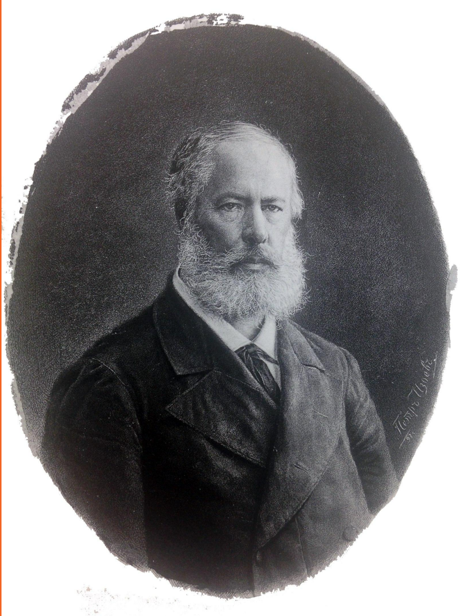


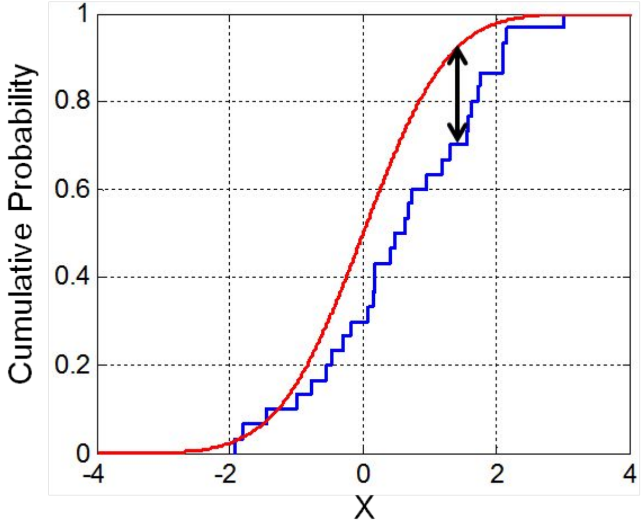
KOLMOGOROV SMIRNOV'S ONE SAMPLE TEST

A Brief Guide to KS One Sample Test

KOLMOGOROV SMIRNOV TEST

The Kolmogorov–Smirnov test is a nonparametric goodness-of-fit test

- It is used to determine whether two distributions differ, or whether an underlying probability distribution differs from a hypothesized distribution.



WHEN TO USE KOLMOGOROV SMIRNOV TEST??

- **It is used when we have two samples coming from two populations that can be different.**

- **It can also be used as a goodness-of-fit test. In this case, we have only one random sample obtained from a population where the distribution function is specific and known.**

What Is One-Sample Kolmogorov-Smirnov Test?

- The Kolmogorov Smirnov's one sample test is concerned with the degree of agreement between the distribution of the observed sample values and some specified theoretical distribution.

- It determines whether or not the values in a sample can reasonably be thought to have come from a population having a theoretical distribution.

Tip

This test is used as a test of **goodness of fit** and is ideal when the size of the sample is small.

Formula:

$$D = \text{Maximum}|\ Fo(X) - Fr(X)\ |$$

Where ,

- $Fo(X)$ = Observed cumulative frequency distribution of a random sample of n observations.
  $Fo(X) = k/n$

= (Number of observations ≤ X)/(Total number of observations)

- $Fr(X)$ = The theoretical frequency distribution.

The critical value of D is found from the K-S table values for one sample test.

Decision Criteria

➔ **Acceptance Criteria**

If calculated value is less than critical value accept null hypothesis

➔ **Rejection Criteria**

If calculated value is greater than table value reject null hypothesis

Example 1:

In a study done from various streams of a college 60 students, with equal number of students drawn from each stream, are we interviewed and their intention to join the Drama Club of college was noted.

