CSE 544

PROBABILITY AND STATISTICS FOR DATA SCIENCE

ASSIGNMENT 5

Team members:

Name	SBU ID
Venkata Satwik Chilukuri	113263262
Akshay Somayaji	113322316
Shardul Churi	114403283

1)

- Given 10 samples,

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

Null hypothesis (H_0): The samples are from the uniform(0,3) distribution.

Alternate hypothesis: The samples are not from the uniform(0,3) distribution.

Sorting the given data,

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

Given,

$\hat{F}_{x^-}(x)$ = ecdf to the left of x

$\hat{F}_{x^+}(x)$ = ecdf to the Right of x

$F_Y(x)$ = cdf of uniform(0,3) at x

So, $F_Y(x) = \frac{x-0}{3-0}$ (since, cdf of unif(a,b) at x $= \frac{x-a}{b-a}$)

$= \frac{x}{3}$

x	$f_Y(x)$	$\hat{F}_{x_1}(x)$	$\hat{F}_{x_2}(x)$	$\lvert\hat{F}_{x_1}(x) - f_Y(x)\rvert$	$\lvert\hat{F}_{x_2}(x) - f_Y(x)\rvert$
0.04	0.0133	0	$\frac{1}{10}$	0.0133	0.0867
0.74	0.2466	$\frac{1}{10}$	$\frac{2}{10}$	0.1466	0.0466
0.84	0.28	$\frac{2}{10}$	$\frac{3}{10}$	0.08	0.02
1.19	0.3966	$\frac{3}{10}$	$\frac{4}{10}$	0.0966	0.0034
1.88	0.626	$\frac{4}{10}$	$\frac{5}{10}$	0.226	0.126
1.99	0.663	$\frac{5}{10}$	$\frac{6}{10}$	0.163	0.063
2.23	0.743	$\frac{6}{10}$	$\frac{7}{10}$	0.143	0.043
2.57	0.856	$\frac{7}{10}$	$\frac{8}{10}$	0.156	0.056
2.65	0.883	$\frac{8}{10}$	$\frac{9}{10}$	0.083	0.017
2.78	0.926	$\frac{9}{10}$	$\frac{10}{10}$	0.026	0.074
				Maximum value =	0.226

let ecdf represent $\hat{F}_D$

Now,

$$d = \max_{\alpha} \left| \hat{f}_D(\alpha) - F_x(\alpha) \right|$$

$$= 0.226$$

Given threshold is 0.25

Here, d is less that given threshold so Accept the null hypothesis (H_0)

The given samples are from the uniform(0,3) distribution

b)

Given data,

$$x = \{5\} \quad , \quad y = \{2, 7\}$$

So,

$$N = |x| + |y| = 3$$

$T_{observed}$ is the sampled mean difference of

x and y.

i.e,

$$T_{obs} = |\bar{x} - \bar{y}| = |5 - (\tfrac{2+7}{2})|$$

$$= |5 - 4.5|$$

$$= 0.5$$

In the table below, we will list the permutations for each permutation we have listed the sample values of x_i, y_i, T_i

i	x_i	y_i	$T_i = \lvert \bar{x}_i - \bar{y}_i \rvert$
1	5	{2, 7}	0.5
2	5	{7, 2}	0.5
3	2	{5, 7}	4
4	2	{7, 5}	4
5	7	{2, 5}	3.5
6	7	{5, 2}	3.5

Now,

we know the Indicator R.V for an Event E

$$I(E) = \begin{cases} 0 & \text{if } E \text{ does not occur} \\ 1 & \text{if } E \text{ occurs} \end{cases}$$

So, $T_{obs} = 0.5$

$T_i = \lvert \bar{x}_i - \bar{y}_i \rvert$	$I(T_i > T_{obs})$
0.5	0
0.5	0
4	1
4	1
3.5	1
3.5	1

Now,

$$\text{p-value} = \frac{4}{6} = \frac{2}{3} = 0.666$$

So

if the threshold is 0.05 and the p-value here is not less that 0.05. So we cant reject the null hypothesis

Thus x and y are from same distribution.

