CSE 544: Probability & Statistics for Data Science,
Spring 2021

Assignment - 5

Submitted by

Harmanpreet Singh Khurana, SBU ID: 113262379,
Mayank Jain, SBU ID: 113263864,
Rajadorai DS, SBU ID: 113259773,
Venkatesan Ravi, SBU ID: 113263484

Assignment 5

1) Hypothesis testing for a single population:

$D = \{0.04, 0.74, 0.84, 1.19, 1.88, 1.99, 2.23, 2.57, 2.65, 2.78\}$

$Y = Unif(0,3)$, Critical threshold, $\alpha = 0.25$

$H_0 : F_D \equiv F_Y \qquad H_1 : F_D \neq F_Y$

$$F_{unif(a,b)}\, d = \frac{d-a}{b-a} = \frac{d-0}{3-0} = \frac{d}{3}$$

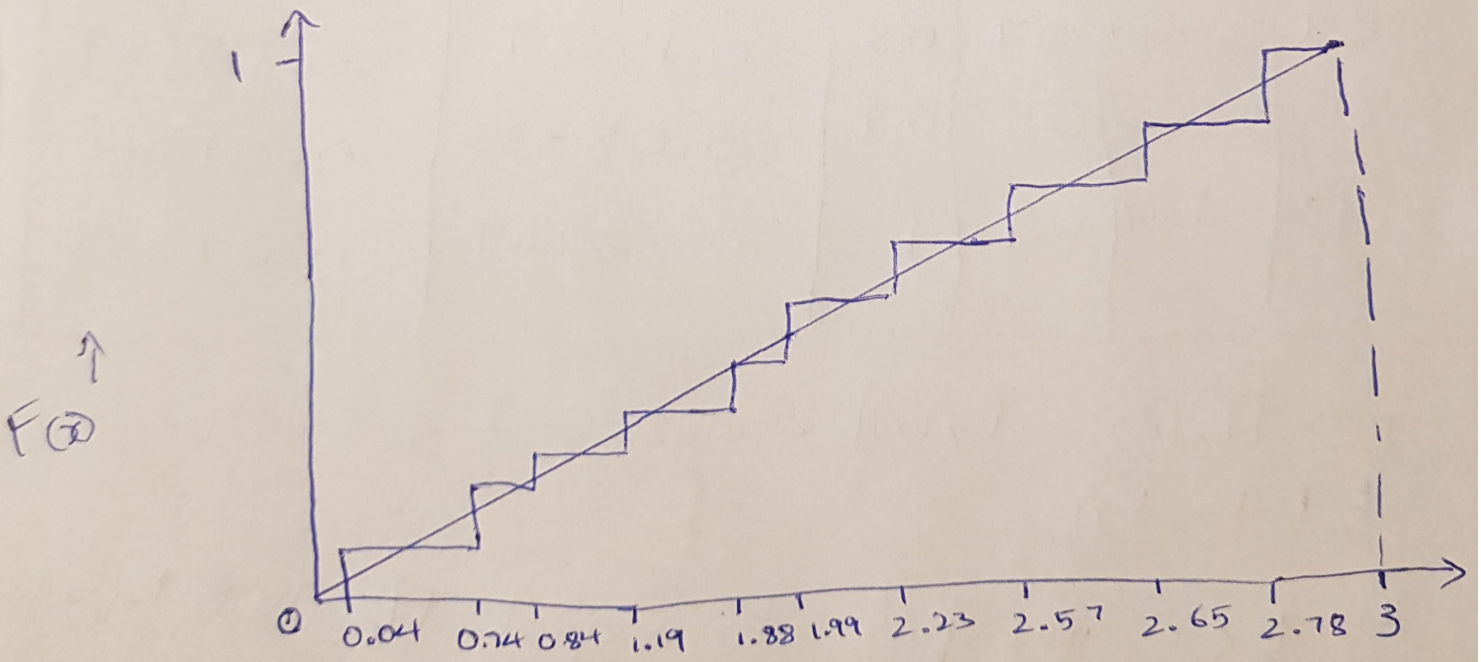


Table for K-S Test:

x	$F_Y(x)$	$\hat{F}_x^-$	$\hat{F}_x^+$	$\lvert\hat{F}_x^- - F_Y(x)\rvert$	$\lvert\hat{F}_x^+ - F_Y(x)\rvert$
0.04	0.01333	0	0.1	0.01333	0.08667
0.74	0.24667	0.1	0.2	0.14667	0.04667
0.84	0.28	0.2	0.3	0.08	0.02
1.19	0.39667	0.3	0.4	0.09667	0.00333
1.88	0.62667	0.4	0.5	0.22667	0.12667
1.99	0.66333	0.5	0.6	0.16333	0.06333
2.23	0.74333	0.6	0.7	0.14333	0.04333
2.57	0.85667	0.7	0.8	0.15667	0.05667
2.65	0.88333	0.8	0.9	0.08333	0.01667
2.78	0.92667	0.9	1	0.02667	0.07333

