1) KS Test

Given $p(x) = 2x$
Let this target distribution be Y.
Given threshold C = 0.25

Given Samples X = {0.592, 0.774, 0.245, 0.424, 0.685, 0.436, 0.648, 0.959, 0.842, 0.995}.

Null Hypothesis H_0: The given samples X follow the distribution Y.
Alternate Hypothesis H_1: The given samples X do not follow the distribution Y.

CDF of Y = $F_Y(X) = Pr(X<x)$

$$= \int_0^x p(x)\,dx$$

$$= \int_0^x 2x\,dx$$

$$= 2(x^2 / 2)$$

$$= x^2$$

X	$F_Y(X) = X^2$	$F^{\wedge -}(X)$	$F^{\wedge +}(X)$	$\lvert F(X) - F^{\wedge -}(X) \rvert$	$\lvert F(X) - F^{\wedge +}(X) \rvert$
0.245	0.06	0	0.1	0.06	0.04
0.424	0.179	0.1	0.2	0.079	0.02
0.436	0.19	0.2	0.3	0.01	0.11
0.592	0.35	0.3	0.4	0.05	0.05
0.648	0.42	0.4	0.5	0.02	0.08
0.685	0.47	0.5	0.6	0.03	0.13
0.774	0.59	0.6	0.7	0.01	0.11
0.842	0.709	0.7	0.8	0.0089	0.091
0.959	0.92	0.8	0.9	0.12	0.02
0.995	0.99	0.9	1	0.1	0.01

$\text{Max}\ (F_Y(X) - F^{\wedge}(X)) = 0.13$

$\text{Max}\ (F_Y(X) - F^{\wedge}(X)) < C$

Therefore, the null hypothesis is accepted.

2) Permutation Test

Should we consider/not consider the repeated permutations while performing the permutations test?

Actually, it does not matter if we consider/did not consider the repeated permutations, as the obtained p_value will be same in both scenarios.

How?

Consider a permutation [a,b,a,c]. This permutation will be repeated two times. Let m and n be the lengths of the two samples we consider in this scenario. We have to note that if [a,b,a,c] is repeated two times, symmetrically [c,a,b,a] is also repeated two times. If T is greater than T_{obs} in first case, then T will be less than or equal to T_{obs} is the second case. This is true in all the possible cases of m and n i.e m>n, m<n, m=n.

So, if we remove all the repeated permutations, if one 1 is removed, one 0 will also be removed. So, the final $Pr(T>T_{obs})$ will not change.

Therefore we can conclude that the P_Value is irrespective of the decision that we are including/not including the duplicate permutations. So to reduce the calculational complexity, it is advised not to include the duplicate permutations.

Given Problem

Given X = {3,2} and Y = {2,5}

Let us use the difference of the means of the distributions as the measure for this experiment.

$X_{mean} = 2.5$

$Y_{mean} = 3.5$

Null Hypothesis H_0: The two samples follow similar distributions
Alternate Hypothesis H_1: The two samples do not follow similar distributions.

$T_{obs} = | 2.5 - 3.5 | = 1$

Let us generate all the permutations of these distributions(exclusing the repeated permutations. These repetitions are caused by the element 2, which is repeated twice).

Index	X	Y	X_mean	Y_mean	T = |X_mean - Y_mean |	I (T > T_obs)
0	[3, 2]	[2, 5]	2.5	3.5	1	0
1	[3, 2]	[5, 2]	2.5	3.5	1	0
2	[3, 5]	[2, 2]	4	2	2	1
3	[2, 3]	[2, 5]	2.5	3.5	1	0
4	[2, 3]	[5, 2]	2.5	3.5	1	0
5	[2, 2]	[3, 5]	2	4	2	1
6	[2, 2]	[5, 3]	2	4	2	1
7	[2, 5]	[3, 2]	3.5	2.5	1	0
8	[2, 5]	[2, 3]	3.5	2.5	1	0
9	[5, 3]	[2, 2]	4	2	2	1
10	[5, 2]	[3, 2]	3.5	2.5	1	0
11	[5, 2]	[2, 3]	3.5	2.5	1	0
Total						4

P_Value = 4/12 = 0.33
P_Value >0.05

Therefore the Null Hypothesis is Accepted.

③ Independente Tests to Save Your Boxing Ring

(a)

Null Hypothesis H_0: Outcome of the match should be independent of the judge.