Question 3.

(a) Independence Test To Save Your Casino

Lets assume the dealers and outcome to be independent of each other.

∴ H_0 : outcome of the table independent of the dealer.

H_1 : outcome of the table not independent of the dealer

The OBSERVED OUTCOME.

	A	B	C	Total.
Win	48	54	19	121
Draw	7	5	4	16
Lose	55	50	25	130
Total	110	109	48	267

Considering the total values, we calculate the expected values.

	A	B	C	Total
win	49.85	49.4	21.75	121
Draw	6.59	6.53	2.88	16
Lose	53.56	53.07	23.37	130
Total	110	109	48	267

Now, calculating χ^2_{obs}

$$= \sum_{row=i}^{n} \sum_{col=j}^{n} \frac{(E_{ij} - O_{ij})^2}{E_{ij}}$$

$$= \frac{(49.85-48)^2}{49.85} + \frac{(49.4-54)^2}{49.4} + \frac{(21.75-19)^2}{21.75}$$
$$+ \frac{(6.59-7)^2}{6.59} + \frac{(6.53-5)^2}{6.53} + \frac{(2.88-4)^2}{2.88}$$
$$+ \frac{(53.56-55)^2}{53.56} + \frac{(53.07-50)^2}{53.07} + \frac{(23.37-25)^2}{23.37}$$

$$= 0.0689 + 0.428 + 0.348$$
$$+ 0.024 + 0.358 + 0.314$$
$$+ 0.038 + 0.188 + 0.106$$
$$= 1.8733$$

Calculating df : degree of freedom.

$$= (\#\text{ of rows} - 1)(\#\text{ of columns} - 1)$$
$$= (3-1)(3-1)$$
$$= 2 \times 2$$
$$= 4$$

Calculating p value for 1.8733 and df = 4

$$= 1 - 0.241 \quad [\text{USING TABLES}]$$
$$= 0.759$$

As $p > 0.05$ [α], we can't reject H_0.

∴ we conclude that the outcome of the table is independent of the dealer.

(b) To find: Pearson Corelation coefficient

DAY	A	B	C
1	48	54	19
2	40	48	40
3	58	51	35
4	53	47	41
5	65	62	38
6	25	35	32
7	52	70	32
8	34	20	37
9	30	25	37
10	45	40	15

$$\bar{A} = \frac{\sum A_i}{n} = \frac{450}{10} = 45$$

$$\bar{B} = \frac{\sum B_i}{n} = \frac{452}{10} = 45.2$$

$$\bar{C} = \frac{\sum C_i}{n} = \frac{326}{10} = 32.6$$

$$\hat{\rho}_{A,B} = \frac{\sum_{i=1}^{n} \{ (A_i - \bar{A})(B_i - \bar{B}) \}}{\sqrt{(\Sigma (A_i - \bar{A})^2) \cdot (\Sigma (B_i - \bar{B})^2)}}$$

$$= \frac{1396}{\sqrt{1462 \times 2193.6}} = \frac{1396}{1790.82} = 0.779$$

$$\hat{\rho}_{B,C} = \frac{\sum_{i=1}^{n} \{ (B_i - \bar{B})(C_i - \bar{C}) \}}{\sqrt{(\Sigma (B_i - \bar{B})^2) \cdot (\Sigma (C_i - \bar{C})^2)}}$$

$$= \frac{-96.2}{\sqrt{2193.6 \times 694.4}} = \frac{-96.2}{1234.19} = -0.078$$

$$\hat{\rho}_{A,C} = \frac{\sum_{i=1}^{n} \{(A_i - \bar{A})(C_i - \bar{C})\}}{\sqrt{(\Sigma (A_i - \bar{A})^2 (\Sigma (C_i - \bar{C})^2}}$$

$$= \frac{22}{\sqrt{1402 \times 694.4}} = \frac{22}{1007.58} = 0.022$$

$$\hat{\rho}_{A,B} = 0.779 \quad \hat{\rho}_{B,C} = -0.078 \quad \hat{\rho}_{A,C} = 0.022$$

There seems to be a +ve correlation between Dealers (A and B)

However there seems to be almost no correlation between (A and C) or (B and C)

With a value of +0.779, the correlation between Dealer A and Dealer B appears to be quite strong with value of A increasing or decreasing in accordance with value of B.

