Maths and Statistics for AI and Data Science

Practical Assessment – 2

by Uttara Naidu

Submitted to

University of Liverpool

UNIVERSITY OF LIVERPOOL

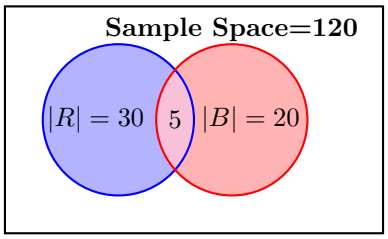
Question 1

1. Out of 120 students in a school, 30 play rugby, 20 play badminton, and 5 play both sports.
   a. How many students play neither sport?
   b. What is the probability that a randomly chosen student plays badminton?
   c. Given that a chosen student plays rugby, what is the probability that they also play badminton?
2. Box 1 contains 6 red beans and 4 green beans. Box 2 contains 5 red beans and 3 green beans. A fair six-sided die is thrown. If a '2' is obtained, a bean is selected from box 1; otherwise, a bean is selected from box 2. Given that the bean selected was red, what is the probability that it came from box 1?
3. The probability that a student is late for a lesson is 0.1, independently of any other students. What is the probability that at least one of 10 students in a class is late?

Provide appropriate justification and explanation to all your answers, detailing the methods used.

Solution:

1.a) Find number of students who play neither Badminton nor Rugby



The above Venn diagram represents the given population.

Let the number of students playing Rugby be represented as $|R|$ and for Badminton be represented as $|B|$.

$$|R| = 30$$
$$|B| = 20$$
$$|R \cap B| = 5$$

As per the principle of Inclusion-Exclusion (ACP,2024), we get the number of active sports players as,

$$\begin{aligned} |R \cup B| &= |R| + |B| - |R \cap B| \\ &= 30 + 20 - 5 \\ |R \cup B| &= 45 \end{aligned} \tag{1}$$

Hence the total number of students who play either Rugby or Badminton or both is 45.

We have the sample space $|S|$ (or universal set) of size 120. Calculating the total number of students $|T|$ playing neither sport as below,

$$|T| = |S| - |R \cup B|$$
$$= 120 - 45$$

$$\boxed{|T| = 75} \tag{2}$$

1.b) Find the probability that a randomly chosen student plays badminton

Since the problem statement asks for a "randomly chosen student" (and not "only the students playing sports"), we consider the entire sample space of 120.

Hence, the probability $P(B)$ that a randomly chosen student plays Badminton is given by,

$$P(B) = \frac{number\ of\ students\ playing\ Badminton}{total\ student\ population}$$
$$= \frac{20}{120}$$
$$= 0.1667$$

$$\boxed{P(B) \approx 16.67\%} \tag{3}$$

1.c) Find the probability that a chosen student plays Badminton given that they play Rugby

Given the occurence of event B, the probability of event A occuring is given by below formula (UoL,2025):

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{A}$$

This is called conditional probability.

For the problem statement, let us state the below terms:

$P(R)$ is probability of a chosen student who plays Rugby.

$P(B)$ is probability of a chosen student who plays Badminton.

$P(B \cap R)$ is probability of chosen student who plays both Rugby and Badminton

Using these terms in formula (A), we get,

$$P(B|R) = \frac{P(B \cap R)}{P(R)} \tag{4}$$

where $P(B|R)$ is the probability that a student plays Badminton given that they play Rugby.

From equation (1), let us consider a sub-sample space of student population of 45 i.e. only the students who are active in sports.

Step i: Calculating $P(B \cap R)$

$$\begin{aligned} P(B \cap R) &= \frac{number\ of\ students\ playing\ Rugby\ and Badminton}{|R \cup B|} \\ &= \frac{|B \cap R|}{45} \\ &= \frac{5}{45} \\ P(B \cap R) &= \frac{1}{9} \end{aligned} \tag{5}$$

Step ii: Calculating $P(R)$

$$\begin{aligned} P(R) &= \frac{number\ of\ students\ playing\ Rugby}{|R \cup B|} \\ &= \frac{30}{45} \\ P(R) &= \frac{2}{3} \end{aligned} \tag{6}$$

Subsctituting (5) and (6) into (4), we get,

$$\begin{aligned} P(B|R) &= \frac{\frac{1}{9}}{\frac{2}{3}} \\ &= \frac{1}{9} \times \frac{3}{2} \\ &= \frac{1}{6} \end{aligned}$$

Therefore, probability that a chosen student plays Badminton, given that they also play rugby is,

$$\boxed{P(B|R) = 0.167}$$

2) Finding the probability that the bean came from box 1 given that the bean selected was red

Summarising the given information as below:

	Box 1	Box 2	Total
Red	6	5	11
Green	4	3	7
Total	10	8	

Table 1: Given data

Let,

Event B_i = pick a bean from box i (i=1 or 2)

Event A = a red bean is picked

"Baye's Theorem gives us the probability of an event, based on prior knowledge of conditions related to the event" (UoL,2025).