	B.Sc	B.A.	B.Com	M.A	M . Com
No in Each Class	5	9	11	16	19

- It was expected that 12 students from each class would join the Drama Club. Using the K-S test to find if there is any difference among student classes with regard to their intention of joining the Drama Club

Solution:

- **Ho: There is no difference among students of different streams with respect to their intention of joining the drama club.**

Streams	No. of students interested in joining		FO(X)	FT(X)	|FO(X)−FT(X)||
	Observed (O)	Theoretical (T)			
B.Sc.	5	12	5/60	12/60	7/60
B.A.	9	12	14/60	24/60	10/60
B.COM.	11	12	25/60	36/60	11/60
M.A.	16	12	41/60	48/60	7/60
M.COM.	19	12	60/60	60/60	0
Total	n=60				

Test statistic |D| is calculated as:

D=Maximum|F0(X) −FT(X)|

=11/60=0.183

The table value of D at 5% significance level is given by :

D0.05=1.36/√n=1.36/√60

=0.175

- **Since the calculated value is greater than the critical value, hence we reject the null hypothesis and conclude that there is a difference among students of different streams in their intention of joining the Club.**

Solution Using R Software(Codes):

```r

                              #### KOLMOGROV SMIRNOV'S ONE SAMPLE TEST #####

#Example 1:

#Ho: There is no difference among students of different streams
 #with respect to their intention of joining the drama club
#H1: Not Ho
n=5
N=60
obs=c(5,9,11,16,19)
Theo=60/n
Theo
FO_xi=cumsum(obs/N)
FT_xi=cumsum(Theo/N)
Theo=rep(60/n,5)

Theo
FO_xi=cumsum(obs/N)
FT_xi=cumsum(Theo/N)
dc1=abs(FO_xi-FT_xi)
Table1=data.frame(obs,Theo,FO_xi,FT_xi,dc1)
Table1
D=max(dc1)
D
#D0.05=1.36/√n=1.36/√60=0.175
#Since the calculated value is greater than the critical value, hence we reject the null hypothesis
#ie 0.1833> 0.175
#conclude that there is a difference among students of different streams in their intention of joining the Clu


#-----------------------------------------------------------------------------------------------
```

Output:

Console   Terminal ×   Jobs ×

~/

```
> #Example 1:
>
> #Ho: There is no difference among students of different streams
>  #with respect to their intention of joining the drama club
> #H1: Not Ho
> n=5
> N=60
> obs=c(5,9,11,16,19)
> Theo=60/n
> Theo
[1] 12
> FO_xi=cumsum(obs/N)
> FT_xi=cumsum(Theo/N)
> Theo=rep(60/n,5)
>
> Theo
[1] 12 12 12 12 12
> FO_xi=cumsum(obs/N)
> FT_xi=cumsum(Theo/N)
> dc1=abs(FO_xi-FT_xi)
> Table1=data.frame(obs,Theo,FO_xi,FT_xi,dc1)
> Table1
  obs Theo      FO_xi FT_xi       dc1
1   5   12 0.08333333   0.2 0.1166667
2   9   12 0.23333333   0.4 0.1666667
3  11   12 0.41666667   0.6 0.1833333
4  16   12 0.68333333   0.8 0.1166667
5  19   12 1.00000000   1.0 0.0000000
> D=max(dc1)
> D
[1] 0.1833333
> #D0.05=1.36/vn=1.36/v60=0.175
> #Since the calculated value is greater than the critical value, hence we reject the null hypothesis
> #ie 0.1833> 0.175
> #conclude that there is a difference among students of different streams in their intention of joining the Club
>
```

Example 2:

Using K-S Test, check for the property of uniformity for the input set of random numbers. 0.54, 0.73, 0.98, 0.11, 0.68. Assume level of significance to be 0.05