$\max \lvert\hat{F}_x(x) - F_Y(x)\rvert = 0.22667 < 0.25$ (c)

↳ $D(F_x, F_Y)$

As $D < C$, we fail to reject H_0 (Null hypothesis)

2) Toy example for permutation test:

$X = \{5\}$, $Y = \{2, 7\}$

p-value threshold = 0.05

$H_0 : X = Y$, $H_1 : X \neq Y$

Total number of samples, $N = |X| + |Y|$

$= 1 + 2 = 3$

$$T_{obs} = |\bar{X} - \bar{Y}| = \left|5 - \frac{2+7}{2}\right| = 0.5$$

Step 1: Permute $X \cup Y$ in all $N! = 3!$ ways

$= 6$ ways

To get x_i of size $|X|$ and

Y_i of size $|Y|$

Step 2: Compute $\frac{1}{N!} \sum_{i=1}^{N!} I(T_i > T_{obs}) = \frac{1}{6} \sum_{i=1}^{6} I(T_i > T_{obs})$

Reject if computed value < p-threshold

(given = 0.05)

i	X_i	Y_i	$\bar{X}_i$	$\bar{Y}_i$	$T_i = \lvert \bar{X}_i - \bar{Y}_i \rvert$	$I(T_i > T_{obs})$
1	{5}	{2,7}	5	4.5	0.5	0
2	{5}	{7,2}	5	4.5	0.5	0
3	{2}	{7,5}	2	6	4	1
4	{2}	{5,7}	2	6	4	1
5	{7}	{5,2}	7	3.5	3.5	1
6	{7}	{2,5}	7	3.5	3.5	1

(From Step 2.)

$$\frac{1}{6}\sum_{i=1}^{6} I(T_i > T_{obs}) = \frac{1}{6}(0+0+1+1+1+1) = \frac{4}{6} = 0.6667$$

$0.6667 > 0.05$ (p-threshold value)

∴ We fail to reject H_0 (Null Hypothesis).

Q3) (a) NULL: Dealer I Outcome of Table

	Dealer A	Dealer B	Dealer C	Total
Win	$O_{11}=48$, E_{11}	$O_{12}=54$, E_{12}	$O_{13}=19$, E_{13}	$T_{14}=121$
Draw	$O_{21}=7$, E_{21}	$O_{22}=5$, E_{22}	$O_{23}=4$, E_{23}	$T_{24}=16$
Loose	$O_{31}=55$, E_{31}	$O_{32}=50$, E_{32}	$O_{33}=25$, E_{33}	$T_{34}=130$
Total	$T_{41}=110$	$T_{42}=109$	$T_{43}=48$	$T_{44}=267$

Step I → To find Q_{obs}

$$Q_{obs} = \sum_{row} \sum_{col.} \frac{(E_{rc} - O_{rc})^2}{E_{rc}}$$

Now, in the question we are given all the observed values i.e. O_{rc}.

We need to find all the E_{rc} values.

E_{11} = Expected value when Win occurs and Dealer A is at table.

$$E_{11} = \frac{T_{14} \times T_{41}}{T_{44}} = \frac{121}{267} \times 110 = 49.85$$

$$E_{12} = \frac{T_{14}}{T_{44}} \times T_{42} = \frac{121}{267} \times 109 = 49.39$$