Since we need to find the probability of an event B_i based on the condition that a '2' is obtained on throwing a fair-sided die, Baye's theorem is applied to compute this probability.

Baye's theorem is given by,

$$P(B_i|A) = \frac{P(A|B_i) \times P(B_i)}{P(A)} = \frac{P(A|B_i) \times P(B_i)}{\sum_{i=1}^{m} P(A|B_i) \times P(B_i)} \tag{B}$$

Using this formula for the given problem statement, we get,

$$P(B_1|A) = \frac{P(A|B_1) \times P(B_1)}{P(A)} = \frac{P(A|B_1) \times P(B_1)}{P(A|B_1) \times P(B_1) + P(A|B_2) \times P(B_2)} \tag{7}$$

Calculating $P(B_1)$ and $P(B_2)$

Since the die is fair,

$$P(B_1) = P(2\ is\ obtained) = \frac{1}{6} \tag{8}$$

$$P(B_2) = P(1, 3, 4, 5, 6\ is\ obtained) = \frac{5}{6} \tag{9}$$

Calculating $P(A|B_1)$ and $P(A|B_2)$

Using values from $Table$ 1.,

$$P(A|B_1) = \frac{no.\ of\ red\ beans\ in\ Box\ 1}{total\ no.\ of\ beans\ in\ Box\ 1} = \frac{6}{10}$$

$$P(A|B_1) = \frac{3}{5} \tag{10}$$

$$P(A|B_2) = \frac{no.\ of\ red\ beans\ in\ Box\ 2}{total\ no.\ of\ beans\ in\ Box\ 2} = \frac{5}{8}$$

$$P(A|B_2) = \frac{5}{8} \tag{11}$$

Calculating $P(A)$

Substituing values from $(8), (9), (10), (11)$ into denominator of equation (7), we have,

$$\begin{aligned} P(A) &= P(A|B_1) \times P(B_1) + P(A|B_2) \times P(B_2) \\ &= \frac{3}{5} \times \frac{1}{6} + \frac{5}{8} \times \frac{5}{6} \\ &= \frac{1}{10} + \frac{25}{48} \end{aligned}$$

$$P(A) = \frac{149}{240} \tag{12}$$

Calculating $P(B_1|A)$

Substituting values from $(8), (10), (12)$ into equation (7)

$$\begin{aligned} P(B_1|A) &= \frac{\frac{6}{10} \times \frac{1}{6}}{\frac{149}{240}} \\ &= \frac{1}{10} \times \frac{240}{149} = \frac{24}{149} \\ &= 0.161 \end{aligned}$$

$$\boxed{P(B_1|A) \approx 16.11\%}$$

3) Finding the probability that at least one of 10 students in a class is late

Let the probability that a student is late be $P(L)$

$P(L) = 0.1$

The probability of an event is always between $0 \leq P(E) \leq 1$. Each event A has a corresponding negation event which is denoted by $(not\ A)$ or (A'). This principle is called a complement rule (Dukkipati,2013). These two event never occur together but one of them always has to occur (UoL,2025). Therefore, the complement rule is given by,

$$\begin{aligned} P(A) + P(not\ A) &= 1 \\ P(A) + P(A') &= 1 \\ P(A') &= 1 - P(A) \end{aligned} \tag{C}$$

Applying ccomplement rule to the problem statement, we have,

Event L= a student being late with a probability $P(L)$

Event not L= a student not late (i.e. is on time) with a probability $P(L')$

Event A= not a single student is late i.e. all 10 students are not late with a probability $P(A)$

Event not A= at least one student is late with a probability $P(A')$

Since we already have a probability of a student being late, let us calculate the probability of a student not being late (on time) using complement rule (C).

$$\begin{aligned} P(L') &= 1 - 0.1 \\ P(L') &= 0.9 \end{aligned} \tag{13}$$

The events of a student being late are independent from other students. The multiplication rule for independent events is given by,

$$P(A \text{ and } B \text{ and } C \text{ and } ...) = P(A) \cdot P(B) \cdot P(C)... \tag{D}$$

Applying this multiplication rule for calculating the probability of all 10 students being late,

$$\begin{aligned} P(A) &= P(L' \text{ for student } 1) \cdot P(L' \text{ for student } 2) \cdot P(L' \text{ for student } 3)... \cdot P(L' \text{ for student } 10) \\ &= 0.9^{10} \\ P(A) &= 0.3487 \end{aligned} \tag{14}$$