Hence, we can say that Dealer C may not be loyal, and could be the cause of the money loss.

4) Refer to q4.py, and the eCDF graph q4_KS_eCDF.png

5) Given, $D_1 = \{x_1, x_2 \dots x_n\} \overset{iid}{\sim} \text{Normal}(\mu_1, \sigma_1^2)$

$D_2 = \{y_1, y_2 \dots y_m\} \overset{iid}{\sim} \text{Normal}(\mu_2, \sigma_2^2)$

x's $\perp$ y's, n and m are large.

S_x and S_y are the sample std deviations of D_1 & D_2.

→ Also, $H_0: \mu_1 > \mu_2$ and $H_1: \mu_1 \leq \mu_2$ with $\delta > 0$ as the critical value

a) To calculate: Type-1 & Type-2 errors for the unpaired T-test using D_1 & D_2.

→ Let's define $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$, $\bar{Y} = \frac{\sum_{i=1}^{m} y_i}{m}$

$\Rightarrow \bar{D} = \bar{X} - \bar{Y}$ (i)

→ Now, since n & m are assumed to be large, we have

$\bar{X} \sim \text{Normal}\left(\mu_1, \frac{\sigma_1^2}{n}\right)$ ($\because$ via CLT)

$\bar{Y} \sim \text{Normal}\left(\mu_2, \frac{\sigma_2^2}{m}\right)$

$\Rightarrow \bar{D} \sim \text{Normal}\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}\right)$ (ii)

($\because \bar{X} - \bar{Y}$ follows weighted sum of $\perp$ Normals)

→ Now, T-statistic is given by

$$T = \frac{\bar{D}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$$ for an unpaired T-test

i) For Type-I error, we have

$Pr(\text{Type-I error}) = Pr(\text{False Positive})$

$= Pr(\text{Reject } H_0 \mid H_0 \text{ is true})$

For a one-tailed T-test with $H_0: \mu_1 > \mu_2$ & $H_1: \mu_1 \leq \mu_2$, we reject H_0 when $T < -\delta$

$\Rightarrow Pr(\text{Type I error}) = Pr(T < -\delta \mid \mu_1 > \mu_2)$

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

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

$= F_{\bar{D}}\left(-\delta\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{n}}\right)$ (CDF of dist^n $\bar{D}$ at point $D = -\delta\sqrt{\frac{S_x^2}{n} + \frac{S_x^2}{m}}$)

(iii)

$\rightarrow$ Now, since final result is in terms of Φ, let's transform $\bar{D}$ into standard normal $Z(0,1)$. From transformation property, we have at any point D, $Z = \frac{D - \mu_{\bar{D}}}{\sigma_{\bar{D}}}$ (iv)

From (ii), $\mu_{\bar{D}} = \mu_1 - \mu_2$, $\sigma_{\bar{D}} = \sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}$

$\rightarrow$ Now since m & n are large, we can plugin true, unknown std deviations σ_1 & σ_2 with the MLE estimator for std deviation of normal dist^n, which is the sample std deviations S_x & S_y respectively, via Asymptotic Normality.

