SHUBHAM AGRAWAL - 113166701
RANJAN KUMAR - 113262786
AMEYA SANKHE - 113219562
PRATIK NAGELIA - 114122014

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

Assignment 5: Hypothesis Testing Due: 4/20, 1:15pm, via Blackboard

(6 questions, 70 points total)

I/We understand and agree to the following:
(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.
(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.

(write down the name of all collaborating students on the line below)

SHUBHAM AGRAWAL, RANJAN KUMAR, AMEYA SANKHE, PRATIK NAGELIA

1. Hypothesis Testing for a single population **(Total 7 points)**

Consider the 10 samples: {2.78, 0.84, 1.88, 2.23, 1.99, 0.04, 2.65, 0.74, 1.19, 2.57}. Use the K-S test to check whether these samples are from the Uniform(0, 3) distribution. First, set up the hypotheses. Then, create a 10 X 6 table with entries: $[x, F_Y(x), \hat{F}_X^-(x), \hat{F}_X^+(x), |\hat{F}_X^-(x) - F_Y(x)|, |\hat{F}_X^+(x) - F_Y(x)|]$, where $\hat{F}_X^-(x)$ and $\hat{F}_X^+(x)$ are the values of the eCDF to the left and right of x, and $F_Y(x)$ is the CDF of Uniform(0, 3) at x; this is the same notation as in class. Finally, compare the max difference with the threshold of 0.25 to Reject/Accept. Show all rows and columns.

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

y = Critical threshold, $c = 0.25$

$$F_{uniform(0,3)}(x) = \frac{x-0}{3-0} = \frac{x}{3}$$

$H_0 : F_D = F_y$ vs $H_1 : F_D \neq F_y$

x	$F_y(x)$	F_X^-	F_X^+	$\lvert \hat{F}_X^- - F_y(x) \rvert$	$\lvert F_X^+ - \hat{F}_y(x) \rvert$
0.04	0.0133	0.0	0.1	0.0133	0.0867
0.74	0.2466	0.1	0.2	0.1466	0.0466
0.84	0.28	0.2	0.3	0.08	0.02
1.19	0.3966	0.3	0.4	0.0966	0.0034
1.88	0.6266	0.4	0.5	$\boxed{0.2266}$	0.1266
1.99	0.6633	0.5	0.6	0.1633	0.0633
2.23	0.7433	0.6	0.7	0.1433	0.0433
2.57	0.8566	0.7	0.8	0.1566	0.0566
2.65	0.8833	0.8	0.9	0.0833	0.0167
2.78	0.9266	0.9	1	0.0266	0.0734

From the above table, we can see that maximum difference is 0.2266. Since $0.2266 < (c = 0.25)$. Hence we accept H_0. (Null Hypothesis)

2. Toy Example for Permutation Test (Total 5 points)

Let X = {5} and Y = {2, 7}. The null hypothesis is that X and Y are from the same distribution. Use the permutation test to decide this using a p-value threshold of 0.05. Please show all steps for each permutation clearly.

Ans. $H_0 : X \equiv Y$

$X = \{5\}$

$Y = \{2, 7\}$

the difference of Means for the given Dataset :

$T_{obs} = |4.5 - 5| = 0.5$

The difference of Means for 6 possible permutations are as follows:-

Sl.No	Permutation	$T_i : \lvert\bar{X} - \bar{Y}\rvert$
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

$$P_{value} = \frac{1}{N!} \sum_{i=1}^{N!} I \left(T_i > T_{obs} \right)$$

$$= \frac{1}{3!} \times \left[0 + 0 + 1 + 1 + 1 + 1 \right] = \frac{1}{6} \times 4 = \frac{2}{3}$$

$$= 0.66$$

As pvalue = 0.66 > 0.05 (threshold).

So we will reject Ho (Null Hypothesis).

3. Independence Tests to Save Your Casino (Total 15 points)

Being the owner of Casino 544, you are concerned that you are losing a lot of money because of the dealers at the blackjack tables. The Null hypothesis is that the outcome of the tables should be independent of the dealer, but you aren't sure.