Similarly, $E_{13} = 21.75$

$E_{21} = 6.59$, $E_{22} = 6.53$, $E_{23} = 2.88$

$E_{31} = 53.56$, $E_{32} = 53.07$, $E_{33} = 23.37$

$$\therefore Q_{obs} = \frac{(E_{11}-O_{11})^2}{E_{11}} + \frac{(E_{12}-O_{12})^2}{E_{12}} + \frac{(E_{13}-O_{13})^2}{E_{13}} +$$

$$\frac{(E_{21}-O_{21})^2}{E_{21}} + \frac{(E_{22}-O_{22})^2}{E_{22}} + \frac{(E_{23}-O_{23})^2}{E_{23}} +$$

$$\frac{(E_{31}-O_{31})^2}{E_{31}} + \frac{(E_{32}-O_{32})^2}{E_{32}} + \frac{(E_{33}-O_{33})^2}{E_{33}}$$

$$= 0.068 + 0.430 + 0.348 +$$
$$0.025 + 0.358 + 0.436 +$$
$$0.039 + 0.178 + 0.113$$

$$Q_{obs} = 1.995$$

degree of freedom $= (3-1)(3-1) = \underline{4}$

$$\text{P-value} = Pr(X^2_{df} > Q_{obs})$$
$$= 1 - Pr(X^2_{df} < Q_{obs})$$
$$= 1 - Pr(X^2_4 < 1.995)$$
$$= 1 - 0.2633$$

$\Rightarrow$ P-value $= 0.7367$

Given, $\alpha = 0.05$

As, P-value $>$ α $\therefore$ we can accept the NULL.

(b) Pearson correlation co-efficient

$$\hat{\rho}_{XY} = \frac{\sum_{i=1}^{n} \{(X_i - \bar{X})(Y_i - \bar{Y})\}}{\sqrt{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)\left(\sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)}}$$

We are given data of 3 dealers, Dealer A, Dealer B and Dealer C.

We need to calculate ρ_{AB}, ρ_{BC} and ρ_{CA}

mean of $A = \bar{A} = 45$

mean of $B = \bar{B} = 45.2$

mean of $C = \bar{C} = 32.6$

$$\hat{\rho}_{AB} = \frac{\sum_{i=1}^{10} \{(A_i - \bar{A})(B_i - \bar{B})\}}{\sqrt{\left(\sum_{i=1}^{10}(A_i - \bar{A})^2\right)\left(\sum_{i=1}^{10}(B_i - \bar{B})^2\right)}}$$

$$= \frac{(26.4 + (-14) + 75.4 + 14.4 + 336 + 204 + 173.6 + 277.2 + 303 + 0)}{\sqrt{1462}\ \sqrt{2193.6}}$$

$$= 0.779$$

$\Rightarrow \hat{\rho}_{AB} > 0.5 \Rightarrow$ No. of wins of Dealer A and B are Positively correlated.

Similarly,

$$\hat{\rho}_{BC} = \frac{-96.2}{\sqrt{2193.6}\sqrt{694.4}} = -0.0779$$

$$\Rightarrow |\hat{\rho}_{BC}| = 0.0779$$

$|\hat{\rho}_{BC}| < 0.5 \Rightarrow$ No correlation is there among Dealer A and Dealer B wins.

$$\hat{\rho}_{AC} = \frac{22}{1007.58} = 0.022$$

$\Rightarrow |\hat{\rho}_{AC}| < 0.5 \Rightarrow$ No correlation among A and C wins

The conclusion that we can get from the above observed correlation coefficients among the dealers is that as the Dealer C is not correlated to Dealer A and Dealer B ∴ The Dealer C can be said to causing Loss to the owner of casino, as ideally they all should be correlated because win probality for every game is equal.

4)

