Correlation & Regression

1. A researcher is studying the relationship between hours of study per week and exam scores for a group of students. The data collected from 10 students is as follows:

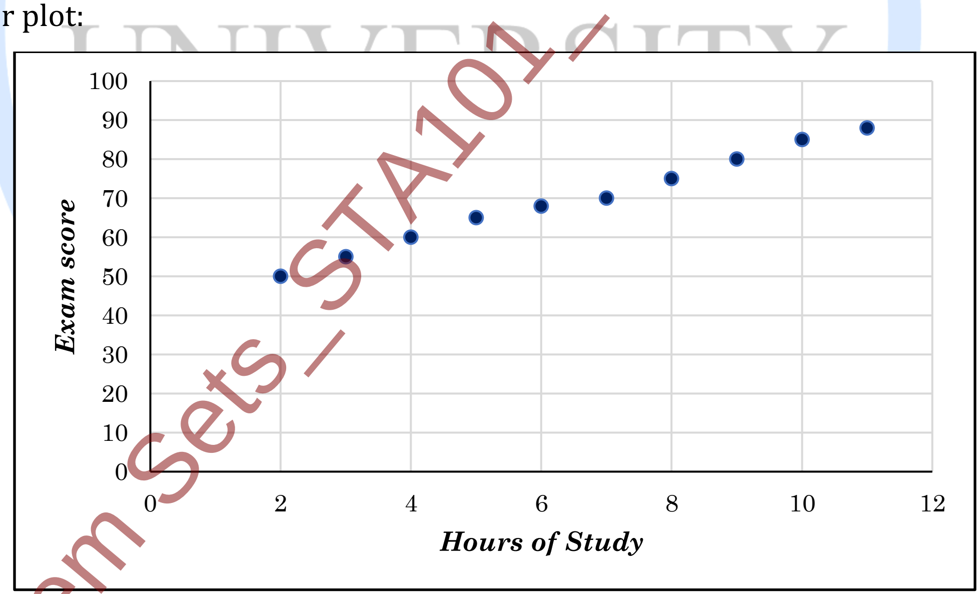
Student ID	1	2	3	4	5	6	7	8	9	10
Hours of Study	5	7	10	3	8	6	9	4	2	11
Exam Score	65	70	85	55	75	68	80	60	50	88

a) Determine the direction of association using appropriate graphical method.

b) Are "Hours of study" and "Exam score" correlated? If yes, comment about the strength of their relationship.

Solution:

a) Scatter plot:



Comment:

b) From graph it is clear that "Hours of study" and "Exam score" are correlated. For comment about the strength of their relationship, we have to calculate the Karl Pearson's Correlation Coefficient.

Here, X = Hours of Study, Y = Exam score

$$\sum X = 65, \sum Y = 696, \sum X^2 = 505, \sum Y^2 = 49868, \sum XY = 4866$$

$$\therefore r = 0.997$$

2. A researcher collects data on two variables: annual rainfall (in cm) and crop yield (in tons per hectare) over a 10-year period from a specific region.

Year	1	2	3	4	5	6	7	8	9	10
Rainfall	120	135	150	115	160	145	110	170	105	130
Crop yield	4.8	5.2	6.0	4.5	6.3	5.8	4.2	6.5	4.0	5.0

a) Apply appropriate correlation measures method to determine the strength of association between Rainfall and Crop yield. Is the relationship between rainfall and crop yield strong or weak?

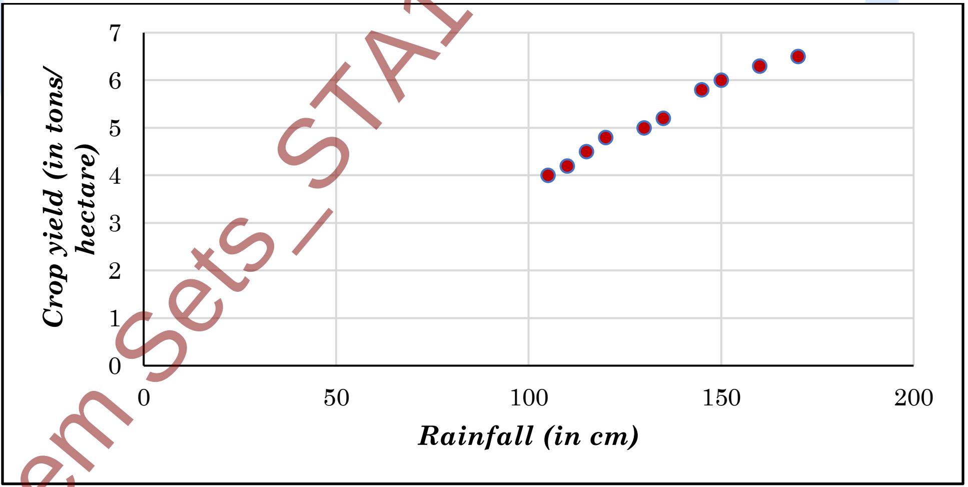
b) Create a scatter plot for the given data. Does the scatter plot suggest the same relationship direction as the you got in (a)?

c) If the correlation is strong, does it imply that an increase in rainfall will always lead to an increase in crop yield? Can we assume causality from correlation in this case?

Solution:

a) Hints: Karl Pearson's correlation coefficient, $r = 0.99$

b) Scatter plot:



Comment:

c) Even if the correlation coefficient is strong, it does not imply causality. While more rainfall generally corresponds to higher crop yield, other factors like soil quality, farming techniques, and pest control could affect the crop yield. Correlation only measures the linear relationship, not cause and effect.

