION OF THE MANN-WHITNEY U TEST.
- *Sometimes we call it the* ONE WAY ANOVA ON RANKS.
- *The kruskal wallis test will tell us if there is a significant difference between groups.*
- *We use the sums of the ranks of the different samples to compare the distributions.*
- *A* SIGNIFICANT KRUSKAL-WALLIS TEST INDICATES THAT AT LEAST ONE SAMPLE STOCHASTICALLY DOMINATES ONE OTHER SAMPLE.

Assumptions:

- *Each sample must be randomly selected.*
- *The size of the each sample must be at least 5.*
- *Observations should be independent.*
- VARIABLES SHOULD BE MEASURED ON AN ORDINAL SCALE OR A CONTINUOUS SCALE.

K-W one way ANOVA test and multiple comparison:

The Kruskal-Wallis test is the natural extension of the wilcoxon test for location with two independent samples to the situation of k mutually independent samples from continuous populations. The null hypothesis is that the k populations are same. But when we assume the location model this hypothesis can be written in terms of the respective location parameters as :-

$$\mathscr{H}_0 : \theta_1 = \theta_2 = \cdots = \theta_k$$
$$\mathscr{H}_1 : \text{At least two } \theta' s \text{ differ}$$

To perform the test all $n_1 + n_2 + ...n_k = N$ observations are pooled into a single array and ranked from 1 to N.

Method :~

- *Since under $\mathscr{H}_0$ we have essentially a single sample of size N from the common population, combine the N observations into a single ordered sequence from smallest to largest and assign the ranks $1, 2, .., N$ to the sequence. If adjacent ranks are well distributed among the k samples, the total sum of ranks $\sum\limits_{i=1}^{N} i = \frac{N(N+1)}{2}$, would be divided proportionally according to sample size among the k samples and will be denoted by,*

$$R_i = \sum_{j=1}^{n_i} r_{ij}$$

- *For the i th sample which contains n_i observations , under null hypothesis$\mathscr{H}_0$ the expected sum of ranks would be :-*

$$E\left(R_i\right) = E\left(\sum_{j=1}^{n_i} r_{ij}\right) = \sum_{j=1}^{n_i} E\left(r_{ij}\right) = \sum_{j=1}^{n_i} \frac{N+1}{2} = \frac{n_i\left(N+1\right)}{2}$$

- *which can also be thought alternatively as the proportion of total rank sum for each sample of size n_i i.e,*

$$\frac{n_i}{N} \frac{N\left(N+1\right)}{2} = \frac{n_i\left(N+1\right)}{2}$$

- *Since the deviation for each group from its expected rank sum i.e., $R_i - \frac{n_i(N+1)}{2}$ can be thought as a measure of deviation from the null assumption, a reasonable test statistic could be based on a function of the all these deviations. Since deviations in either direction indicate disparity between the samples and absolute $(|.|)$ values are not particularly tractable mathematically, the sum of squares of these deviations can be employed as,*

$$S = \sum_{i=1}^{k} \left[R_i - \frac{n_i\left(N+1\right)}{2}\right]^2$$

Hence, the null hypotheis should be rejected for large value of S.

Null Distribution of S (no tie case)

- *In order to determine the null probability distribution of S, we first consider all the possible arrangements of ranks $1, 2, \ldots, N$ into k groups of size n_i each.*
- *This can be done in $\frac{N!}{\prod\limits_{i=1}^{k} n_i!}$.*
- *Then for each of these arrangements, we calculate the value of the S statistic and let us denote by $t\left(s\right)$ number of arrangements for which $S = s$, then the corresponding probability of S taking the value s is,*

$$f\left(s\right) = \frac{t\left(s\right)}{\frac{N!}{\prod\limits_{i=1}^{k} n_i!}} = t\left(s\right) \frac{\prod\limits_{i=1}^{k} n_i!}{N!}$$

Drawbacks of S Statistic:

- *First of all the calculation for S becomes very tedious for even $n_i \geq 5$ as the number of such arrangements rapidly increase with increasing values of n_i's.*
- *Also* <u>THERE IS NO STANDARD ASYMPTOTIC DISTRIBUTION FOR S WHICH CAN BE USED FOR LARGE SAMPLE TESTS</u>.
- *S only consider the sum of square of deviations of R_i from its mean but it do not standarize the observations R_i.*

Kruskal-Wallis Test Statistic:

- Due to all the drawbacks of the S statistic, A BETTER STATISTIC COULD BE A WEIGHTED SUM OF SQUARES OF DEVIATIONS WITH THE RECIPROCALS SAMPLE SIZE USED AS WEIGHTS,THEN THE TEST WILL BE MORE USEFUL AND SIGNIFICANT.This test statistic,due to Kruskal and Wallis (Kruskal -Wallis H Statistic) is defined as :-

$$\begin{aligned} H &= \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{1}{n_i}\left[R_i - \frac{n_i(N+1)}{2}\right]^2 \\ &= \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{1}{n_i}\left[n_i\overline{R}_i - \frac{n_i(N+1)}{2}\right]^2 \\ &= \frac{12}{N(N+1)}\sum_{i=1}^{k}n_i\left[\overline{R}_i - \frac{(N+1)}{2}\right]^2 \end{aligned}$$

- Here by R_i, we denote the i^{th} average rank sum $\overline{R}_i = {}^{R_i}/_{n_i}, i = 1(1)k$.