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

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

$$\therefore Z = -\delta - \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \quad (v)$$

Applying this transformation, (iii) becomes

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

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

ii) Type 2 error probability can be written as

$$Pr(\text{Type 2 error}) = Pr(\text{False Negative})$$
$$= Pr(\text{Accept } H_0 \mid H_0 \text{ is False})$$

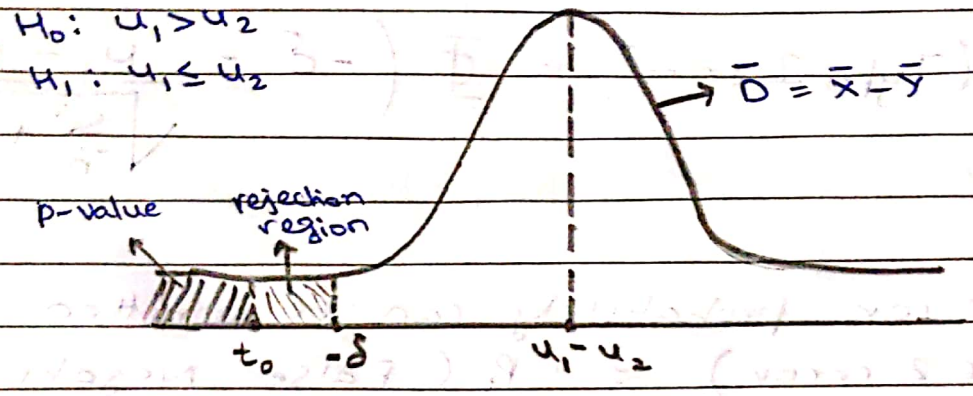
For a one-tailed T-test with $H_0: \mu_1 > \mu_2$ & $H_1: \mu_1 \leq \mu_2$, we accept H_0 when $T \geq -\delta$

$$\Rightarrow Pr(\text{Type 2 error}) = Pr(T \geq -\delta \mid \mu_1 \leq \mu_2)$$
$$= 1 - Pr(T < -\delta \mid \mu_1 \leq \mu_2)$$
$$= 1 - Pr\left(\frac{\bar{D}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} < -\delta \mid \mu_1 \leq \mu_2\right)$$
$$= 1 - Pr\left(\bar{D} < -\delta\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}} \mid \mu_1 \leq \mu_2\right)$$
$$= 1 - F_{\bar{D}}\left(-\delta\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}\right) \quad (vi)$$

Using Z-transformation properties as in part i) and using results of (iii), (iv) & (v), we get

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

b) For a 1-tailed T-test, we can represent the region of p-value as below -



Now, p-value = Pr (Finding more extreme statistic than t_0 | H_0 is true)

$= Pr(T < t_0)$

Let $t_0 = \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$

$\Rightarrow$ p-value $= Pr\left(\dfrac{\bar{D}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} < \dfrac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \right)$

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

Using (iv), we can transform into std normal by

$$Z = \frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \quad (\text{at } D = \bar{X} - \bar{Y})$$

$\therefore$ p-value $= F_Z\left(\dfrac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \right)$

$$\boxed{\text{p-value} = \Phi\left(\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}} \right)}$$

6a) Given, $X \sim \text{Normal}(1.5, 1)$ and $Y \sim \text{Normal}(1, 1)$

i) Z-Test for 20 sample x_1 & y_1 -

$\sigma_{x_1} = 1$, $\sigma_{y_1} = 1$, $\bar{x}_1 = 1.5987$, $\bar{y}_1 = 1.0625$, $n = m = 20$

sample pool std deviation $= \sqrt{\frac{\sigma_{x_1}^2}{n} + \frac{\sigma_{y_1}^2}{m}} = \sqrt{\frac{1}{20} + \frac{1}{20}}$

$= \sqrt{0.1} = 0.3162$

Since we're assuming null hypothesis, the hypothesized difference between sample means = 0

$$\Rightarrow z = \frac{\bar{x}_1 - \bar{y}_1}{\text{sample pool std dev}} = \frac{1.5987 - 1.0625}{0.3162}$$

$$\Rightarrow z_{20} = 1.6957 \quad (i)$$

$$z\ p\text{-value}_{20} = 2 * (1 - \Phi(|z|)) = 2 * (1 - \Phi(1.6957)) = 0.0899 \quad (ii)$$