By applying complement rule to $P(A)$, we calculate the probability that at least one student is late,

$$\begin{aligned} P(A') &= 1 - P(A) \\ &= 1 - 0.3487 \\ &= 0.6513 \end{aligned}$$

Therefore the probability of at least one student, in a class of 10, is late is,

$$\boxed{P(A') \approx 65.13\%}$$

Question 2

1. For the dataset {5.78, 6.71, 6.84, 7.23, 8.20, 9.65, 13.44, 14.71, 16.39, 24.37}, give your answer to the below questions in 2 decimal places.

a. Find sample mean, median and standard deviation

b. Find 1st quartile (Q1), 3rd quartile (Q3) and Interquartile Range (IQR). Identify the outliers of the dataset

2. The main body weight of 500 male student at a university is 72kg, and the standard deviation is 10kg. Assuming the weights are normally distributed, find how many students weigh between 65kg to 75kg. Table 1 is the z-table in which entries for z represent the area under the bell curve to the left of z

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981

Figure 1: z-table

Provide appropriate justification and explanation to all your answers, detailing the methods used.

Solution:

1.a) Calculating sample mean, median and standard deviation

The given dataset represents ungrouped data. The mean, also called measures of central tendency or average, for an ungrouped data is given by below formula (Dukkipati,2013),

$$Mean\ for\ a\ population\ data,\ \mu = \frac{\sum X}{N}$$
$$Mean\ for\ a\ sample\ data,\ \bar{X} = \frac{\sum X}{n}$$

where μ is Mean of population data

$\sum X$ is sum of all variables

N is population size

$\bar{X}$ is sample Mean

n is sample size

Since we have only a sample of 10 data points, we will use below formula for mean calculation,

$$\bar{X} = \frac{\sum X}{n} \tag{B}$$

Using formula (B) for the given dataset as below:

$$\bar{X} = \frac{5.78 + 6.71 + 6.84 + 7.23 + 8.20 + 9.65 + 13.44 + 14.71 + 16.39 + 24.37}{10}$$

$$\boxed{\bar{X} = 11.33} \tag{15}$$

"Median is the value of the middle term in a dataset that has been ranked in either increasing or decreasing order." (Dukkipati,2013) It can be calculated as,

$$Median\ of\ ungrouped\ data = \frac{n+1}{2}\ th\ term\ in\ ranked\ dataset \tag{C}$$

Since the given dataset has 10 elements, an even number, the median will be an average of the values of two middle terms.

The given dataset is already sorted in an increasing order. Hence, from definition (C) we get,

$$Median = \frac{8.20 + 9.65}{2}$$

$$\boxed{Median = 8.93} \tag{16}$$

Standard deviation is a measure of how much the data points in a sample or population deviates from the mean of the sample or population. (W3S, 2025).

The standard deviation is given by,

$$The\ population\ Standard\ deviation,\ \sigma = \sqrt{population\ variance} = \sqrt{\sigma^2}$$
$$The\ sample\ Standard\ deviation,\ s = \sqrt{sample\ variance} = \sqrt{s^2}$$

We will use below formula to calculate stadard deviation of a mean,

$$s = \sqrt{sample\ variance} = \sqrt{s^2}$$

The sample variance, in turn, is given by,

$$s^2 = \frac{\sum(x - \bar{X})^2}{n - 1} \tag{C}$$

where $(x - \bar{X})$ is the deviation of x value from mean

s is sample variance

n is sample size

Applying formula (C) to the given dataset and substituting the sample Mean value from (15), we get,

$$\begin{aligned} s^2 &= \frac{(5.78 - 11.33)^2 + (6.71 - 11.33)^2 + ... + (16.39 - 11.33)^2 + (24.37 - 11.33)^2}{10 - 1} \\ &= \frac{313.25796}{9} \\ &= 34.80644 \\ s &= \sqrt{34.80644} \\ &= 5.8996981 \end{aligned}$$

$$\boxed{s = 5.90}$$

1.b) Identifying outliers of the dataset

Quartiles are three summary measures that splits the dataset into four equal parts (Dukkipati,2013).

First quartile (Q1) is the 25th percentile of the dataset.

Third Quartile (Q3) is the 75th percentile of the dataset.