- Solution:
  Ho: Random no. are not uniform V/S H1: Not Ho N=5

i	1	2	3	4	5
Ri	0.11	0.54	0.68	0.73	0.98

Example 2:

$$D^{+} = \max_{1 \le i \le N} \{i/N - R_i\} = \boxed{0.09}$$

$$D^{-} = \max_{1 \le i \le N} \{ R_i - (i-1/n)\} = \boxed{0.34}$$

$$D = \max \{ D^{+}, D^{-}\}$$

$$D = \max \{0.09, 0.34\}$$

$$\boxed{D = 0.34}$$

Example 2:

$D+= \max \{i/N- R_i\} = \boxed{0.09} \quad 1 <- i <- N$

$D- = \max \{ R_i- (i-1/n)\} = \boxed{0.34} \quad 1 <- i <- N$

$D = \max \{ D+, D- \}$

$D = \max \{0.09, 0.34\}$

$\boxed{D = 0.34}$

$\alpha = 0.05$

$\therefore$ **$D\alpha$=0.565** (using KS Table)

$\because$ 0.34< 0.565

$\because$ $D < D\alpha$

$\therefore$ we do not reject Ho

Conclusion: Random numbers are not uniform

Solution Using R Software(Codes):

```r
#Example 2:

#Ho: Random no. are not uniform   V/S   H1: Not Ho
N=5
i=c(1:5)
Ri=c(0.54, 0.73, 0.98, 0.11, 0.68)
Ri=sort(Ri)
D_plus=(i/N-Ri)
D_mins=(Ri-(i-1)/N)
Table2=data.frame(i,Ri,D_plus,D_mins)
Table2
max(D_plus)
max(D_mins)
D=max(D_plus,D_mins)
D



#Da=0.565  (using KS Table)   0.34< 0.565
#D < D a
#we do not reject Ho
#Conclusion:  Random numbers are not uniform.
```

OutPut:

Console Terminal Jobs

Project: (Nor

```
~/
> #Example 2:
>
> #Ho: Random no. are not uniform   V/S   H1: Not Ho
> N=5
> i=c(1:5)
> Ri=c(0.54, 0.73, 0.98, 0.11, 0.68)
> Ri=sort(Ri)
> D_plus=(i/N-Ri)
> D_mins=(Ri-(i-1)/N)
> Table2=data.frame(i,Ri,D_plus,D_mins)
> Table2
  i   Ri D_plus D_mins
1 1 0.11   0.09   0.11
2 2 0.54  -0.14   0.34
3 3 0.68  -0.08   0.28
4 4 0.73   0.07   0.13
5 5 0.98   0.02   0.18
> max(D_plus)
[1] 0.09
> max(D_mins)
[1] 0.34
> D=max(D_plus,D_mins)
> D
[1] 0.34
>
>
>
> #Da=0.565  (using KS Table)   0.34< 0.565
> #D < D a
> #we do not reject Ho
> #Conclusion:  Random numbers are not uniform.
>
```

Example 3:

Each person in random sample say n =10 employees was asked about X, the daily time wasted at work doing non work activities, such as surfing the internet and emailing to friends. The resulting data, in minutes, are as follows:

110,102,98,95,108,105,103,90,92,100

Is it okay assume that these data come from a normal distribution with mean 100 and standard deviation 10?

Solution:

- **H_o : given sample comes from normal distribution with $\mu=100$ and $\sigma=10$ i.e $X \sim N(\mu=100, \sigma=10)$**

- **H_1: given sample does not come from normal distribution with $\mu=100$ and $\sigma=10$ i.e $X\sim$ does not follow normal with $\mu=100$ and $\sigma=10$**

Test statistic:

Under H_o,

$D = Max\,|Fn(Xi)-Fo(Xi)|$

Where

Fn C.D.F of empirical distribution and Fo is C.D.F , $N(\mu=100, \sigma=10)$

Solution:

$F_n(x_i) = i/n = i/10$; $i = 1,2,3,\ldots\ldots 10$

Where observation Xi are written in ascending order

$F_o(Z_i) = P(X \le X_i)$

$= P(x-\mu/\sigma \le X_i-\mu/\sigma)$

$= P(Z \le X_i-100/10)$

where ,

Z is $X-\mu/\sigma \sim N(0,1)$

$= F(X_i-100/10)$

S.r no (i)	Xi	Fn(Xi)	Xi-100/10	Fo(Xi)= F(Xi-100/10)	| Fn(Xi)-Fo(Xi) |
1	90	1/10 =0.1	-1	0.15866	0.05866
2	92	2/10=0.2	-0.8	0.21186	0.01186
3	95	3/10=0.3	-0.5	0.30854	0.00854
4	98	4/10=0.4	-0.2	0.42074	0.02074
5	100	5/10=0.5	0	0.5	0
6	102	6/10=0.6	0.2	0.57926	0.02074
7	103	7/10=0.7	0.3	0.61791	0.08209
8	105	8/10=0.8	0.5	0.69146	0.10854
9	108	9/10=0.9	0.8	0.78814	0.11186
10	110	10/10=1	1	0.84134	0.15866 » Dn

Example 3:

$D = \text{Max} \mid F_n(X_i) - F_o(X_i) \mid$

$= \mathbf{0.15866}$

Decision: As D_n = 0.15866<0.410, accept H_o at 5 % level of significance

Critical region

$= \{ D \mid D > D_n \; ; \alpha = D_{10,0.05} = 0.410 \}$ at α% l.o.s

Conclusion: Given sample may have come from normal distribution with parameter(μ=100 , σ=10)

Solution Using R Software(Codes):

```r

#-----------------------------------------------------------------------------------------------------------#
#Example 3


#H??? : given sample comes from normal distribution with µ=100 and s=10 i.e  X~N (µ=100 , s=10)
v/s
#H1: given sample does not come from normal distribution with  µ=100 and s=10

xi=c(110,102,98,95,108,105,103,90,92,100)
x_bar=100
sd=10
N=10

xi_new=sort(xi)
i=c(1,2,3,4,5,6,7,8,9,10)
Fn_xi=i/N
z=(xi_new-x_bar)/sd
F0_xi=pnorm(z)
dc=abs(Fn_xi-F0_xi)
dc
Table3=data.frame(xi_new,Fn_xi,z,F0_xi,dc)
table
max(dc)
#Decision: As  Dn = 0.15866<0.410,  Dn is calculated form Table a=D10·0·05=0.410
#accept H??? at 5 % level of significance.
#Conclusion: given sample may have come from normal distribution with parameter(µ=100 , s=10)

#---------------------------------------------------------------------------------------------------------#
```

OutPut:

```
> xi_new=sort(xi)
> i=c(1,2,3,4,5,6,7,8,9,10)
> Fn_xi=i/N
> z=(xi_new-x_bar)/sd
> F0_xi=pnorm(z)
> dc=abs(Fn_xi-F0_xi)
> dc
 [1] 0.058655254 0.011855399 0.008537539 0.020740291 0.000000000 0.020740291 0.082088578 0.108537539
 [9] 0.111855399 0.158655254
> Table3=data.frame(xi_new,Fn_xi,z,F0_xi,dc)
> table
   xi_new Fn_xi    z     F0_xi          dc
1      90   0.1 -1.0 0.1586553 0.058655254
2      92   0.2 -0.8 0.2118554 0.011855399
3      95   0.3 -0.5 0.3085375 0.008537539
4      98   0.4 -0.2 0.4207403 0.020740291
5     100   0.5  0.0 0.5000000 0.000000000
6     102   0.6  0.2 0.5792597 0.020740291
7     103   0.7  0.3 0.6179114 0.082088578
8     105   0.8  0.5 0.6914625 0.108537539
9     108   0.9  0.8 0.7881446 0.111855399
10    110   1.0  1.0 0.8413447 0.158655254
> max(dc)
[1] 0.1586553
> #Decision: As  Dn = 0.15866<0.410,  Dn is calculated form Table a=D10□0·05=0.410
> #accept H??? at 5 % level of significance.
> #Conclusion: given sample may have come from normal distribution with parameter(μ=100 , s=10)
>
```