H_1: Outcome of the match is dependent on the judge

	Judge A	Judge B	Judge C	Total
Player 1 Wins	72	50	24	146
Draw	8	5	3	16
Loses	20	8	9	37
Total	100	63	36	199

From the above table, grand total = 199

$$P(\text{Player 1 Wins}) = \frac{\text{Total Wins}}{\text{Grand Total}} = \frac{146}{199} = 0.73$$

$$P(\text{Judge A}) = \frac{\text{Total A}}{\text{Grand Total}} = \frac{100}{267} = 0.374$$

Expected frequency of ~~Win~~ player 1 & Judge A is = Grand Total * P(win) * P(Judge A)

Similarly
$$\Rightarrow E_{ij} = \frac{T_{4j} \times T_{i4}}{T_{44}}$$
row column

Similarly, we populate the below table with expected frequencies:

	Judge A	Judge B	Judge C
Wins	73.36	46.22	26.41
Draw	8.04	5.065	2.89
Player 1 Loses	18.54	11.71	6.693

$$Q_{obs} = \sum_{r}\sum_{c} \frac{(E_{rc} - O_{rc})^2}{E_{rc}}$$

Observed	Excepted	$\frac{(E-O)^2}{E}$
72	73.36	0.0252
50	46.22	0.309
24	26.41	0.2199
8	8.04	0.00099
5	5.065	0.00083
3	2.89	0.0041
20	18.54	0.1069
8	11.71	1.175
9	6.693	0.7951

$Q_{obs} = 2.63702$

df (Degree of Freedom) $= (3-1) * (3-1) = 4$

P-value $= P_r(\chi^2_4 > Q_{obs}) = 1 - P_r(\chi^2_4 < 2.63702)$

$= 0.620283$

As p-value $(0.620283) > 0.05$, we fail to reject H_0

We Accept H_0.

3) b) Pearson Correlation Coefficient

$$\hat{S}_{X,Y} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

Let

$X_i \rightarrow$ number of player1 wins for Judge A for 10 days

$Y_i \rightarrow$ number of player wins for Judge B

$Z_i \rightarrow$ Number of player wins for Judge C.

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} = 42.6 \qquad \bar{Y} = 43 \qquad \bar{Z} = 40.2$$

Correlation between Judges A, B

$$\hat{S}_{\cancel{X},\cancel{Y}} = \frac{\sum_{i=1}^{10}(X_i - 42.6)(Y_i - 43)}{\sqrt{\sum_{i=1}^{10}(X_i-42.6)^2}\sqrt{\sum_{i=1}^{10}(Y_i-43)^2}}$$

$= 0.5726 \ (> 0.5)$ - Positive linear Correlation

Correlation between Judges B, C

$$\hat{S}_{\cancel{Y},Z} = \frac{\sum_{i=1}^{10}(Y_i - 43)(Z_i - 40.2)}{\sqrt{\sum_{i=1}^{10}(Y_i-43)^2}\sqrt{\sum_{i=1}^{n}(Z_i-40.2)^2}} = -0.0837$$

$= |-0.0837| \leq 0.5 \Rightarrow$ ~~No~~ No linear Calculation

Correlation between Judges A, C

$$\hat{S}_{\cancel{X},Z} = \frac{\sum_{i=1}^{10}(X_i - 42.6)(Z_i - 40.2)}{\sqrt{\sum_{i=1}^{10}(X_i-42.6)^2}\sqrt{\sum_{i=1}^{10}(Z_i-40.2)^2}} = 0.0913$$

$\hat{S}_{A,C} \Rightarrow \hat{S}_{X,Z} = 0.0913$

$|\hat{S}_{X,Z}| \leq 0.5 \rightarrow$ No linear Corelation

As the probability of winning each game is same; the results for each Judges should be Correlated.
We observe from the results of Judge C that it is not linearly Correlated with Judge A & Judge B.
From this we can infer that Judge C is not doing their job.

4 a) Step 1: Define Null hypothesis

$H_0 : \mu \leq \mu_0$

$H_1 : \mu > \mu_0$

Given $X = [21.5, 24.5, 18.5, 17.2, 14.5, 23.2, 22.1, 20.5, 19.4, 18.1, 24.1, 18.5]$

$H_0 : \bar{X} \leq 20$

$H_1 : \bar{X} > 20$

$n = 12$ Since this is One Sample T Test, the degrees of freedom $= n-1 = 12-1 = 11$

$\alpha = 0.05$ to meet 95% Confidence interval

Step (2) Calculate Test Statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$$\bar{X} = \frac{x_1 + x_2 + \cdots x_n}{n} = \frac{21.5 + 24.5 + 18.5 + 17.2 + 14.5 + 23.2 + 22.1 + 20.5 + 19.4 + 18.1 + 24.1 + 18.5}{12}$$

$$= 20.175$$

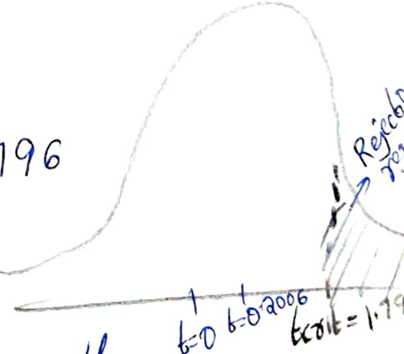
$$S = \sqrt{\frac{\Sigma(X_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1.768 + 18.74 + 2.78 + 8.82 + 32.14 + 9.18 + 3.72 + 0.108 + 0.59 + 4.28 + 15.44 + 2.78}{11}}$$

$$= 3.0211$$

$$T = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{(20.175 - 20)}{3.0211/\sqrt{12}} = 0.2006$$