We will use an exclusive method to identify Q1 and Q3 quartiles i.e. the median is excluded from our computing.(Bhandari,2021)

$$5.78, 6.71, \underbrace{6.84}_{\text{Q1}}, 7.23, 8.20 \mid 9.65, 13.44, \underbrace{14.71}_{\text{Q3}}, 16.39, 24.37$$

$$\boxed{Q1 = 6.84} \tag{17}$$

$$\boxed{Q3 = 14.71} \tag{18}$$

Second Quartile (Q2) is the Median of the dataset. From equation (16) we have,

$$\boxed{Q2 = 8.93} \tag{19}$$

Interquartile is the difference between the third and first quartiles i.e.

$$\begin{aligned} IQR &= Q3 - Q1 \\ &= 14.71 - 6.84 \end{aligned}$$

$$\boxed{IQR = 7.87} \tag{20}$$

In order to determine the outliers, we calculate the lower and upper boundaries (UoL, 2025).

$$\begin{aligned} Lower\ limit &= Q1 - 1.5 \times IQR \\ &= 6.84 - 1.5 \times 7.87 \\ &= -4.965 \end{aligned}$$

$$\begin{aligned} Upper\ limit &= Q3 + 1.5 \times IQR \\ &= 14.71 + 1.5 \times 7.87 \\ &= 26.515 \end{aligned}$$

This means all values in out dataset must fall within the limits -4.965 to 26.515

Our dataset is:

$\{\ 5.78, 6.71, 6.84, 7.23, 8.20, 9.65, 13.44, 14.71, 16.39, 24.37\ \}$

Therefore, we can conclude that there are no outliers in this dataset.

2.) Calculating the number of students that weigh between 65kg to 75kg.

The given details are:

$$\mu = 72 \tag{21}$$

$$\sigma = 10 \tag{22}$$

$$Total\ no.\ of\ students = 500 \tag{23}$$

The z-score (or z-statistic) is a measure of the number of standard deviations away from a mean for a data point. (UoL,2025)

It is given by,

$$z = \frac{x - \mu}{\sigma} \tag{D}$$

Since the weights are normally distributed, calculating the number of students that weigh between 65kg and 75kg as below:

Using formula (D) for calculating for 65kgs:

$$\begin{aligned} z_1 &= \frac{65 - 72}{10} \\ &= \frac{-7}{10} \\ &= -0.7 \end{aligned}$$

$$Area\ under\ normal\ distribution\ curve\ for\ z_1 : Area(P \leq -0.7) = 0.2420 \tag{24}$$

Calculating for 75kgs:

$$\begin{aligned} z_2 &= \frac{75 - 72}{10} \\ &= \frac{3}{10} \\ &= 0.3 \end{aligned}$$

$$Area\ under\ normal\ distribution\ curve\ for\ z_2 : Area(P \leq 0.3) = 0.6179 \tag{25}$$

Calculating the probability of weight being in between 65kg and 75kg using (24) and (25):

$$\begin{aligned} P(-0.7 \leq Z \leq 0.3) &= 0.6179 - 0.2420 \\ &= 0.3759 \end{aligned}$$

which also means,

$$P(65 \leq X \leq 75) = 0.3759 \tag{26}$$

Calculating the number of students whose weight is within 65kg and 75ks:

$$Number\ of\ students = 500 \times 0.3759 = 187.95$$

Therefore, approximately:

$$\boxed{No.\ of\ students\ having\ weight\ between\ 65kg\ and\ 75ks\ \approx 188}$$

Question 3

1. For a probability distribution shown in table below:

X	8	12	16	20	24
P(X)	3/8	1/12	1/8	1/4	1/6

Find the expected value of X and variance of X

2. 320 students in a school were asked about their favorite sports. It was found that 185 students like tennis, 65 students like cycling, 58 students like swimming and 12 students like football. According to national statistics, the numbers should be in proportion 9:3:3:1. The table below shows the empirical distribution of the numbers. Complete the table by filling in the values in the row of "Expected frequency"

	Tennis (i=1)	Cycling (i=2)	Swimming (i=3)	Football (i=4)
Observed frequency (	185	65	58	12
Expected frequency				

3. Calculate the value of Chi-squared ( ) and give all values to one decimal place. Does the observed distribution differ significantly from the expected distribution, using a significance level of 0.05? Table 2 shows the Chi-square distribution, where the areas given across the top of the table are the areas to the right of the critical value

df	0.995	0.99	0.975	0.95	0.90	0.10	0.05	0.025
1	---	---	0.001	0.004	0.016	2.706	3.841	5.024
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736

Figure 2: Chi-squared distribution

Provide appropriate justification and explanation to all your answers, detailing the methods used.

Solution:

The given probability distribution is discrete. The mean or expected value of X denoted by $E(x)$, for a discreate random variable is given by (Dukkipati,2013),

$$E(x) = \mu = \sum_{all\ i} x_i \cdot P(x_i) \tag{A}$$

Variance is a measure of how distributed or disperesed the data points are from the mean i.e. whether they are clustered near average or spread out (GG,2025).