- H can also be written as $H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1)$.

Mean and Variance of $\overline{R}_i$:

- *Under null assumption, the k groups can be thought as* SRSWOR SAMPLES *of size n_i each from a $U\{1,2,\ldots,N\}$ population. Hence,*

$$\textit{Population Mean}: \mu = \frac{N+1}{2}$$
$$\textit{Population Variance}: \sigma^2 = \frac{N^2-1}{12}$$

- *Similarly, $\overline{R}_i$ can be thought as the sample mean of a SRSWOR sample of size n_i hence,*

$$\begin{aligned} E\left(\overline{R}_i\right) &= \mu = \frac{N+1}{2} \\ V\left(\overline{R}_i\right) &= \frac{\sigma^2}{n_i}\frac{N-n_i}{N-1} = \frac{N^2-1}{12n_i}\frac{N-n_i}{N-1} = \frac{(N+1)(N-n_i)}{12n_i} \end{aligned}$$

Asymptotic Distribution of H:~

If n_i is large, the standarized random variable

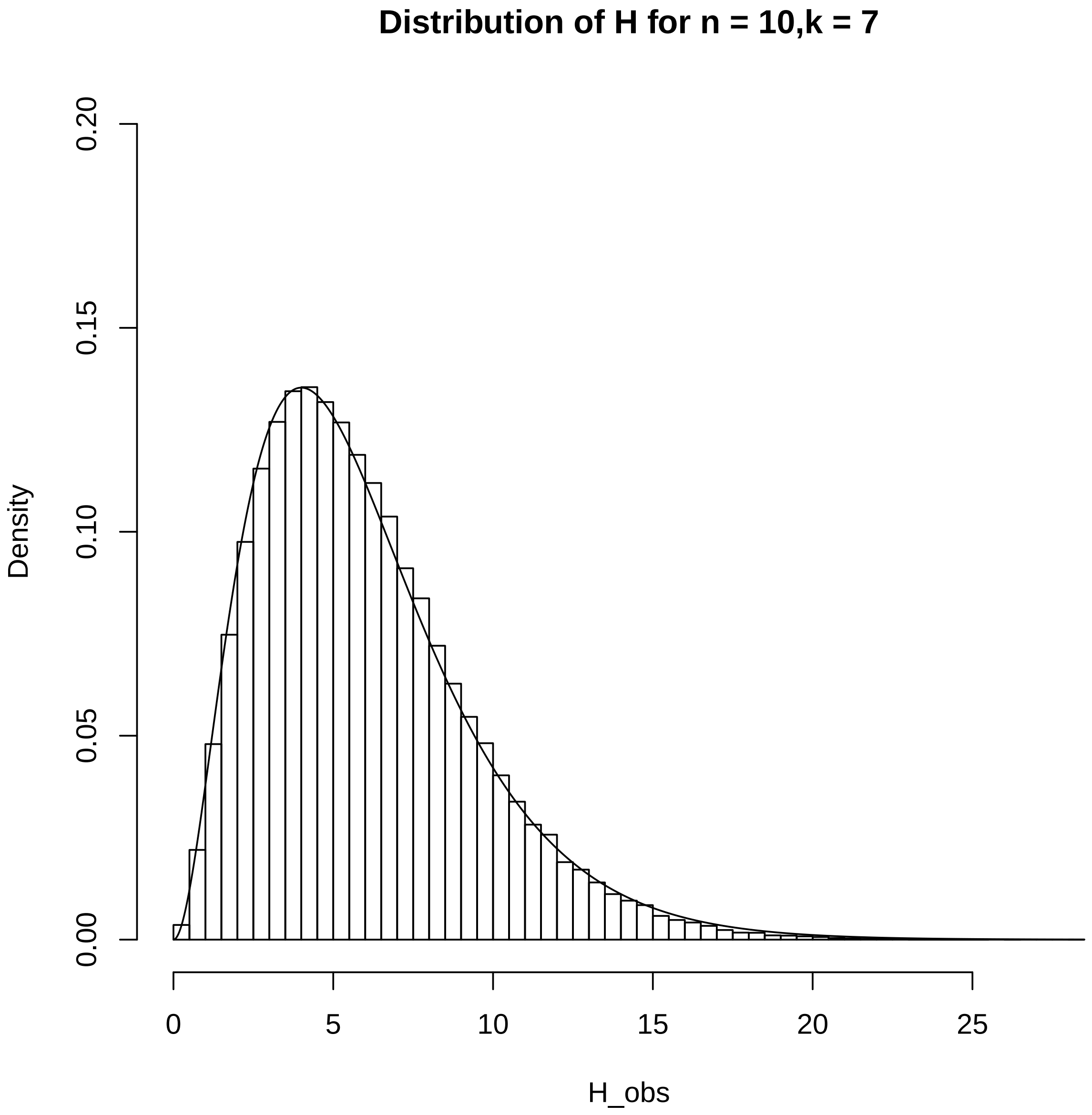
$$z_i = \frac{\overline{R}_i - \frac{N+1}{2}}{\sqrt{\frac{(N+1)(N-n_i)}{12n_i}}} \overset{d}{\sim} N(0,1)$$