Mann–Whitney U test

A Brief Guide to Mann-Whitney U Test

MANN-WHITNEY U TEST

- In statistics, the Mann–Whitney *U* test is a nonparametric test of the null hypothesis that, for randomly selected values *X* and *Y* from two populations, the probability of *X* being greater than *Y* is equal to the probability of *Y* being greater than *X*.
- A similar nonparametric test used on *dependent* samples is the Wilcoxon signed-rank test.

WHEN TO USE MANN-WHITNEY U TEST??

- The Mann-Whitney U test is used to compare whether there is a difference in the dependent variable for two independent groups.

- It compares whether the distribution of the dependent variable is the same for the two groups and therefore from the same population.

What Is Mann-Whitney U Test?

- The Mann-Whitney U test is a non-parametric test that can be used in place of an unpaired t-test. It is used to test the null hypothesis that two samples come from the same population (i.e. have the same median) or, alternatively, whether observations in one sample tend to be larger than observations in the other.

- The Mann-Whitney U Test is a statistical test used to determine if 2 groups are significantly different from each other on your variable of interest.

Tip:

Test use to compare the difference in the dependent variable for two independent groups.

Formula:

$$U_1 = R_1 - \frac{n_1(n_1+1)}{2}$$

or

$$U_2 = R_2 - \frac{n_2(n_2+1)}{2}$$

Either of the two formulas are valid for the Mann Whitney U Test.

Where,
R : sum of ranks in the sample
n : number of items in the sample.

Decision Criteria

➔ **Acceptance Criteria**

if the observed value of U exceeds the critical value we do not reject H_0.

➔ **Rejection Criteria**

if the observed value of U is less than or equal to the critical value, we reject H_0 in favor of H_1

Example 1:

We want to measure the reaction time of male and female and check whether there is difference in both the group.

Gender	F	F	F	F	F	F	M	M	M	M	M
Reaction time	34	36	41	43	44	37	45	33	35	39	42

Hypothesis:

H0: there is no significant difference between reaction time of male and female.

H1: not H0.

Solution:

Gender	F	F	F	F	F	F	M	M	M	M	M
Reaction time	34	36	41	43	44	37	45	33	35	39	42
Rank	2	7	4	9	10	5	11	1	3	6	8

Test statistic:

n1= 6; n2=5

R1: It is the sum of rank of female

2+4+7+9+10+5= 37

R2: It is the sum of rank of male

11+1+3+6+8=29

Example 1:

$U1 = n1n2 + [n1(n1+1)]/2 - \sum R1$

$\quad = 6*5+[6(6+1)]/2 - 37$

$U1 \quad = 14$

$U2 = n1n2+[n2(n2-1)]/2 - \sum R2$

$\quad = 6*5+[5(5+1)]/2-29$

$U2 \quad = 16$

$U = \min(U1, U2)$

$\quad = 14$

Decision:

$Utab = U_{n1,n2,\alpha}$

$= U_{6,5,0.05}$

$= 3$

Since Ucal>Utab

Conclusion: Therefore we reject null hypothesis and conclude that there is significant difference in reaction time of male and female

Solution Using R Software(Codes):

```
#Ex 1

#H0: there is no significant difference between reaction time of male and female.
#H1: Not H0


Gender=c('Female','Female','Female','Female','Female','Female','Male','Male','Male','Male','Male')
Reaction_time=c(34,36,41,43,44,37,45,33,35,39,42)
Data=data.frame(Gender,Reaction_time)
View(Data)
str(Data)
wilcox.test(Reaction_time~Gender,data = Data,mu=0,alt="two.sided",conf.int=T,conf.level=0.95,exact=F)

#DC:Null Hypothesis is  rejected
#There is Significant difference between recation time of Male and Female
```