The variance of X denoted by σ^2 or $V(x)$ is defined as,

$$\sigma^2 = V(x) = (\sum_x x^2 \cdot P(x)) - \mu^2 \tag{B}$$

Calculating for all given values of X and $P(x)$ in table below:

x	$P(x)$	$xP(x)$	$x^2P(x)$
8	3/8	3	24
12	1/12	1	12
16	1/8	2	32
20	1/4	5	100
24	1/6	4	96
Total	1	15	264

Table 2: Probability Distribution breakdown

Formula (A) is used in *Table* 2. column 3. Hence we have the expected value of x as,

$$\boxed{E(x) = \mu = 15} \tag{27}$$

Calculating variance of the given probability distribution by substituting (27) and *Table* 2. column 4 summation into formula (B) as below:

$$\begin{aligned} \sigma^2 = V(x) &= 264 - (15^2) \\ &= 264 - 225 \\ \boxed{\sigma^2 = 39} & \end{aligned} \tag{28}$$

2) Calculating Expected Frequencies from Observed Frequencies

For the given problem statement, we follow the ratio-based expected frequency approach. Expected frequency for a test of independence is given by (Dukkipati,2013),

$$E = \frac{(row\ total) \cdot (column\ total)}{sample\ size} \tag{C}$$

Given proportial ratio is 9 : 3 : 3 : 1.

Category	Tennis (i=1)	Cycling (i=2)	Swimming (i=3)	Football (i=4)	Total
Observed frequency O_i	185	65	58	12	320

Table 3: Given Observed frequencies

The ratio represents how many parts of the total i.e. 320, that each category represents. By computing the sum of all parts, the expected values are scaled in proportion to the actual number of observations i.e. 320 students. (EBSCO,2022)

$$\begin{aligned} Total\ parts &= 9 + 3 + 3 + 1 \\ &= 16 \end{aligned} \tag{29}$$

Applying the formula (C) to the problem statement, we get,

$$Expected\ Frequency,\ E = \frac{No.\ of\ parts \times Total\ no.\ of\ students}{Total\ parts} \tag{30}$$

Using the values from the $Total$ column in $Table.$ 3, we calculate the expected frequencies as below:

Category	Tennis (i=1)	Cycling (i=2)	Swimming (i=3)	Football (i=4)
Observed frequency O_i	185	65	58	12
No. of parts	9	3	3	1
Expected frequency E_i	$\frac{9\times320}{16} = 9 \times 20$	$\frac{3\times320}{16} = 3 \times 20$	$\frac{3\times320}{16} = 3 \times 20$	$\frac{1\times320}{16} = 20$

Table 4: Calculation of Expected Frequencies

Therefore we get the final values as below:

Category	Tennis (i=1)	Cycling (i=2)	Swimming (i=3)	Football (i=4)
Observed frequency O_i	185	65	58	12
Expected frequency E_i	180	60	60	20

Table 5: Contingency Table - Expected Frequency corresponding to Observed frequency

3) Calculating value of Chi-squared χ^2

We have a contingency table *table* 5.

Let,

H_0 : observed distribution is same as expected distribution

H_a : observed distribution differs from expected distribution

We first calculate Degree of Freedom df. Degree of freedom helps to determine the number of elements in a sample that can fluctuate independently.

A null hypothesis test i.e. test of independence is used to prove that two attributes of a population are not related. The degree of freedom for such a test is given by (Dukkipati,2013),

$$df = (R-1)(C-1) \tag{D}$$

where R is number of rows in a contingency table

C is number of columns in a contingency table

From *Table* 5. we have,

$$\begin{aligned} R &= 2 \\ C &= 4 \\ df &= (2-1)(4-1) \\ df &= 3 \end{aligned} \tag{31}$$

Referring to the Chi-square distribution table, the critical value cv, corresponding to $df = 3$ and the given significance level of 0.05 is,

$$cv = 7.8 \tag{32}$$

The next step is to calculate the χ^2 value. Dukkipati states that the value of test statistic χ^2 for a test of independence can be acquired from (2013):

$$\chi^2 = \sum \frac{(O-E)^2}{E} \tag{E}$$

where O is Observed Frequency for a cell, and

E is Expected Frequency for a cell.

Using corresponding values for both O and E from *Table* 5. in formula (E), we get,

$$\begin{aligned} \chi^2 &= \frac{(185-180)^2}{180} + \frac{(65-60)^2}{60} + \frac{(58-60)^2}{60} + \frac{(12-20)^2}{20} \\ &= \frac{5^2}{180} + \frac{5^2}{60} + \frac{(-2)^2}{60} + \frac{(-8)^2}{20} \\ &= \frac{25}{180} + \frac{25}{60} + \frac{4}{60} + \frac{64}{20} \\ \chi^2 &= 3.8 \end{aligned} \tag{33}$$

Comparing values from (32) and (33), we observe that, $\chi^2 < cv$ i.e. $3.8 < 7.8$

Hence, we fail to reject the null hypothesis. Thus we can conclude that the observed distribution does not differ from the expected distribution at significance level of 0.05.