3. A scientist is analyzing the relationship between temperature (in degrees Celsius) and electricity consumption (in kilowatt-hours) in a residential area over 12 days. The aim is to determine how changes in temperature affect electricity usage. To achieve this aim, he did bellow calculation:

$$\sum Temperature = 272; \sum Electricity\ Consumption = 4630$$

$$\sum (Temperature)^2 = 6876; \sum (Electricity\ Consumption)^2 = 1835100$$

$$\sum (Temperature \times E.Consumption) = 110790$$

a) Calculate correlation coefficient using above information.

b) Suppose 12th value of Temperature is 33 and Electricity consumption is 470. After calculating correlation coefficient, the scientist notices that the 12th value of electricity consumption was misreported and should be 490. Calculate the corrected correlation coefficient.

c) Using above information, show that correlation coefficient is a symmetric measure.

Solution:

a) Try yourself.

b) Corrected correlation coefficient:

Here, X = Temperature, Y = Electricity consumption

Before correction	**After correction**
$\sum X = 272$	$\sum X = 272$
$\sum Y = 4630$	$\sum Y = (4630 - 470) + 490 = ???$
$\sum X^2 = 6876$	$\sum X^2 = 6876$
$\sum Y^2 = 1835100$	$\sum Y^2 = (1835100 - 470^2) + 490^2 = ???$
$\sum XY = 110790$	$\sum XY = \left(110790 - (33 \times 470)\right) + (33 \times 490) = ???$

[Using these corrected values, calculate the corrected correlation coefficient]

c) Hints: Show that, $r_{xy} = r_{yx}$

4. A water resources engineer is investigating how annual rainfall (in mm) affects the water level (in meters) of a major river in Bangladesh over 10 years. The engineer collects data from different hydrological stations located near the river to understand the impact of varying rainfall on the river's water level. The data collected is as follows:

Year	1	2	3	4	5	6	7	8	9	10
Water level	8.5	9.0	9.5	8.7	9.2	10.0	8.0	10.5	9.8	11.0
Rainfall	1200	1300	1400	1250	1350	1500	1100	1600	1450	1700

a) Determine how strongly and in which direction annual rainfall (in mm) is related to the water level (in meters). [$Ans: r = 0.9995$]

b) Fit an appropriate regression model using variable "Annual Rainfall (in mm)" and "Water Level (in meters)". [$Ans: intercept = 2.4417, Slope = 0.005$]

c) Provie interpretation about regression parameters.

d) Predict the water level value when the annual rainfall is 800 mm. [Ans: 6.4417]

e) Does your model predict well? Provide a mathematical explanation. [Ans: $R^2 = 0.999$]

5. A transportation planner is analyzing how traffic volume (in thousands of vehicles) affects the average travel time (in minutes) on a major highway over 10 days. The goal is to understand how an increase in traffic volume impacts travel time and to develop a predictive model to estimate travel times based on traffic data. The data collected is as follows:

Day	1	2	3	4	5	6	7	8	9	10
Traffic Volume	40	45	50	42	48	55	38	60	53	65
Travel Time	30	35	38	32	36	42	28	46	40	50

a) Determine how strongly and in which direction traffic volume (in thousands of vehicles) is related to average travel time (in minutes). [Ans: 0.998]

b) Fit an appropriate regression model using the variables "Traffic Volume (in 000s)" and "Average Travel Time (in minutes)". [Ans: Intercept = -1.7143, Slope = 0.7946]

c) Provide interpretation of regression parameters.

d) Predict the travel time when the traffic volume is 30,000 vehicles. **[Try yourself]**

e) Does your model predict well? Provide a mathematical explanation based on model fit metrics. [Ans: $R^2 = 0.996$]

6. A company maintains a database that tracks employee performance based on two variables: **Hours of training (Training Hours)**: The number of hours an employee spends on training each month. **Monthly Sales (Sales)**: The sales (in dollars) generated by each employee in a month. A random sample of 10 employees provides the following data:

Employee ID	101	102	103	104	105	106	107	108	109	110
Training hours (X)	5	7	6	8	10	4	9	7	6	5
Sales ($) (Y)	3000	3500	3200	4000	4500	2800	4200	3800	3400	3100

a) Calculate the correlation coefficient between the number of training hours and the monthly sales. Interpret the result.

b) Perform a simple linear regression analysis to model the relationship between **Training Hours (X)** and **Sales (Y)**. Use the data provided to find the regression equation.

c) Based on the regression equation, predict the monthly sales for an employee who trains for **8 hours.**

d) Discuss whether the correlation and regression analyses imply a causal relationship between training hours and sales performance.

Solution:

a) **Correlation Coefficient:**

To calculate the correlation coefficient (r), we use the formula:

$$r = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{n}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{n}\right]}}$$

Where: n is the number of data points (10 in this case), X is the number of training hours, Y is the monthly sales. After applying the data, suppose the calculated correlation coefficient is $r = 0.91$. This suggests a strong positive correlation between training hours and monthly sales, indicating that employees who undergo more training tend to generate higher sales.

b) **Simple Linear Regression:**

The formula for the linear regression equation is:

$$Y = b_0 + b_1 X$$

Where: $b_1 = \frac{\Sigma XY - \frac{\Sigma X \Sigma Y}{n}}{\Sigma X^2 - \frac{(\Sigma X)^2}{n}}$ (slope), $b_0 = \bar{Y} - b_1\bar{X}$ (intercept).