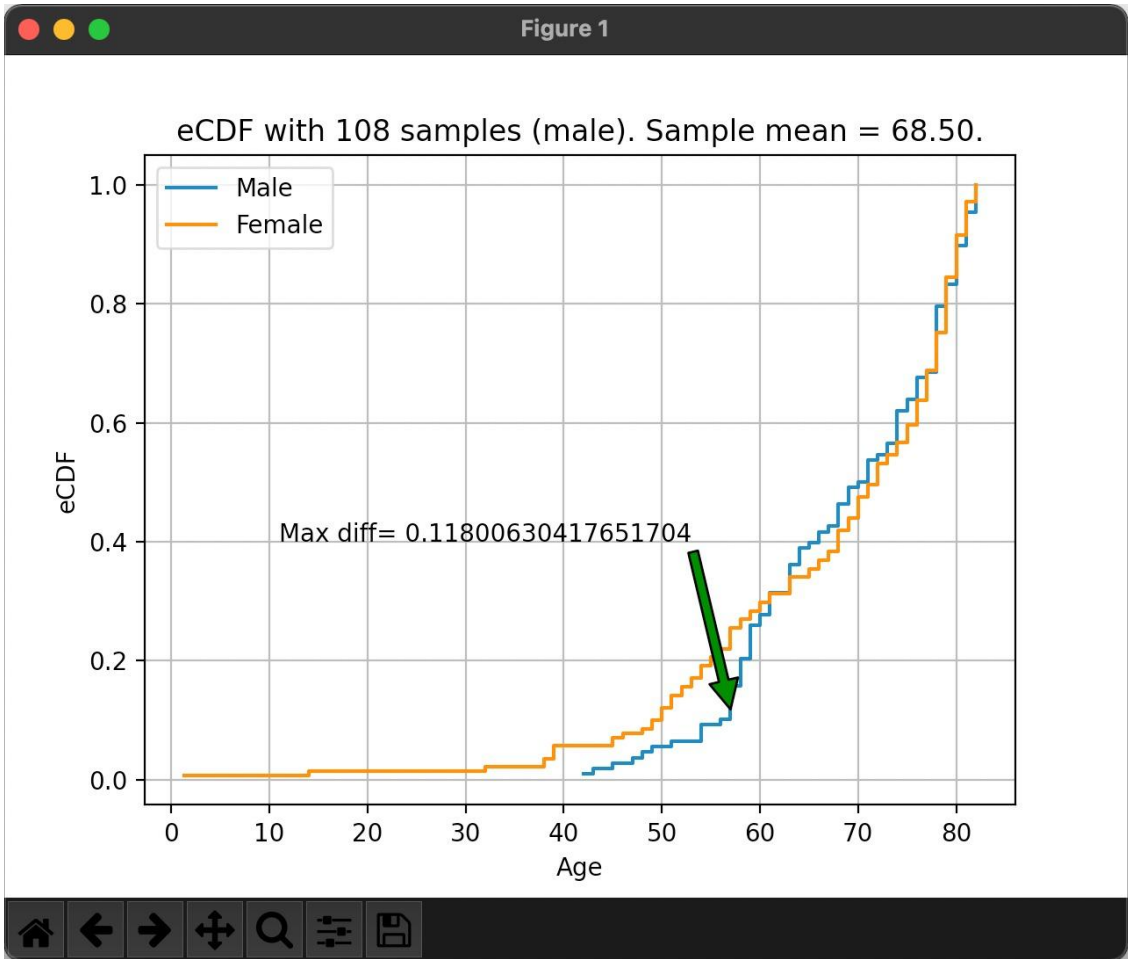
a)

```
harman@harmans-MacBook-Pro Hw5 % python3 q4a.py
observed_T=  21.584579138471682
alpha =  0.05
For n = 200 random permutations, p_value:  0.0
Therefore, NULL hypothesis for  200 permutations can be rejected as p-value is less than alpha
For n = 1000 random permutations, p_value:  0.0
Therefore, NULL hypothesis for  1000 permutations can be rejected as p-value is less than alpha
```

b)

```
harman@harmans-MacBook-Pro Hw5 % python3 q4b.py
mean of male age data:  68.5
mean of female age data:  67.13702127659575
observed_T=  1.3629787234042539
alpha =  0.05
For n = 1000 random permutations, p_value:  0.422
Therefore, NULL hypothesis for  1000 permutations can be accepted as p-value is more than alpha
```

c)



5) (a) $\{X_1, X_2 \cdots X_n\} \sim Nor(\mu_1, \sigma_1^2) \Rightarrow (i)$

$\{Y_1, Y_2, \cdots Y_m\} \sim Nor(\mu_2, \sigma_2^2) \Rightarrow (ii)$

$H_0 : \mu_1 > \mu_2$ vs $H_1 : \mu_1 \leq = \mu_2$

$\downarrow$ $\downarrow$

$H_0 : \mu_1 - \mu_2 > 0$ vs $H_1 : \mu_1 - \mu_2 \leq 0$

T statistic for unpaired t-test, $T = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$

where $\bar{D} = \bar{X} - \bar{Y}$, $\leftarrow$ $= \dfrac{\bar{D}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$

S_x : sample standard deviation for $\{X_1 \cdots X_n\}$

S_y : sample standard deviation for $\{Y_1 \cdots Y_m\}$

For one sides test with $H_0 : \mu_1 > \mu_2$, we have : Reject H_0 if $T < -\delta$