(a) Validate your claim based on the dealer observations for a day, using the χ^2 test. Use α=0.05. You can use tools/online resources to find the CDF of χ^2; one such tool is https://www.danielsoper.com/statcalc/calculator.aspx?id=62. (10 points)

	Dealer A	Dealer B	Dealer C
Win	48	54	19
Draw	7	5	4
Loose	55	50	25

(b) You want to be more certain about the loyalty of your dealers, so you collect more data: number of wins from each dealer for 10 days. Find the Pearson correlation coefficient for each pair of dealers. What can you conclude? (5 points)

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

3.(a) Given: Observed Values:

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

Expected Values:

	Dealer A	Dealer B	Dealer C
Win	$\frac{110}{267} \times 121 = 49.85$	$\frac{109}{267} \times 121 = 49.40$	$\frac{48}{267} \times 121 = 21.75$
Draw	$\frac{110}{267} \times 16 = 6.59$	$\frac{109}{267} \times 16 = 6.53$	$\frac{48}{267} \times 16 = 2.88$
Loose	$\frac{110}{267} \times 130 = 53.56$	$\frac{109}{267} \times 130 = 53.07$	$\frac{48}{267} \times 130 = 23.37$

Evaluating the χ^2 test:

$$Q_{obs} = \sum_r \sum_c \frac{(E_{rc} - O_{rc})^2}{E_{rc}}$$

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

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

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

$$= \frac{(49.85 - 48)^2}{49.85} + \frac{(6.59 - 7)^2}{6.59} + \frac{(53.56 - 55)^2}{53.56}$$

$$+ \frac{(49.40 - 54)^2}{49.40} + \frac{(6.53 - 5)^2}{6.53} + \frac{(53.07 - 50)^2}{53.07}$$

$$+ \frac{(21.75 - 19)^2}{21.75} + \frac{(2.88 - 4)^2}{2.88} + \frac{(23.37 - 25)^2}{48}$$

putting values from expected & observed tables

$= 0.0687 + 0.0255 + 0.0387 + 0.4283 + 0.3584$
$+ 0.1776 + 0.3477 + 0.43556 + 0.1137$

$\therefore Q_{obs} = 1.9942$

degrees of freedom (df) = (# datarows − 1) (# data column − 1)
$= (3-1) \times (3-1) = 2 \times 2 = 4$

p-value $= \Pr(\chi^2_4 > 1.9942) = 1 - 0.263 = 0.737$

As p-value $(0.737) > 0.05$, we fail to reject H_0

3(b) To find the Pearson correlation coefficient:

$$S_{xy} = \frac{\sum_{i=1}^{n} \{(X_i - \bar{X})(Y_i - \bar{Y})\}}{\sqrt{\left(\sum_{i=1}^{n} (X_i - \bar{X})^2\right)\left(\sum_{i=1}^{n} (Y_i - \bar{Y})^2\right)}}$$

Dealer A : $\bar{A} = \sum_{i=1}^{n} X_i \times \frac{1}{n} = \frac{450}{10} = 45$

Dealer B : $\bar{B} = \sum_{i=1}^{n} X_i \times \frac{1}{n} = \frac{452}{10} = 45.2$

Dealer C : $\bar{C} = \sum_{i=1}^{n} X_i \times \frac{1}{n} = \frac{326}{10} = 32.6$

Day	A	$(A-\bar{A})$	$(A-\bar{A})^2$	B	$(B-\bar{B})$	$(B-\bar{B})^2$	C	$(C-\bar{C})$	$(C-\bar{C})^2$
1	48	3	9	54	8.8	77.44	19	-13.6	184.96
2	40	-5	25	48	2.8	7.84	40	7.4	54.76
3	58	13	169	51	5.8	33.64	35	2.4	5.76
4	53	8	64	47	1.8	3.24	41	8.4	70.56
5	65	20	400	62	16.8	282.24	38	5.4	29.16
6	25	-20	400	35	-10.2	104.24	32	-0.6	0.36
7	52	7	49	70	24.8	615.04	32	-0.6	0.36
8	34	-11	121	20	-25.2	635.04	37	4.4	19.36
9	30	-15	225	25	-20.2	408.04	37	4.4	19.36
10	45	0	0	40	-5.2	27.04	15	-17.6	309.76
Total	450		1462	452		2193.8	326		694.4