After calculation, the regression equation may look like:

$$\text{Sales} = 2000 + 250 \times \text{Training\_Hours}$$

This means that for every additional hour of training, the sales increase by $250, starting from a base sales amount of $2000.

c) **Prediction:** To predict the sales for an employee who trains for 8 hours, plug in $X = 8$ into the regression equation: $\text{Sales} = 2000 + 250 \times 8 = 2000 + 2000 = 4000$

Thus, an employee who trains for 8 hours is predicted to generate $4000 in sales.

7. Imagine a small electronics company in Dhaka that has been tracking its daily sales and profits over a period of 6 consecutive days. The company sells various electronic items like mobile phones, laptops, and accessories. The following are the sales ('000 BDT) and profits ('000 BDT) for the last 6 days:

Day	1	2	3	4	5	6
Sales	6	7	8	11	12	10
Profit	1	1	3	5	6	4

a) Calculate the Pearson's correlation coefficient between sales and profit, and interpret your findings. $[Ans: r = 0.9820]$

b) Use an appropriate graphical method to find the relationship between sales and profit.

c) Fit a least-squares regression line of profit on sales and interpret the regression parameters. $[Intercept = -4.3810, Slope = 0.8571]$

d) If the sale is 15,000 BDT, estimate the profit.

e) Evaluate how well the regression line fits the data using the coefficient of determination. $[R^2 = 0.9643]$

8. A department of transportation's study on driving speed and mileage for midsize automobile resulted in the following table:

Driving speed	30	40	50	55	25
Mileage	27	25	30	35	22

a) Is there any relationship between Driving speed and Mileage? Verify your answer. $[Ans: 0.8879]$

b) Find the regression equation of driving speed on mileage. $[Intercept\ = -23.3097, Slope\ =\ 2.2773]$

c) What will be mileage when speed is 45? **[Try yourself]**

d) Test the fitness of your regression model with explanation. $[R^2\ = 0.7883]$

<u>Basic Concepts of Probability</u>

1. In the Engineering department, the academic paths of 348 students, across different years and specializations, are summarized in the contingency table below:

	Subjects				
	Mechanical	**Electrical**	**Civil**	**CS**	***Total***
1st Year	$6x$	$5x$	$4x$	$3x$	90
2nd Year	35	A	15	B	90
3rd Year	$10y$	$15y$	$12y$	$7y$	88
4th Year	15	15	20	C	
Total		90			

a) Complete the table.

b) Determine the probability (with interpretation) that a randomly selected student is either in the “4th Year” or choosing “Computer Science (CS)”.

<u>Solution:</u>

a) Calculate the value x,

$6x + 5x + 4x + 3x = 90$

$\therefore x = 5$

Calculate the value y,

$10y + 15y + 12y + 7y = 88$

$\therefore y = 2$

Now, we reconstruct the table,

	Subjects				
	Mechanical	**Electrical**	**Civil**	**CS**	***Total***
1st Year	30	25	20	15	90
2nd Year	35	20	15	20	90
3rd Year	20	30	24	14	88
4th Year	15	15	20	30	80
Total	100	90	79	79	348

b) Let,

A = Event of 4th year student

B = Event of student choosing CS

Here, $P(A) = \frac{80}{348}$; $P(B) = \frac{79}{348}$; $P(A \,\&\, B) = \frac{30}{348}$

$$\therefore P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.37\ (Approx.)$$

2. In a survey of business students, respondents were asked about their preferences for studying three major subjects: Marketing (M), Finance (F), and Operations (O). Each student could choose to study one or more subjects. The probabilities for each subject are as follows: The probability of a student choosing Marketing is 60%. The probability of a student choosing Finance is 50%. The probability of a student choosing Operations is 40%. The probability of a student choosing both Marketing and Finance is 30%. The probability of a student choosing both Marketing and Operations is 25%. The probability of a student choosing both Finance and Operations is 20%. The probability of a student choosing all three subjects (Marketing, Finance, and Operations) is 10%.

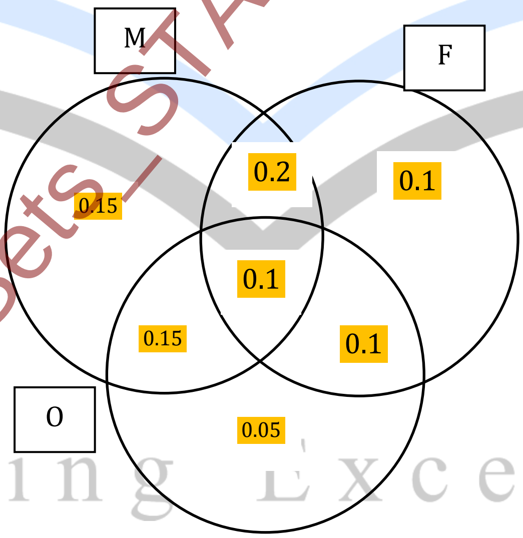
a) Construct a Venn diagram to represent the probabilities.

b) If a business student is selected randomly, what is the probability that he/she is studying:

i. exactly one subjects?
ii. At least one subject?
iii. At most one subject?
iv. No subject?