Type I error : Pr (Test rejects H_0 | H_0 is true)

$= Pr(T < -\delta \mid H_0 \text{ is true})$

$= Pr\left(\dfrac{\bar{D} - 0}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} < -\delta\right)$

Denominator is positive, hence, multiplying both sides by $\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$

Type I error : $Pr\left(\bar{D} < -\delta\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}\right)$ → (iii)

Now we know, $\bar{D} = \bar{X} - \bar{Y} = \bar{X} + (-1)\bar{Y}$

$\sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$

Transforming D to a standard normal to get the result in terms of Φ, we have,

$$Z = \frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}$$

(Using transformation property)

↓

Equation II (seen in (b) ⇒ p-value calculation)

As n & m are large by our assumption, sample standard deviation is the same as actual standard deviation

$$\therefore Z = \frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} = \frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \rightarrow (iv)$$

From (3),

$$\text{Type 1 error} = Pr\left(\bar{D} < -\delta\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}\right)$$

→ Subtracting $(\mu_1 - \mu_2)$ from both sides, ↦ (Step A)

$$= Pr\left(\bar{D} - (\mu_1 - \mu_2) < -\delta\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}} - (\mu_1 - \mu_2)\right)$$

→ Dividing both sides by $\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}$ ↦ (Step B) (used again for type II error)

$$= Pr\left(\frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} < -\delta - \left(\frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)\right)$$

$$= Pr\left(z < -\delta - \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

As $\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \sim Nor(0, 1)$,

$$\text{Type-I error} = \phi\left(-\delta - \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

Hence, Proved.

Type II error:

Pr (Accept H_0 | H_0 is false),

$T \geq -\delta \rightarrow$ Accept H_0.

$$= Pr\left(\frac{\bar{D}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \geq -\delta\right)$$

$$= Pr\left(\bar{D} \geq -\delta\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}\right)$$

Using step A, Step B from Type 1 error calculation,

$$= Pr\left(\frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \geq -\delta - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

$$= 1 - Pr\left(Z < -\delta - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

As stated earlier, assuming n & m are large, Sample standard deviation is the same as actual Standard deviation here.

$$= 1 - \phi\left(-\delta - \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

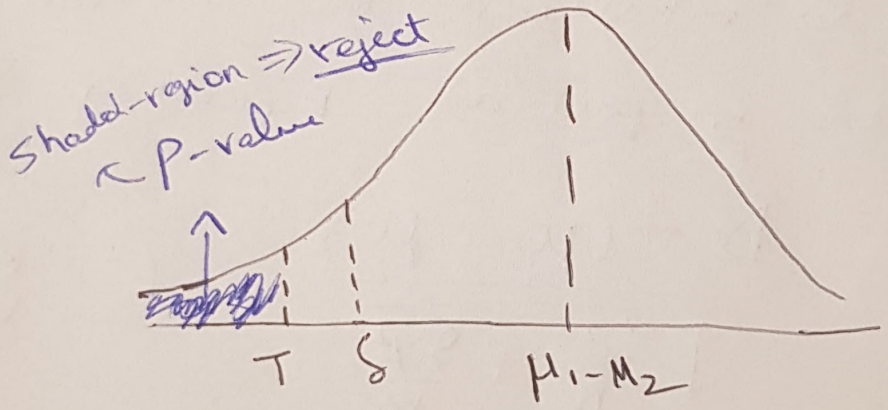
$\therefore$ Type 2 error : $1 - \Phi\left(-\delta - \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$

Hence, Proved.

(b) As stated in (a),

For the unpaired test we get the T statistic, $T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$

If we reject H_0, then $T < -\delta$.



-val
$= P(T < t_{obs})$

We will reject H_0 if T_{obs} falls in the shaded region in the graph shown.

P-value = Area to the left of $T_{obs} = \phi(T_{obs})$

The graph shown above is of the standard normal distribution and we had already computed in part (a) ⇒ Equation V

$$\hookrightarrow Z = \frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

(∵ n & m are large, ∴ sample standard deviation is same as true standard deviation)

$$\text{P-value} = \Phi(z) = \phi\left(\frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

$$\bar{D} = \bar{X} - \bar{Y}$$

$$\therefore \text{P-value} = \phi\left(\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

Hence, Proved.