(i) Finding correlation coefficient between Dealer A and Dealer B:

$$S_{AB} = \frac{\sum_{i=1}^{n} \{(A_i - \bar{A})(B_i - \bar{B})\}}{\sqrt{\left(\sum_{i=1}^{n} (A_i - \bar{A})^2\right)\left(\sum_{i=1}^{n} (B_i - \bar{B})^2\right)}}$$

From the table;

$$\sum_{i=1}^{n} (A-\bar{A})(B-\bar{B}) = 1396 \quad \text{——} \quad (1)$$

$$\sum_{i=1}^{n} (B-\bar{B})(C-\bar{C}) = -96.20 \quad \text{——} \quad (2)$$

$$\sum_{i=1}^{n} (A-\bar{A})(C-\bar{C}) = 22.0 \quad \text{——} \quad (3)$$

$$\sqrt{\sum_{i=1}^{n} (A-\bar{A})^2} = 38.2361 \quad \text{——} \quad (4)$$

$$\sqrt{\sum_{i=1}^{n} (B-\bar{B})^2} = 46.836 \quad \text{——} \quad (5)$$

$$\sqrt{\sum_{i=1}^{n} (C-\bar{C})^2} = 26.35 \quad \text{——} \quad (6)$$

(i) $$\rho_{AB} = \frac{1396}{38.2361 \times 46.836} = 0.7795$$ [Putting values from (1), (4) and (5)]

(7)

(ii) Finding correlation coffecient between Dealer B and Dealer C:

$$\rho_{BC} = \frac{\sum_{i=1}^{n} \{(B_i - \bar{B})(C_i - \bar{C})\}}{\sqrt{\left(\sum_{i=1}^{n} (B_i - \bar{B})^2\right)\left(\sum_{i=1}^{n} (C_i - \bar{C})^2\right)}}$$

$$= \frac{-96.20}{46.836 \times 26.35} = -0.0779$$ [Putting values from (2), (5) and (6)]

(8)

(iii) Finding correlation coefficient between Dealer A and Dealer C:

$$\rho_{AC} = \frac{\sum_{i=1}^{n} \{(A_i - \bar{A})(C_i - \bar{C})\}}{\sqrt{\left(\sum_{i=1}^{n} (A_i - \bar{A})^2\right)\left(\sum_{i=1}^{n} (C_i - \bar{C})^2\right)}}$$

$$= \frac{22.0}{38.2361 \times 26.35} = 0.0218$$ [Putting values from (3), (4) and (6)] — (9)

From (7): $\rho_{A,B} = 0.7795 \ (> 0.5) \rightarrow$ Positive linear correlation

From (8): $\rho_{BC} = |-0.0779| \ (< 0.5) \rightarrow$ No linear correlation

From (9): $\rho_{AC} = |0.0218| \ (< 0.5) \rightarrow$ No linear correlation

From the results above, we can conclude that Dealer A is positively correlated with B. However Dealer C is not correlated with Dealer A and Dealer B.

5. Type-1 and Type-2 error for one-sided unpaired T-test (Total 10 points)

Let $\{X_1, X_2, \ldots, X_n\}$ be i.i.d. from Normal(μ_1, σ_1^2) and $\{Y_1, Y_2, \ldots, Y_m\}$ be i.i.d. from Normal(μ_2, σ_2^2). Also suppose X's and Y's are independent, and $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$ are unknown. Let S_x and S_y be the sample standard deviations of the two populations. Assume that n and m are large. Let H_0: $\mu_1 > \mu_2$ be the null hypothesis and H_1: $\mu_1 <= \mu_2$ be the alternate hypothesis. Consider the T statistic for the unpaired T test, as in class, with $\delta > 0$ being the critical value.