Solution:

a) Venn Diagram:



b) $P(Exactly\ one\ subject)\ =\ P(Only\ M) + P(Only\ F) + P(Only\ O) = ???$

$P(At\ least\ one) =\ P(M \cup F \cup O) = ???$

$P(At\ most\ one) = P(Only\ M) + P(Only\ F) + P(Only\ O) + P(None) = ???$

$P(No\ subject) = 1 - P(At\ least\ one) = ???$

3. In a certain residential suburb, 60% of all households get Internet service from the local cable company, 80% get television service from that company, and 50% get both services from that company. If a household is randomly selected, what is the probability that it gets at least one of these two services from the company, and what is the probability that it gets exactly one of these services from the company? $[Ans: 0.9;\ 0.4]$

4. A and B are two weak students in Statistics. A can answer correctly 15% of the questions and B can answer correctly 10% of the questions. A and B both can answer 2% of the questions. A question is selected at random. Find the probability that (a) at least one of them can answer correctly, (b) No one can answer correctly, (c) only one can answer correctly.

<u>Solution:</u>

Let us denote the events,

$A: A\ can\ answer\ the\ question\ correctly$

$B: B\ can\ answer\ the\ question\ correctly$

Here, $P(A) = 0.15; P(B) = 0.10; P(A \cap B) = 0.02$

a) $P(at\ least\ one) = P(A \cup B) = P(A) + P(B) - P(A \cap B) = ???$

b) $P(No\ one) = P(\overline{A \cup B}) = 1 - P(A \cup B) = ???$

c) $P(Only\ one) = P(A \cap \bar{B}) + P(B \cap \bar{A}) = ???$

5. In a large corporation, two senior managers, Ms. Smith and Mr. Johnson, are frequently required to travel abroad for business meetings. Ms. Smith travels abroad 70% of the time during a given year, while Mr. Johnson travels abroad 45% of the time during the same period.

a) Calculate the probability that on a randomly selected day in the year, both Ms. Smith and Mr. Johnson will be abroad. $[Ans: 0.315\ or\ 31.5\%]$

b) Determine the probability that either Ms. Smith or Mr. Johnson will be abroad on a given day. $[Ans: 0.835\ or\ 83.5\%]$

c) What is the probability that neither Ms. Smith nor Mr. Johnson will be abroad on a randomly selected day? $Hints: [P(\bar{A} \cap \bar{B}) = 1 - P(A \cup B)]$

6. Tickets are numbered from 1 to 100. They are well shuffled and a ticket is drawn at random. What is the probability that the drawn ticket has

a) An old number

b) A number 4 or multiple of 4

c) A number which is greater than 70

d) A number which is a square?

Solution:

Here are 100 tickets, the total number of exhaustive, mutually exclusive and equally likely cases is 100.

a) Let A denote the event that the ticket drawn has an odd number. Since there 50 odd numbered tickets, the number of cases favorable to the event A is 50,

$$\therefore P(A) = \frac{50}{100} = 0.5$$

Comment:

b) Let B denote the event that the drawn ticket has a number 4 or multiple of 4. The numbers favorable to event "B" are 4, 8, 12, 16, 20 .... 92, 96, 100. The total number of cases will be 25.

$$\therefore P(B) = \frac{25}{100} = \frac{1}{4}$$

Comment:

c) Let C be the event that the drawn ticket has a no greater than 70. Since the No. greater than 70 are 71, 72 ...... 100. Therefore 30 cases are favorable to the event C.

$$\therefore P(C) = \frac{30}{100}$$

Comment:

d) Let D be the event that the drawn ticket has a number which is a square. Since the squares between 1 to 100 are 1, 4, 9, 16, 25, 36, 49, 64, 81, 100.

$$\therefore P(D) = \frac{10}{100}$$

Comment:

7. Afsana feels that the probability that she will pass mathematics is 2/3 and statistics is 5/6. If the probability that she will pass both the course is 3/5, what is the probability that she will pass at least one of the courses? [Ans: 9/10]

8. Suppose A and B are two mutually exclusive events with P(A) = 0.35 and P(B) = 0.15. Find

a) $P(A \cup B)$ [Ans: 0.5]
b) $P(A')$ [Ans: 0.65]
c) $P(A \cap B)$ [Ans: 0]
d) $P(A' \cup B')$ [Ans: 1]
e) $P(A' \cap B')$ [Ans: 0.5]

9. Sample Space Related Questions

a) Write the total number of outcomes in sample space for rolling a die until a 4 appears. **[Ans: Infinity (Depends on when 4 appears)]**
b) What is the sample space for selecting an even number between 1 and 10? [$\boldsymbol{S} = \{\mathbf{2, 4, 6, 8, 10}\}$]
c) Write the sample space for rolling two dice. **[Try yourself]**
d) Write the sample space of toss a coin and rolling a dice simultaneously. **[Try yourself]**
e) Some people are in favor of reducing federal taxes to increase consumer spending and others are against it. Two persons are selected and their opinions are recorded. Assuming no one is undecided, list the possible outcomes. **[$\boldsymbol{S} = \{(\mathbf{F, F}), (\mathbf{F, A}), (\mathbf{A, F}), (\mathbf{A, A})\}$]**
f) A quality control inspector selects a part to be tested. The part is then declared acceptable (A), repairable (R), or scrapped (S). Then another part is tested. List the possible outcomes of this experiment regarding two parts. [$\mathrm{S} = \{(\mathrm{A, A}), (\mathrm{A, R}), (\mathrm{A, S}), (\mathrm{R, A}), (\mathrm{R, R}), (\mathrm{R, S}), (\mathrm{S, A}), (\mathrm{S, R}), (\mathrm{S, S})\}$]