Question 4

4. A random sample of 10 students in a university were taken for the measurement of their body height. The results in the unit of centimeter are 172.35, 174.37, 184.38, 164.36, 154.39, 174.34, 170.33, 175.35, 174.36, 177.37. Give your answers to 3 decimal places for the questions below.

a. Calculate the mean and standard deviation of the students' height from the above sample

b. Construct a 95% confidence interval for the mean of the students' height. Table 3 shows the t-distribution

t Table

cum. prob	$t_{.50}$	$t_{.75}$	$t_{.80}$	$t_{.85}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$	$t_{.999}$	$t_{.9995}$
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073

Figure 3: t-distribution

c. Suppose the average body height in the whole university 170.20cm. Through hypothesis testing method, assess whether this sample result suggest that average body height of the sample students is different from that of the general student population in the university. Use the significance level $\alpha = 0.05$

Provide appropriate justification and explanation to all your answers, detailing the methods used.

Solution:

4.a) Calculating sample mean and standard deviation

As per the definition of sample Mean mentioned earlier in *Question* 2 formula (B), the sample mean is given by,

$$\bar{x} = \frac{\sum X}{n}$$

where $\sum X$ and n are the sum of all variables and sample size respectively.

Applying this formula to the given dataset, we get,

$$\bar{x} = \frac{172.35 + 174.37 + 184.38 + 164.36 + 154.39 + 174.34 + 170.33 + 175.35 + 174.36 + 177.37}{10}$$
$$= \frac{1721.6}{10}$$

$$\boxed{\bar{x} = 172.16} \tag{34}$$

Similarly, referring to the formula (C) for sample standard deviation s in *Question* 2, the standard deviation is given by,

$$s = \sqrt{sample\ variance} = \sqrt{s^2}$$
$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

where $(x - \bar{x})$ is the deviation of x from mean value.

Using this formula to calculate variance s^2 for the given dataset,

$$s^2 = \frac{(172.35 - 172.16)^2 + (174.37 - 172.16)^2 + ... + (174.36 - 172.16)^2 + (177.37 - 172.16)^2}{10 - 1}$$
$$s^2 = \frac{581.123}{9}$$
$$= 64.569$$
$$s = \sqrt{64.569}$$

$$\boxed{s = 8.036} \tag{35}$$

4.b) Calculating confidence interval

Step i: Computing SE

Dukkipati states that a standard error of a statistic is the standard deviation of its sampling distribution (2013). In other words, a Standard Error (SE) is a measure of how much the average of a sample varies from the average of an entire given population.

SE is given by,

$$SE = \frac{s}{\sqrt{n}} \tag{A}$$

Calculating SE with formula (A) and s from (35):

$$SE = \frac{8.036}{\sqrt{10}}$$
$$= 2.541 \tag{36}$$

Step ii: Computing t-score

In hypothesis testing, significance level is represented by alpha, α (UoL, 2025). It is an acceptable section of distribution outside the confidence interval. So when we take the equation $1-\alpha$, we calculate the probability of a parameter being within the interval (SP,2015) i.e. we accept a certain percentage of chance that the interval doesn't contain the parameter.

Calculating α for 95% confidence interval as below,

$$\begin{aligned} \alpha &= 1-\alpha \\ &= 1-0.95 \\ \alpha &= 0.05 \end{aligned} \tag{37}$$

Next, we find degrees of freedom. "Degrees of freedom are the maximum number of logically independent values which may vary in a data sample."(Ganti,2024)

It is given by,

$$\begin{aligned} degrees\ of\ freedom, \nu &= n-1 \\ &= 10-1 \\ \nu &= 9 \end{aligned} \tag{38}$$

Now we calculate t-score. "A $t-score$ (or $t-value$) is the number of standard deviations from the mean in a $t-distribution$" (UoL,2025). It is an indicator of how much the mean of sample population deviates from the mean of entire population.