Q6)

Given,

$X \sim N(1.5,\ 1)$

$X \sim N(1,\ 1)$

$\therefore$ We have, true variance of $X\ =\ 1$, true variance of $Y\ =\ 1$

We have

$$var(\overline{D})\ =\ var(\overline{X}\ -\ \overline{Y})$$

$$By\ LOV,\ var(\overline{D})\ =\ var(\overline{X})\ + var(\overline{Y})$$

$$var(\overline{D})\ =\ \frac{\sigma_1^2}{len(X)}\ +\ \frac{\sigma_2^2}{len(Y)} \tag{1}$$

Whenever we use, Z test, we will use true variance of the samples.

Since true variance of $X\ =\ 1$, true variance of $Y\ =\ 1$, $len(X)\ =\ len(Y)$

$$var(\overline{D})\ =\ \frac{1}{len(X)}\ +\ \frac{1}{len(Y)}\ =\ \frac{2}{len(X)}$$

$$\therefore\ std(\overline{D})\ =\ \sqrt{\frac{2}{len(X)}}$$

$$\therefore z\ statistic\ =\ \frac{(\overline{X})\ -\ (\overline{Y})}{std(\overline{D})}\ =\ \frac{(\overline{X})\ -\ (\overline{Y})}{\sqrt{\frac{2}{len(X)}}} \tag{2}$$

Whenever we use, T test, we will use sample variance of the samples

Since sample variance of $X = \frac{\sum_{i=1}^{n}(x_i - \overline{X})}{len(X) - 1}$, sample variance of $Y = \frac{\sum_{i=1}^{n}(y_i - \overline{Y})}{len(Y) - 1}$,

$$var(\overline{D}) = \frac{\frac{\sum_{i=1}^{n}(x_i - \overline{X})}{len(X) - 1}}{len(X)} + \frac{\frac{\sum_{i=1}^{n}(y_i - \overline{Y})}{len(Y) - 1}}{len(Y)}$$

$$\therefore std(\overline{D}) = \sqrt{\left[\frac{\frac{\sum_{i=1}^{n}(x_i - \overline{X})}{len(X) - 1}}{len(X)}\right] + \left[\frac{\frac{\sum_{i=1}^{n}(y_i - \overline{Y})}{len(Y) - 1}}{len(Y)}\right]}$$

$$\therefore t\ statistic = \frac{(\overline{X}) - (\overline{Y})}{std(\overline{D})} = \frac{(\overline{X}) - (\overline{Y})}{\sqrt{\left[\frac{\frac{\sum_{i=1}^{n}(x_i - \overline{X})}{len(X) - 1}}{len(X)}\right] + \left[\frac{\frac{\sum_{i=1}^{n}(y_i - \overline{Y})}{len(Y) - 1}}{len(Y)}\right]}} \quad (3)$$

6a)

We will use $X = X_1, Y = Y_1$

$\overline{X_1} = 1.59873856$

$\overline{Y_1} = 1.06250437$

$H_0 : Means\ are\ equal.\ i.e,\ \overline{D} = 0$
$H_1 : Means\ are\ unequal.\ i.e,\ \overline{D} \neq 0$

$Using\ (2),\ z\ statistic = 1.69572141$

$p\ calculated\ from\ table = 0.08993865$

$We\ will\ accept\ H_0.$

$Using\ (3),\ t\ statistic = 1.57496970$

$p\ calculated\ from\ table = 0.13176797$

$We\ will\ accept\ H_0.$

6b)

We will use $X = X_2, Y = Y_2$

$\overline{X_1} = 1.46125898$

$\overline{Y_1} = 0.98352020$

$H_0 : Means\ are\ equal.\ i.e,\ \overline{D} = 0$
$H_1 : Means\ are\ unequal.\ i.e,\ \overline{D} \neq 0$

$Using\ (2),\ z\ statistic = 10.68256398$

$p\ calculated\ from\ table = 0.0$

$We\ will\ not\ accept\ H_0.$

$Using\ (3),\ t\ statistic = 10.71342881$

$p\ calculated\ from\ table = 0.0$

$We\ will\ not\ accept\ H_0.$

Is there a significant advantage of using Z-test in a small sample?

No, Not for small samples

Reason:

As we see from the outputs, there is no much change in the t statistic and z statistic in both the cases. So we can say there is no much difference when input sizes are small. This is true because Z test works better when ($n \geqslant 30$) as CLT will be valid only from then.