10. In each of the following cases, indicate whether classical, empirical/frequency, or subjective probability is used.

a) A baseball player gets a hit in 30 out of 100 times at bat. The probability is 0.3 that he gets a hit in his next at bat. **[Empirical Probability]**

b) A seven-member committee of students is formed to study environmental issues. What is the likelihood that any one of the seven is randomly chosen as the spokesperson? **[Classical Probability]**

c) You purchase a ticket for the Lotto Canada lottery. Over 5 million tickets were sold. What is the likelihood you will win the $1 million jackpot? **[Classical Probability]**

d) The probability of an earthquake in northern California in the next 10 years above 5.0 on the Richter Scale is 0.8. **[Subjective Probability]**

11. A sample of 40 oil industry executives was selected to test a questionnaire. One question about environmental issues required a yes or no answer.

a) What is the experiment?

Solution: The experiment involves selecting 40 oil industry executives and asking them a yes/no question about environmental issues.

b) List one possible event.

Solution: One possible event is that a randomly selected executive responds "yes" to the question.

c) Ten of the 40 executives responded yes. Based on these sample responses, what is the probability that an oil industry executive will respond yes? $\left[\boldsymbol{Ans: P(Yes) = \frac{10}{40} = 0.25}\right]$

d) What concept of probability does this illustrate? **[This illustrates empirical probability because it is based on observed data from the sample.]**

e) Are each of the possible outcomes equally likely and mutually exclusive?

Solution:

Equally Likely: No, the outcomes are not equally likely because more respondents may answer "no" than "yes," as evidenced by the data.

Mutually Exclusive: Yes, the outcomes are mutually exclusive because a respondent can only provide one answer: either "yes" or "no" (not both).

12. Consider the experiment of rolling a pair of dice. Suppose that we are interested in the sum of the face values showing on the dice.

a) How many sample points are possible? **[Try yourself]**

b) List the sample points. **[Try yourself]**

c) What is the probability of obtaining a value sum of 7? **[Try yourself]**

d) What is the probability of obtaining a sum of 9 or greater? **[Try yourself]**

e) What method did you use to assign the probabilities requested? **[Classical Probability was used]**

13. The events A and B are mutually exclusive. Suppose P(A) = 0.30 and P(B) = 0.20. What is the probability of either A or B occurring? What is the probability that neither A nor B will happen? [*Ans*: 0.5; 0.5]

14. The probabilities of the events A and B are 0.20 and 0.30, respectively. The probability that both A and B occur is 0.15. What is the probability of either A or B occurring? **[Try yourself]**

15. A student is taking two courses, history and math. The probability the student will pass the history course is 0.60, and the probability of passing the math course is 0.70. The probability of passing both is 0.50. What is the probability of passing at least one? **[Try yourself]**

16. A survey by the American Automobile Association (AAA) revealed 60% of its members made airline reservations last year. Two members are selected at random. What is the probability both made airline reservations last year? [*Ans*: 0.36]

Conditional Probability

1. Suppose a balanced die is rolled once.

a. Find the probability that a number divisible by 3 is rolled given that the die comes up even.

b. Find the probability that the die comes up even given that a number divisible by 3 is rolled. [Ans: 0.5]

c. Find the probability that a number divisible by 3 is rolled given that die comes up at most 4. [Ans: 0.25]

d. Find the probability that the die comes up at most 4 given that a number divisible by 3 is rolled. [Ans: 0.5]

Solution:

a) Let,

D3 = Event of number division by 3 = {3,6}

E = Event of even number = {2,4,6}

Here, the sample space, $S = \{1,2,3,4,5,6\}$

$$P(D3) = \frac{2}{6}; P(E) = \frac{3}{6}; P(D3 \cap E) = \frac{1}{6}$$

$$\therefore P(D3|E) = \frac{P(D3 \cap E)}{P(E)} = 0.333$$

Comment: Given that die comes up even, there is a 33.3% chance of rolling a number divisible by 3.

b) Try yourself

c) Try yourself

d) Try yourself

2. In a certain coastal city, during the hurricane season, it is known that hurricanes occur on 60% of the days. When a hurricane occurs, there is an 85% chance that it will also lead to heavy rainfall. Additionally, it is observed that heavy rainfall occurs on 50% of the days during the hurricane season.

a) Calculate the probability that on a given day during the hurricane season, both a hurricane and heavy rainfall will occur.

b) Determine the probability that a hurricane has occurred given that heavy rainfall was observed on a particular day.

c) Determine the probability that there is no heavy rainfall on a given day given that a hurricane has occurred. [$Hints: P(R'|H) = 1 - P(R|H)$]

Solution:

a) Let,

H = Event of hurricanes occur

R = Event of heavy rainfall

$$P(H) = 0.6; P(R|H) = 0.85$$

$$\therefore P(R \cap H) = P(R|H) \times P(H) = ???$$

b) Try yourself.

c) Try yourself.