$t-score$ is calculated as below (UoL,2025):

Critical values for the tails of a t-distribution is denoted by $t_{\frac{\alpha}{2}}$

Caculating the $t-score$ with the degree of freedom is given by,

$$t_{\frac{\alpha}{2}}(n-1) = t_{\frac{\alpha}{2}}\nu \tag{B}$$

Substituting values of α and ν from (37) and (38) respectively, into (B), we get,

$$t_{\frac{0.05}{2}}(9) = t_{0.025}(9)$$

Referring to $t-distribution$ table, we get,

$$t_{0.025}(9) = 2.262 \tag{39}$$

Step iii: Finding confidence interval

A confidence interval is an interval estimate for a population parameter. Meaning, an interval is built around a particular point estimate, followed by making a probabilistic statement that this particular interval consists of the corresponding population parameter. (Dukkipati, 2013)

Confidence interval for $\bar{x}$ is specified with a lower limit and upper limit, and these limits can be calculated using (UoL,2025),

$$\bar{x} \pm t_{\frac{\alpha}{2}} \times SE \tag{C}$$

where $\bar{x}$ is sample mean.

Using (C) and values from (34),(36), (39)

$$\begin{aligned} Confidence\ interval : & (\bar{x} - t_{\frac{\alpha}{2}} \times SE, \bar{x} + t_{\frac{\alpha}{2}} \times SE) \\ & (172.16 - 2.262 \times 2.541, 172.16 + 2.262 \times 2.541) \\ & (166.412, 177.908) \end{aligned} \tag{40}$$

$$\boxed{Confidence\ interval = (166.412, 177.908)}$$

4.c) Using Hypothesis testing to evaluate body height of students

"Hypothesis is a claim about a population parameter. Hypothesis test is a formal procedure to check if the Hypothesis it true or not." (W3S,2025)

Since we have the mean of entire population μ, we employ the method for Hypothesis test for population mean (UoL,2025).

Let,

$H_0 : \mu = 170.20$ (null hypothesis i.e. average body height of students in the sample is same as average body height of general student population)

$H_a : \mu \neq 170.20$ (alternative hypothesis i.e. average body height of students in the sample is different as compared to average body height of general student population)

Sample Mean of the population $\bar{x}$ is calculated in (34), $\bar{x} = 172.16$

Sample stadard deviation s is calculated in (35), $s = 8.036$

Step i: Calculating Test statistic:

"The test statistic is a standardized value calculated from the sample." (W3S,2025). It helps us to understand how much the sample statistic deviates from the null hypothesis parameter.

Test statistic t is given by,

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} \tag{D}$$

Substituting values from (34), (35), the given μ and sample size n in formula (D), we get,

$$\begin{aligned} t &= \frac{172.16 - 170.20}{\frac{8.036}{\sqrt{10}}} \\ &= \frac{1.96}{2.541} \\ t &= 0.771 \end{aligned} \tag{41}$$

Step ii: Calculate p-value:

$p - value$, also called as probability value, indicates the probability of the occurance of an event. Larger the $p - value$, stronger is the evidence in favour of the null hypothesis. (CueMath,2025)

For degrees of freedom $\nu = 9$ from (38), we determine the $p - value$ for two tails from $t - distribution\ table$ as below,

$$p - value : \ 0.4603 \tag{42}$$

The given significance level $\alpha = 0.05$.

We see $0.4603 > 0.05$ i.e. $p - value > significance\ level$. Therefore statistical analysis result fails to reject the null hypothesis i.e. the evidence from sample data is inadequate to conclude that the average body height of students is different as compared to the general student population.

Question 5

5. Several years ago, 39% of students in a university were satisfied with the quality of education they received. A recent survey asked 1065 students and 459 of them indicated that they were satisfied. Give your answers to 3 decimal places for the questions below.

a. Find and interpret a 99% confidence interval for the proportion of student population who are satisfied with the quality of education in the recent poll

b. Through hypothesis test method, assess whether students' attitudes toward the quality of education have changed in the recent poll. Use the significance level $\alpha = 0.01$

Solution:

5.a) Calculate confidence interval

We have Categorical data, where the observations are binary i.e. they either fall within the area of interest or outside that area (UoL,2025). We follow below steps to construct a confidence interval for estimating the given population proportion.

Step i: Evaluating population proportion using sample proportion:

Sample proportion is calculated as,

$$\hat{p} = \frac{x}{n} \tag{A}$$

where x is number of successes in the sample, and

n is sample size

From the problem statement, we have $x = 459$ and $n = 1065$.