(a) For the above test, show that the probability of Type-1 and Type-2 errors are given by $\Phi\left(-\delta - \frac{\mu_1-\mu_2}{\sqrt{\frac{Sx^2}{n}+\frac{Sy^2}{m}}}\right)$ and $1-\Phi\left(-\delta - \frac{\mu_1-\mu_2}{\sqrt{\frac{Sx^2}{n}+\frac{Sy^2}{m}}}\right)$, respectively. (5 points)

(b) Show that the p-value is given by $\Phi\left(\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{Sx^2}{n}+\frac{Sy^2}{m}}}\right)$. (5 points)

Given:

$$X_1, X_2 \ldots X_n \overset{iid}{\sim} N(\mu_1, \sigma_1^2)$$

$$Y_1, Y_2 \ldots Y_m \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$$

also $X's \perp Y's$.

$$\bar{X} \sim N\left(\mu_1, \frac{\sigma_1^2}{n}\right)$$

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

$$\bar{X}-\bar{Y} \sim N\left(\mu_1-\mu_2, \frac{\sigma_1^2}{n}+\frac{\sigma_2^2}{m}\right) \quad [\text{Since } X's \perp Y's]$$

$$T = \frac{\bar{X}-\bar{Y}}{\sqrt{\frac{Sx^2}{n}+\frac{Sy^2}{m}}}$$

(a) According to question:

$H_0: \mu_1 > \mu_2$, $H_1: \mu_1 \leq \mu_2$

Type 1 error: $Pr(\text{Test reject } H_0 \mid H_0 \text{ true})$

$$\Rightarrow P(T < -\delta \mid H_0 \text{ true})$$

$$= P\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} < -\delta \;\Big|\; H_0 \text{ true}\right)$$

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

[subtracting $\frac{\mu_1-\mu_2}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}}$ from both sides]

Since n and m are very large, S_x is a consistent estimator of σ_x and S_y is a consistent estimator of σ_y.

$$\frac{\bar{X}-\bar{Y}-(\mu_1-\mu_2)}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} \sim T \sim N(0,1)$$

From above: Type 1 error: $P\left(T < -\delta - \frac{(\mu_1-\mu_2)}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}}\right)$

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

Type 2 error: $Pr(\text{Test accept } H_0 \mid H_0 \text{ false})$

$$= P(T > -\delta \mid H_0 \text{ false})$$

$$= P\left(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_x^2}{n}+\frac{S_y^2}{m}}} > -\delta \;\Big|\; H_0 \text{ false}\right)$$

$$= 1 - P\left(\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} < -\delta \right)$$

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

[Subtracting $\frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}}$ from both sides]

Since n and m are very large, S_x is a consistent estimator of σ_x and S_y is a consistent estimator of σ_y.