3. In a game of chance, you roll two six-sided dice. One die is red and the other is blue.

a) What is the probability that the sum of the numbers rolled is 7 given that the number on the red die is 4? [Ans: 1/6]

b) What is the probability that the number on the blue die is 5 given that the sum of the numbers rolled is 9? [Ans: 1/4]

c) What is the probability of rolling a sum less than or equal to 6 given that the number on the red die is even? [Ans: 1/3]

d) What is the probability that the number on the red die is odd given that the sum of the numbers rolled is 8? [Ans: 2/5]

4. A software company employs 80 developers. Out of these, 50 developers are proficient in **JavaScript** (JS), 40 are proficient in **Python**, and 20 are proficient in **both** JavaScript and Python.

a) **What is the probability** that a randomly selected developer is proficient in either JavaScript or Python?

b) If a developer is known to be proficient in Python, **what is the probability** that they are also proficient in JavaScript (i.e., conditional probability)?

c) **Are the events** "Proficient in JavaScript" and "Proficient in Python" independent? Explain your reasoning using probability concepts.

Solution:

a) Probability of a developer being proficient in JavaScript or Python:

We can use the formula for the union of two events:

$$P(\text{JS} \cup \text{Python}) = P(\text{JS}) + P(\text{Python}) - P(\text{JS} \cap \text{Python})$$

Where:

- $P(\text{JS}) = \frac{50}{80} = 0.625,$
- $P(\text{Python}) = \frac{40}{80} = 0.5,$
- $P(\text{JS} \cap \text{Python}) = \frac{20}{80} = 0.25.$

So, $P(\text{JS} \cup \text{Python}) = 0.625 + 0.5 - 0.25 = 0.875$

The probability that a developer is proficient in either JavaScript or Python is **0.875**.

b) Conditional Probability: Given that the developer is proficient in Python, what is the probability that they are also proficient in JavaScript?

The conditional probability formula is:

$$P(\text{JS}|\text{Python}) = \frac{P(\text{JS} \cap \text{Python})}{P(\text{Python})}$$

Substituting the values:

$$P(\text{JS}|\text{Python}) = \frac{0.25}{0.5} = 0.5$$

So, the probability that a developer is proficient in JavaScript given that they are proficient in Python is **0.5**.

c) **Independence Check**:

Two events A and B are independent if:

$$P(A \cap B) = P(A) \times P(B)$$

Here:

$$P(\text{JS}) \times P(\text{Python}) = 0.625 \times 0.5 = 0.3125$$

Since $P(\text{JS} \cap \text{Python}) = 0.25$, which is not equal to 0.3125, the events "Proficient in JavaScript" and "Proficient in Python" are **not independent**.

5. In rolling two balanced dice, if the sum of the two volumes is 8 what is the probability that one of the values is 3?

Solution: Let A be the event that one of the values is 3. Let B be the event that the sum is 8. If we consider the sample space of the experiment consisting of 2 dice, the event AB consists of the sample points (3, 5) and (5, 3). The event B consists of the sample points (2, 6), (3,5), (4,4), (5, 3) and (6, 2). Now,

$$P(AB) = \frac{2}{36}; P(B) = \frac{5}{36}$$

$$\therefore P(A|B) = \frac{P(AB)}{P(B)} = ???$$

6. Two balanced dice, one black and one red are thrown and the number of dots on their upper faces are noted, let b be the outcomes of the black die and r be the outcomes of the red die. Answer the following:

a) List a sample space of the experiment.

b) What is the probability that $r > 4$ and $b \leq 5$? [Ans: 0.27]

c) What is the probability that the difference of the two dice is less than three? [Ans: 0.66]

7. The probability that a person picked at random from a population will exhibit the symptom of certain disease is 0.2, and the probability that a person picked at random has the disease is 0.23. the probability that a person who has the symptom also has the disease is 0.18. A person selected at random from the population does not have the symptom. What is the probability that the person has the disease? $[Hints: P(D|S'), Ans: 0.0625]$

8. A particular medicine was given to a group of people for a specific disease.

	Cured	**Not cured**
Male	20	15
Female	25	10

One person is selected. What is the probability that,

a) The person is cured? [Ans: 45/70]

b) The person is male? [Ans: 35/70]

c) The person is male and cured? [Ans: 20/70]

d) The person is female and cured? [Ans: 55/70]

e) The person is cured given that the person is male? [Ans: 20/35]

f) The person is cured or not cured? [Ans: 1]

9. Two students A and B are asked to develop a computer programme. Previous knowledge tells that A becomes successful in 60% cases and B becomes successful in 70% cases. If they work independently. What is the probability that

a) The programme will be developed?

b) None will be successful?

c) A will be successful under the condition that B fails?

Solution: P(A) = 0.6, P(B) = 0.7. As they work independently, $P(A \cap B) = 0.6 \times 0.7 = 0.42$

a) $P(A \cup B) = 0.88$

b) $P(A' \cap B') = 1 - P(A \cup B) = 0.12$

c) $P(A|B') = 0.6$

With replacement and Without replacement

1. A bag contains 5 red balls and 3 green balls. You are asked to draw 2 balls from the bag one after another, with replacement (meaning after drawing a ball, you put it back in the bag).

a) What is the probability of drawing two red balls consecutively?

b) What is the probability of drawing one red ball and one green ball in two draws, in any order?

Solution:

a) Probability of drawing a red ball on the first draw is 5/8.

Sine the ball is replaced, the probability of drawing a red ball in the second draw is still 5/8. Now,

$$P(2\ red\ balls) = \frac{5}{8} \times \frac{5}{8} = \frac{25}{64}$$

So, the probability of drawing two red balls consecutively with replace is 25/64.

b) There are two possible favorable outcomes:

Red on the first draw, green on the second draw.