Step-3 ~~$t_{(11, 0}$~~ If $T > t_{n-1,\alpha}$ reject H_0

$$t_{n-1,\alpha} = t_{11, 0.05} = 1.796$$



$T \not> t_{n-1,\alpha}$

$0.2006 \not> 1.796$ we accept H_0

[$\therefore$ We accept the null hypothesis]

A) b

Step 1: Null & Alternative Hypotheses:

$H_0 : \mu \le \mu_0$

$H_1 : \mu > \mu_0$

[$\mu_0 = 20 \rightarrow$ The standard potato yield ~~from~~ ~~the~~ for the given Variety]

a) ~~Given~~ $\therefore H_0 : \mu \le 20$

$H_1 : \mu > 20$

Given that, ~~the~~ farming Company believes that there is an improvement in the standard potato yield by the introduction of a new fertilizer.

In the above mentioned case sample size ($n = 12$) and their mean yield is 22.

Standard deviation of potato yields is 3.

Step 2: Calculate T-statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$\bar{X} = 22$ (Sample mean given) $\mu_0 = 20$

$n = 12$ and $S = 3$

$$T = \frac{22 - 20}{3/\sqrt{12}} = \frac{2}{\frac{3}{\sqrt{12}}} = 2.3094$$

Step 3: If $T > t_{n-1, \alpha}$ reject H_0

4 Continuation

(b) $t_{n-1, \alpha} = t_{11, 0.05} = 1.796$

The test we have taken is One-tailed test & the Critical value is given by 1.796

$2.3094 > 1.796$ We reject H_0

Step (3) We can use P.value to provide Confidence in the rejection region

$P\text{-Value} = P(T > t)$

$P\text{-value} = P(T > 2.3094)$

Using P.value Calculator, we get

$P\text{-value} = 0.020671$

$P \leq 0.05$

If P-value ≤ 0.05 is Considered as a statistically significant region. Therefore, we reject the null hypothesis.

$\therefore$ We accept Alternative hypothesis saying that there is an improvement in the standard potato yield due to the new fertilizer.

5) Given $D_1 = \{x_1, x_2, ..., x_n\}$ be iid from Normal(μ_1, σ_1^2)

$D_2 = \{y_1, y_2, ..., y_m\}$ be iid from Normal (μ_2, σ_2^2)

x's and y's are independent and $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ are unknown

then the hypothesis is

$H_0: \mu_1 > \mu_2$ $\quad H_1: \mu_1 <= \mu_2$

using unpaired T-test with thresold value $\delta > 0$ to Check hypothesis

$$T = \frac{\bar{D}}{\sqrt{\frac{S_x^2}{m} + \frac{S_y^2}{n}}} \quad \text{where } D = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{m} y_i}{m} = \bar{x} - \bar{y}$$

Since this is a one-Sided test, to Validate null hypothesis we need

$$T > -\delta$$

Assume n & m are large

Using CLT we get $\bar{x} \sim Nor\left(\mu_1, \frac{\sigma_1^2}{n}\right)$ and

$\bar{y} \sim Nor\left(\mu_2, \frac{\sigma_2^2}{m}\right)$

$$\therefore \bar{D} = \bar{x} - \bar{y} \sim Nor\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} - \frac{\sigma_2^2}{m}\right)$$

As σ_1^2 and σ_2^2 are unknown, we can replace them with their plugin estimators.

$$\bar{D} \sim Nor\left(\mu_1 - \mu_2, \frac{S_x^2}{n} + \frac{S_y^2}{m}\right)$$

$$Pr(\text{Type I error}) = Pr(\text{reject } H_0 / H_0 \text{ true})$$

$$= P(T < -\delta)$$

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

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

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

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

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

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

$$Pr(\text{Type II error}) = Pr(\text{Accept } H_0 \mid H_1 \text{ true})$$

$$= Pr(T > -\delta)$$

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

Applying same steps as of type I error.

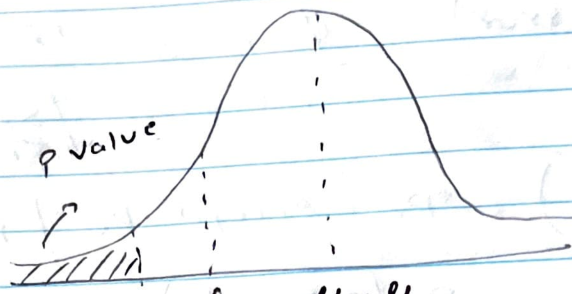
$$Pr(T > -\delta) = Pr\left(\frac{\bar{D} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} > -\delta - \frac{(\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

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

5b) For the unpaired test we get T-Statistic as

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$

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



$$P\text{value} = Pr\left(\frac{\bar{D}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{n}}} < \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}\right)$$

$$= Pr(\bar{D} < \bar{X} - \bar{Y})$$

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

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

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