by LINDEBERG-LEVY CLT.

Since, there is a linear dependence between the quantities $\overline{R}_i$ as, the total of all rank sums $\sum_{i=1}^{k} n_i\bar{R}_i = \frac{N(N+1)}{2}$, so all the k Z_i's can't be independently distributed (atmost $k-1$ of them can). So it can be shown if no n_i is very small then, under $\mathscr{H}_0$, $H = \sum_{i=1}^{k}\frac{N-n_i}{N}Z_i^2$ is approximately distributed as a Chi-Squared Distribution with $k-1$ degrees of freedom $\left(\chi^2_{k-1}\right)$. Hence, we reject $\mathscr{H}_0$ at α level of significance if $H_{obs} \geq \chi^2_{\alpha;k-1}$.

Approximate Sampling Distribution of H:

Using R we plot histogram of 100000 observed values of H for $n = 10, k = 7$:-



Tie Case:

If ties to the extent t are present and are handled by the midrank method, the variance of the finite population is,

$$\sigma^2 = \frac{N^2-1}{12} - \frac{\sum t\left(t^2-1\right)}{12}$$

then the Kruskal-Wallis Statistic becomes,

$$H^{'} = \sum_{i=1}^{k} \frac{N-n_i}{N} \left\{ \frac{\left[\overline{R}_i - \frac{N+1}{2}\right]^2}{\frac{(N+1)(N-n_i)}{12n_i} - \frac{N-n_i}{n_i(N-1)} \frac{\sum t(t^2-1)}{12}} \right\}$$
$$= \sum_{i=1}^{k} \frac{\left[\overline{R}_i - \frac{N+1}{2}\right]^2}{\frac{N(N+1)}{12n_i}\left[1 - \frac{\sum t(t^2-1)}{N(N^2-1)}\right]} = \frac{H}{1 - \frac{\sum t(t^2-1)}{N(N^2-1)}}$$

So, we just need to divide the original H statistic by the factor $1 - \frac{\sum t\left(t^2-1\right)}{N(N^2-1)}$.

Pairwise Comparison:

IF THE NULL HYPOTHEIS IS REJECTED, ONE MAY NATURALLY WANT TO COMPARE DIFFERENT GROUPS PAIRWISE TO CHECK IF THEIR LOCATION PARAMETERS ARE EQUAL OR NOT.

From the asymptotic normal distribution of Z_i*, we can easily make groupwise comparisons using the statistic* $Z_{ij}, 1 \leq i < j \leq k$*,*

$$Z_{ij} = \frac{R_i - R_j}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}} \overset{d}{\sim} N(0,1)$$

We reject our null hypothesis $\mathscr{H}_0^{ij} : \theta_i = \theta_j$ *at* α^* *level of significance if,*

$$\left|Z_{ij(obs)}\right| > \tau_{\alpha^*} \text{ where } \alpha^* = \frac{\alpha}{k(k-1)}$$

since, we are comparing $^{k(k-1)}/_{2}$ *many pairs,*

$$\begin{aligned} P\left(\mathscr{H}_0^{ij} \text{ accepted } \forall i,j\right) &= P\left(\bigcap_{1\leq i<j\leq k} \mathscr{H}_0^{ij} \text{ accepted }\right) \\ &\geq \sum_{1\leq i<j\leq k} P\left(\mathscr{H}_0^{ij} \text{ accepted }\right) - \frac{k(k-1)}{2} \\ &= \sum_{1\leq i<j\leq k} (1-\alpha^*) - \frac{k(k-1)}{2} = 1-\alpha \end{aligned}$$

In words, the probability that all the statements are correct or all the pairs have equal location parameters, is atleast $1-\alpha$*. Hence, we take,* $\alpha \geq 0.20$ *because we are making such large number of statements.*