Green on the first draw, red on the second draw.

$$\therefore P(Red\ first, Green\ Second) = \frac{5}{8} \times \frac{3}{8} = \frac{15}{64}$$

And,

$$P(Green\ first, Red\ second) = \frac{3}{8} \times \frac{5}{8} = \frac{15}{64}$$

Now,

$$P(1\ red, 1\ green\ in\ any\ order) = \frac{15}{64} + \frac{15}{64} = \frac{15}{32}$$

So, the probability of drawing one red ball and one green ball in two draws is 15/32.

2. A jar contains 5 blue marbles, 4 green marbles, and 3 red marbles. You are asked to draw 3 marbles one after another, without replacement. What is the probability that all three marbles are blue?

Solution:

Probability of drawing a blue marble on the first draw is 5/12

After drawing one blue marble, there are 4 blue marbles left and 11 marbles in total. The probability of drawing a second blue marble is 4/11.

After drawing two blue marbles, there are 3 blue marbles left and 10 marbles in total. The probability of drawing a third blue marble is 3/10.

Now, the probability that all three marbles are blue,

$$P(3\ blue\ marbles) = \frac{5}{12} \times \frac{4}{11} \times \frac{3}{10} = \frac{1}{22}$$

Alternative:

Selecting 3 marbles from 12 in 12C3 ways.

Selecting 3 blue marbles from 5 blue in 5C3 ways.

$$\therefore \boldsymbol{P(3\ blue\ marbles) = \frac{5C3}{12C3} = \frac{1}{22}}$$

3. A box contains seven balls- 2 red, 3 blue, and 2 yellow. Consider an experiment that consists of drawing a ball from the box.

a) What is the probability that the first ball drawn is yellow?

b) What is the probability that the same-colored ball is drawn twice without replacement?

c) What is the probability that the same-colored ball is drawn twice with replacement? **[Ans: 0.35]**

Solution:

a) $P(Yellow\ on\ the\ first\ draw) = \frac{Number\ of\ yellow\ balls}{Total\ number\ of\ balls} = \frac{2}{7}$

b) If the first ball is red, $P(Red\ on\ both\ draws) = \frac{2}{7} \times \frac{1}{6}$

If the first ball is blue, $P(Blue\ on\ both\ draws) = \frac{3}{7} \times \frac{2}{6}$

If the first ball is yellow, $P(Yellow\ on\ both\ draws) = \frac{2}{7} \times \frac{1}{6}$

$$\therefore P(Same\ colored\ ball\ twice\ with\ out\ rep.) = \left(\frac{2}{7} \times \frac{1}{6}\right) + \left(\frac{3}{7} \times \frac{2}{6}\right) + \left(\frac{2}{7} \times \frac{1}{6}\right)$$

c) Try yourself.

4. A box contains 20 bulbs, of which 5 are defective. If 3 of the bulbs are selected at random without replacement, what is the probability that all three bulbs are defective? [Ans: 0.0088]

5. A shelf contains 8 fiction books and 12 non-fiction books. You randomly pick 2 books, one after the other, with replacement. What is the probability of picking a fiction book first and a non-fiction book second? [Ans: 6/25]

Measures of Dispersion (Grouped Data)

1. A botanist is studying the lifespan (in weeks) of two species of plants in a controlled environment. The data collected on the lifespan of the two species is summarized as follows:

Lifespan (weeks)	Species A	Species B
0-5	3	2
5-10	8	7
10-15	12	15
15-20	10	6
20-25	5	8

a) Which species has a higher average lifespan?

b) Calculate standard deviation for both species and comment about which species exhibits greater variation in lifespan?

c) Based on the relative measure of dispersion, which species would you prefer to cultivate and why?

Solution: First, we construct the table with necessary calculations:

Lifespan	f_A	f_B	*Mid value* (X)	f_AX	f_BX	X^2	f_AX^2	f_BX^2
0-5	3	2	2.5	7.5	5	6.25	18.75	12.5
5-10	8	7	7.5	60	52.5	56.25	450	393.75
10-15	12	15	12.5	150	187.5	156.25	1875	2343.75
15-20	10	6	17.5	175	105	306.25	3062.5	1837.5
20-25	5	8	22.5	112.5	180	506.25	2531.25	4050
Total	38	38		505	530		7937.5	8637.5

Now,

a) Here,

Mean of A,

$$\bar{X}_A = \frac{\sum f_A X}{\sum f_A} = \frac{505}{38} = 13.29\ weeks$$

Mean of B,

$$\bar{X}_B = \frac{\sum f_B X}{\sum f_B} = \frac{530}{38} = 13.95\ weeks$$

Comment: Species B provides higher average lifespan.

b)

Variance of A,

$$S_A^2 = \frac{\sum f_A X^2 - \frac{(\sum f_A X)^2}{\sum f_A}}{\sum f_A - 1} = 33.14$$

$$\therefore SD\ of\ A = \sqrt{33.14} = 5.76\ weeks$$

Variance of B,

$$S_B^2 = \frac{\sum f_B X^2 - \frac{(\sum f_B X)^2}{\sum f_B}}{\sum f_B - 1} = 33.66$$

$$\therefore SD\ of\ B = \sqrt{33.66} = 5.80\ weeks$$

Comment:

c)

CV of A,

$$CV_A = \frac{SD_A}{\bar{X}_A} \times 100 = \frac{5.76}{13.29} \times 100 = 43.34\%$$

CV of B,

$$CV_B = \frac{SD_B}{\bar{X}_B} \times 100 = \frac{5.80}{13.95} \times 100 = 41.58\%$$

Comment: Since the average lifespan of species B is higher and it exhibits lower variability, species B would be the preferred choice for cultivation.