OutPut:

```
> #H0: there is no significant difference between reaction time of male and female.
> #H1: Not H0
>
>
> Gender=c('Female','Female','Female','Female','Female','Female','Male','Male','Male','Male','Male')
> Reaction_time=c(34,36,41,43,44,37,45,33,35,39,42)
> Data=data.frame(Gender,Reaction_time)
> View(Data)
> str(Data)
'data.frame':   11 obs. of  2 variables:
 $ Gender       : chr  "Female" "Female" "Female" "Female" ...
 $ Reaction_time: num  34 36 41 43 44 37 45 33 35 39 ...
> wilcox.test(Reaction_time~Gender,data = Data,mu=0,alt="two.sided",conf.int=T,conf.level=0.95,exact=F)

        Wilcoxon rank sum test with continuity correction

data:  Reaction_time by Gender
W = 16, p-value = 0.9273
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -7.999979  8.000019
sample estimates:
difference in location
             0.9999627

>
> #DC:Null Hypothesis is not rejected
> #There is no Significant difference between recation time of Male and Female
> 
```

Example 2:

The effectiveness of advertising for two rival products (Brand X and Brand Y) was compared. Market research at a local shopping centre was carried out, with the participants being shown adverts for two rival brands of coffee, which they then rated on the overall likelihood of them buying the product (out of 10, with 10 being "definitely going to buy the product"). Half of the participants gave ratings for one of the products, the other half gave ratings for the other product.

Brand X Rating	3	4	2	6	2	5
Brand Y Rating	9	7	5	10	6	8

Solution:

Brand	X	X	X	X	X	X	Y	Y	Y	Y	Y	Y
Rating	3	4	2	6	2	5	9	7	5	10	6	8
Rank	3	4	1.5	7.5	1.5	5.5	11	9	5.5	12	7.5	10

Test statistic:

- Add up the ranks for Brand X, to get T1

Therefore T1= 3 + 4 + 1.5 + 7.5 + 1.5 + 5.5 = 23

- Add up the ranks for Brand Y, to get T2

Therefore, T2 = 11 + 9 + 5.5 + 12 + 7.5 + 10 = 55

Example 2:

Find U

(Note: Tx is the larger rank total)

In our case it's T2

$$U = n1 \times n2 + nx \times \frac{(nx+1)}{2} - Tx$$

$$U = 6 \times 6 + 6 \times \frac{(6+1)}{2} - 55$$

U=2

Decision:

Utab= Un1,n2,α

= $U_{6,6,0.05}$

= 5

Since Ucal<Utab

Conclusion: We can say that there is a highly significant difference between the ratings given to each brand in terms of the likelihood of buying the product.

Solution Using R Software(Codes):

```r

#H0:- Brand X and Brand Y are not significant (equal)
#H1: Brand X and Brand Y are significant (not equal)


Brand=c(rep("X",6),rep("Y",6))
Brand
Rating=c(3,4,2,6,2,5,9,7,5,10,6,8)
BData=data.frame(Brand,Rating)
head(BData)
wilcox.test(Rating~Brand,data =BData,mu=0,alt="two.sided",conf.int=T,conf.level=0.95,exact=F)
#DC:  We can say that there is a highly significant difference between the ratings
# given to each brand in terms of the likelihood of buying the product.