Substituting these values in formula (A), sample proportion is calculated as,

$$\begin{aligned} \hat{p} &= \frac{459}{1065} \\ &= 0.431 \end{aligned} \tag{43}$$

Step ii: Calculating Standard error SE:

"SE for a population proportion is an estimated standard deviation of the sample distribution" (UoL,2025). It is given by,

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{B}$$

Using formula (B) and substituting (43) , we calculate SE as,

$$\begin{aligned} SE &= \sqrt{\frac{0.431(1-0.431)}{1065}} \\ &= \sqrt{\frac{0.245}{1065}} \\ SE &= 0.015 \end{aligned} \tag{44}$$

Step iii: Calculating confidence interval:

Given 99% confidence interval, calculating α as

$$\begin{aligned} \alpha &= 1-\alpha \\ &= 1-0.99 \\ \alpha &= 0.01 \end{aligned} \tag{45}$$

Since the sample size is more than 30 (Shubh,2025), we calculate the $.z-score$.

$$\begin{aligned} z-score &= z_{\frac{\alpha}{2}} \\ &= z_{\frac{0.01}{2}} \\ &= z_{0.005} \end{aligned} \tag{46}$$

Referring to the $z-table$, we get,

$$z_{\frac{\alpha}{2}} = 2.5758 \tag{47}$$

Calculating confidence interval as,

$$\begin{aligned} Confidence\ Interval : &(\hat{p} - z_{\frac{\alpha}{2}} \times SE, \hat{p} + z_{\frac{\alpha}{2}} \times SE) \\ &= (0.431 - 2.5758 \times 0.015, 0.431 + 2.5758 \times 0.015) \\ &= (0.392, 0.4696) \end{aligned} \tag{48}$$

$$\boxed{Confidence\ Interval : (0.392, 0.470)}$$

5.b) Hypothesis Test

Since we have the proportions of data, we employ the method for calculating Hypothesis test for population proportion (UoL,2025).

Let,

$H_0 : p_o = 0.39$ (null hypothesis i.e. students' attitudes toward the quality of education remains the same)

$H_a : p_o \neq 0.39$ (students' attitudes toward the quality of education have changed)

i) Calculate test statistic

From (43) we have the population proportion.

Test static is a measure of how much the sample proportion $\hat{p}$ deviates from the null hypothesis value p_o (UoL,2025). The test statistic for a population proportion is given by,

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1-p_o)}{n}}} \tag{A}$$

Using this formula and substituting values from (43), given p_o and given n, we get,

$$\begin{aligned} z &= \frac{0.431 - 0.39}{\sqrt{\frac{0.39(1-0.39)}{1065}}} \\ &= \frac{0.041}{0.0149} \\ &= 2.752 \end{aligned} \tag{49}$$

From (47) we have the results of a two tailed test with $\alpha = 0.01$.

Therefore the critical values are : ± 2.576 $(approx)$

From (49) we have $z = 2.752$

Comparing these two values, we have, $2.752 > 2.576$. Therefore we reject the null hypothesis. We can conclude that (at a significance level of 0.01) the hypothesis test results show that the students' attitude towards the quality of education have changed in the recent poll.

References:

Dukkipati, R. V. (2013), Probability and Statistics: For Scientists and Engineers, New Academic Science.

UoL - University of Liverpool

ACP - Algorithms for Competitive Programming (2024). 'The Inclusion-Exclusion Principle'. Available: https://cp-algorithms.com/combinatorics/inclusion-exclusion.html. (Accessed on: 26 May 2025)

CueMath (2025). 'Hypothesis Testing'. Available: https://www.cuemath.com/p-value-formula/. (Accessed on: 26 May 2025)

W3S - W3Schools (2025). 'Statistics - Hypothesis Testing'.

Available: https://www.w3schools.com/statistics/statistics_hypothesis_testing.php. (Accessed on: 26 May 2025)

UoND - University of Notre Dame (2025). 'Confidence Intervals'. Available:https://www3.nd.edu/ rwilliam/stats1/x23.pdf. (Accessed on: 26 May 2025)

EBSCO (2022) 'Proportionality (mathematics)'.

Available: https://www.ebsco.com/research-starters/science/proportionality-mathematics. (Accessed on: 26 May 2025)

GG (2025). 'Variance'. Available: https://www.geeksforgeeks.org/variance/. (Accessed on: 26 May 2025)

SM - Survey Monkey (2025) 'What is a margin of error?'.Available: https://www.surveymonkey.com/mp/margin-of-error-calculator/. (Accessed on: 26 May 2025)

SP - StudyPug (2015). 'Understanding Confidence and Significance Levels in Statistics'.

Available:https://www.studypug.com/statistics-help/confidence-levels-significance-levels-and-critical-values. (Accessed on: 26 May 2025)

Ganti, A. (2024). 'Degrees of Freedom in Statistics Explained: Formula and Example'.

Available:https://www.investopedia.com/terms/d/degrees-of-freedom.asp. (Accessed on: 26 May 2025)

Bhandari, P. (2021) 'How to Find Outliers — 4 Ways with Examples Explanation'.

Available: https://www.scribbr.com/statistics/outliers/ (Accessed on: 26 May 2025)

Shubh (2025). 'Difference Between Z-Test and T-Test'.

Available: https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/. (Accessed on: 26 May 2025)