Given, critical threshold = 1.962 = c

since $|z| = 1.6975 < c$, we fail to reject the null hypothesis, and accept that the mean values of X and Y are equal.

ii) Unpaired T-test for 20 samples of x_1 & y_1 -

$\Rightarrow s_{x_1}^2 = 1.1772$, $s_{y_1}^2 = 1.1412$

$$\text{now, } T = \frac{\bar{x}_1 - \bar{y}_1}{\sqrt{\frac{s_{x_1}^2}{n} + \frac{s_{y_1}^2}{m}}} = \frac{1.5987 - 1.0625}{\sqrt{\frac{1.1772}{20} + \frac{1.1412}{20}}}$$

$$\Rightarrow T_{20} = 1.5750 \quad (iii)$$

$$T_{p\text{-value}_{20}} = 2 * (1 - F_{T_{df}}(|T|))$$

Now, degrees of freedom $df = n + m - 2 = 38$

$$\Rightarrow T_{p\text{-value}} = 2 * (1 - F_{T_{38}}(1.5750)) = 0.1236 \quad (iv)$$

Notes: $s_{x_1}^2 = \sum_{i=1}^{20} (x_i - \bar{x}_1)^2 / n-1 = 1.1772$

$s_{y_1}^2 = \sum_{i=1}^{20} (y_i - \bar{y}_1)^2 / m-1 = 1.1412$

Given, critical threshold $c = 2.086$

since $|T| = 1.5750 < c$, we fail to reject the null hypothesis, and accept that the mean values of X and Y are equal.

b) i) Z-test for 1000 sample X_2 & Y_2

$\sigma_{X_2} = 1$, $\sigma_{Y_2} = 1$, $\bar{X}_2 = 1.4613$, $\bar{Y}_2 = 0.9835$

$m = n = 1000$, sample pool std dev $= 0.0447$

$$\Rightarrow z_{1000} = \frac{1.4613 - 0.9835}{0.0447} = 10.6826 \quad \text{(v)}$$

$$z\ p\text{-value}_{1000} = 2(1 - \phi(10.6826)) \cong 0.0000 \quad \text{(vi)}$$

Given, critical value $c = 1.962$

since $|z| = 10.6826 > c$, we reject the null hypothesis and accept the alternate one that the means of X and Y are not equal

ii) T-test for 1000 sample X_2 & Y_2

$S_{X_2}^2 = 1.0299$, $S_{Y_2}^2 = 0.9586$

$$\Rightarrow T_{1000} = \frac{1.4613 - 0.9835}{\sqrt{\frac{1.0299}{1000} + \frac{0.9586}{1000}}} = 10.7134 \quad \text{(vii)}$$

$$T\ p\text{-value}_{1000} = 2(1 - F_{T_{998}}(10.7134)) \cong 0.000 \quad \text{(viii)}$$

Given, critical value $c = 1.962$

Since $|T| = 10.7134 > c$, we reject the null hypothesis and accept the alternate one that the means of X and Y are not equal

→ Observations -

1. We got a True Positive result in part b), as using both z and T tests, we were successfully able to reject the null hypothesis, that is, infer that the means of X and Y are not same, given we already have the ground truth that mean of X is 1.5 and mean of Y is 1. We can hence claim that both z and T tests work well with large data samples

2. There is not much of a significant advantage of using z test on smaller samples, as both tests on the 20-sample data failed to reject the null hypothesis (even though ground truth says otherwise), thereby resulting in a false negative. However, the p-value is much closer to $\alpha = 0.05$ in the case of the z-test (0.0899) than in the T-test (0.1236)

3. We can see that the p-value is significantly lesser than $\alpha = 0.05$ in part b) with 1000 samples, for both z and T tests. Therefore, we can support our rejection of the null hypothesis with extremely high confidence, since $0.0000 << 0.05$.