```

OutPut:

```
> #Ex2:
>
> #H0:- Brand X and Brand Y are not significant (equal)
> #H1: Brand X and Brand Y are significant (not equal)
>
>
> Brand=c(rep("X",6),rep("Y",6))
> Brand
 [1] "X" "X" "X" "X" "X" "X" "Y" "Y" "Y" "Y" "Y" "Y"
> Rating=c(3,4,2,6,2,5,9,7,5,10,6,8)
> BData=data.frame(Brand,Rating)
> head(BData)
  Brand Rating
1     X      3
2     X      4
3     X      2
4     X      6
5     X      2
6     X      5
> wilcox.test(Rating~Brand,data =BData,mu=0,alt="two.sided",conf.int=T,conf.level=0.95,exact=F)

        Wilcoxon rank sum test with continuity correction

data:  Rating by Brand
W = 2, p-value = 0.01259
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 -6.000001 -1.000006
sample estimates:
difference in location
             -3.999954

> #DC:  We can say that there is a highly significant difference between the ratings
> # given to each brand in terms of the likelihood of buying the product.
>
```

Example 3:

With the Help of Mann Whitney U test find if significant difference exist between the score obtained on organizational commitment scale obtain by public and private bank employees (alpha=0.05)

Public Bank employee	Rank R1	Private Bank employee	Rank R2
10	1	34	16
12	2	54	22
21	3	56	23
23	5.5	43	19
34	16	32	10
45	21	23	5.5
32	10	34	16
23	5.5	32	10
34	16	33	13
23	5.5	44	20
		32	10
		34	16
		32	10

Solution:

Hypothesis:

H0: There is no significant difference between public bank and private bank employee

H1: Not H0.

Test statistic:

$n_1=10 \quad n_2=13$

$u_1=n_1n_2+(n_1)(n_1+1)/2-\sum R_1 =99.5$
$u_2=n_1n_2+(n_2)(n_2+1)/2-\sum R_2 =30.5$

Example 2:

U=min(U1,U2)=30.5

U=30.5

Decision:
UFrom U table
U-critical=33
Ucal<Utab

Conclusion: Null Hypothesis is rejected
There is Significant difference between the public bank and Private bank employee

Solution Using R Software(Codes):

```
#Ex3
#H0: There is no significant difference between public bank and private bank employee
#H1: Not H0.


Bank=c(rep("PublicBank",10),rep("PrivateBank",13))
score=c(10,12,21,23,34,45,32,23,34,23,34,54,56,43,32,23,34,32,33,44,32,34,32)
SData=data.frame(Bank,score)
head(SData)
wilcox.test(score~Bank,data =SData,mu=0,alt="two.sided",conf.int=T,conf.level=0.95,exact=F)

#DC:Null Hypothesis is rejected
#There is Significant difference between the public bank and Private bank employee
```

OutPut:

```
> #Ex3
> #H0: There is no significant difference between public bank and private bank employee
> #H1: Not H0.
>
>
> Bank=c(rep("PublicBank",10),rep("PrivateBank",13))
> score=c(10,12,21,23,34,45,32,23,34,23,34,54,56,43,32,23,34,32,33,44,32,34,32)
> SData=data.frame(Bank,score)
> head(SData)
        Bank score
1 PublicBank    10
2 PublicBank    12
3 PublicBank    21
4 PublicBank    23
5 PublicBank    34
6 PublicBank    45
> wilcox.test(score~Bank,data =SData,mu=0,alt="two.sided",conf.int=T,conf.level=0.95,exact=F)

        Wilcoxon rank sum test with continuity correction

data:  score by Bank
W = 99.5, p-value = 0.03275
alternative hypothesis: true location shift is not equal to 0
95 percent confidence interval:
 5.808113e-05 2.100002e+01
sample estimates:
difference in location
                    11


>
> #DC:Null Hypothesis is rejected
> #There is Significant difference between the public bank and Private bank employee
>
```

Presented By:

- Zeenat(ppt & explanation)
- Hunain(example 1,2)
- Jasmin(example 2,3)
- Sidharth(example 3,1)
- Zishan(R codes)

Get All R Codes At:

https://github.com/ZishanSayyed/KS-ONE-SAMPLE-TEST-MANN-WHITNEY-U-TEST