$$\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \sim T \sim N(0, 1)$$

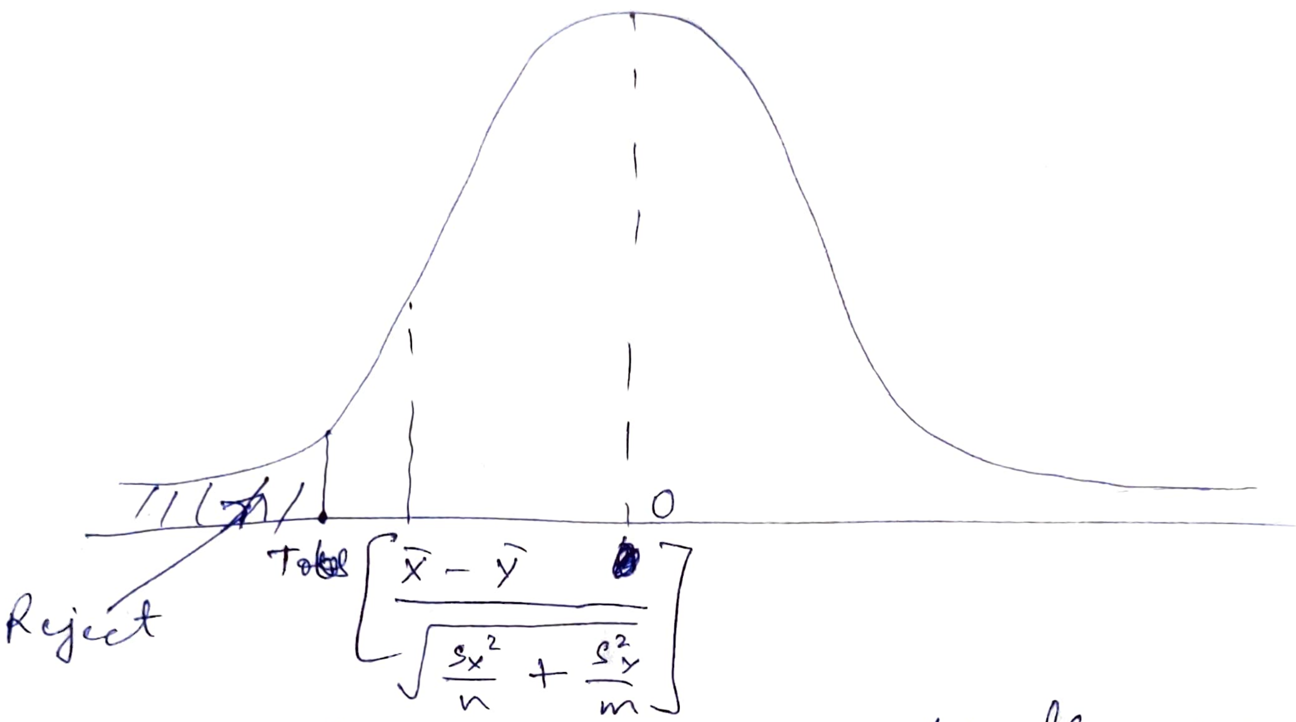
From above: Type 2 error $= 1 - P\left(T < -\delta - \frac{\mu_1 - \mu_2}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \right)$

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

proved!

~~(b)~~

1(b) Ans. The Normal distribution graph is as follows:-



Ho is rejected if Tobs lies all the way to the left in the above graph.

p value = Area to the left of Tobs = ϕ(Tobs)

$$T_{obs} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}$$ be the statistic that is observed.

$$p\text{-value} = P(T < t_{obs}) = P\left(\frac{\bar{x}-\bar{y}}{\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}}\right)$$

On subtracting $(\mu_1-\mu_2)$ from both side of the above inequality, we get:-

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

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

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

{ Hence Proved }

Q6)
Steps:
We load the data, and calculate the mean, standard deviation, statistics
In Z test we use the true variance for the calculation of standard deviation.
True Variance given for both dist. Is 1.
```
diff = np.mean(X1) - np.mean(Y1)
std = np.sqrt( varX1True /len(X1) + varY1True /len(Y1))
```
Now to calculate Z, we use Z = np.absolute(diff/std)

For T, we use sample variance for both the dist.
```
Sample Variance :
var_X1 = np.sum(np.square(X1 - X1_mean))/(len(X1)-1)
var_Y1 = np.sum(np.square(Y1 - Y1_mean))/(len(Y1)-1)
Standard Deviation :
std = np.sqrt( var_X1 /len(X1) + var_Y1 /len(Y1))
```
Now, T = np.absolute(diff/std)

a)
Mean value of X1 : 1.598738564848365
Mean value of Y1 : 1.06250437094748

H0 : X and Y have the same mean value
Std Deviation value : 0.31622776601683794
Using Z test, we get:
Value of Z : 1.6957214119911677
Accept H0
P Value : 0.08993865154335023

H0 : X and Y have the same mean value
Std Deviation value : 0.3404727031767004
Using T test, we get:
Value of T : 1.5749697079903267
Accept H0
P Value : 0.13176797257945405

Q6) b)

Mean value of X1 : 1.461258989664249
Mean value of Y1 : 0.9835202047160342
H0 : X and Y have the same mean value

Std Deviation value : 0.044721359549995794
Using Z test, we get:
Value of Z : 10.682563986323615

Reject H0
P Value : 0.0

H0 : X and Y have the same mean value
Std Deviation value : 0.0445925196297373
Using T test, we get:
Value of T : 10.713428819788561
Reject H0
P Value : 0.0

Is there a significant advantage of using Z-test in a small sample?
Ans : No there is no significant advantage for Z test in a small sample.