2. A network administrator is studying the downtime (in hours) of two different server models over a six-month period. The data collected on downtime is summarized as follows:

Downtime (hours)	Server A	Server B
0-10	3	4
10-20	6	8
20-30	12	10
30-40	9	7
40-50	5	6

a) Which server has a higher average downtime? [$Ans: \bar{X}_A = 27; \bar{X}_B = 134.118$]

b) Calculate the standard deviation for both servers and determine which server has more variation in downtime. [$Ans: S_A^2 = 134.118;\ S_B^2 = 161.008$]

c) Based on the relative measure of dispersion, which server model would you recommend and why? **[Try yourself]**

3. A company is analyzing the tenure (in years) of employees across two departments to assess retention rates. The data is summarized as follows:

Tenure (years)	Department X	Department Y
0-2	7	5
2-4	10	12
4-6	13	15
6-8	10	8
8-10	5	5

a) Which department has a higher average employee tenure?

b) Calculate the standard deviation for both departments to determine which has greater variation in tenure.

c) Based on the relative measure of dispersion, which department would you recommend for focusing on retention strategies and why?

Time Series Analysis

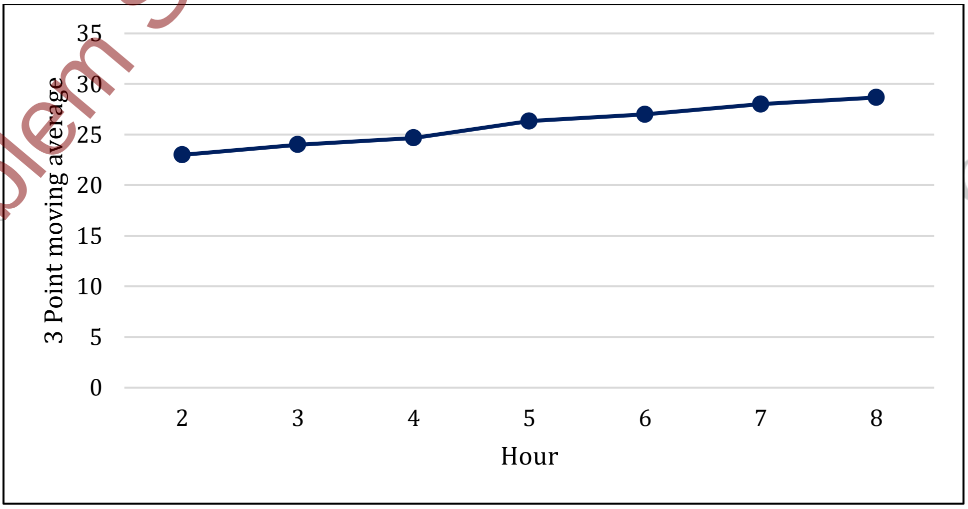
1. A sensor in a weather station record hourly temperature over a day. To smooth fluctuations and identify trend, a three-point moving average is applied. The station collected the following hourly temperature data:

Hour (t)	1	2	3	4	5	6	7	8	9
Temperature (Y_t)	22	24	23	25	26	28	27	29	30

a) Calculate trend using three-point moving average method.

Hour (t)	**Temperature (Y_t)**	**3-point average**
1	22	
2	24	$\frac{22+24+23}{3} = 23$
3	23	$\frac{24+23+25}{3} = 24$
4	25	$\frac{23+25+26}{3} = 24.67$
5	26	$\frac{25+26+28}{3} = 26.33$
6	28	$\frac{26+28+27}{3} = 27$
7	27	$\frac{28+27+29}{3} = 28$
8	29	$\frac{27+29+30}{3} = 28.67$
9	30	

b) Visualize the trend value using appropriate graphical method.



c) Suppose, the sensor experienced a sudden error, causing an outlier $Y_5 = 50$. How does this outlier affect three point moving average values for 4th, 5th, and 6th hour. Compare the results before and after error.

Solution: If Y_5 changes to 50, recalculate moving average for $t = 4,5, and\ 6$

At $t = 4, MA_4 = \frac{23+25+50}{3} = 32.67$

At $t = 5, MA_5 = \frac{25+50+28}{3} = 34.33$

At $t = 6, MA_6 = \frac{50+28+27}{3} = 35$

Comparison: The outlier significantly increases the moving average values for $t = 4,5\ and\ 6$, demonstrating moving averages are sensitive to large outliers.

2. An economist is studying the quarterly GDP growth rates of a country over the past three years. The data (in percentage) is as follows:

Quarter (t)	1	2	3	4	5	6	7	8	9	10	11	12
GDP (%)	3.2	2.9	3.5	2.8	3.1	2.7	3.3	2.6	3.4	2.5	3.0	3.4

a) Calculate the trend using four-quarter moving average method.

b) Fit linear trend model to the data using least squares estimation.

c) Using the linear trend model from (b), predict the GDP growth rate for the next two quarters.

3. An agricultural researcher is examining the annual yield (in metric tons) of a specific crop over the past 8 years to understand the trend. The yield data is as follows:

Year	2000	2001	2002	2003	2004	2005	2006	2007
Yield (Metric Tons)	50	52	54	56	56	60	62	64

a) Calculate the least squares estimate of the trend line.

b) Determine the trend value and visualize using appropriate graphical method.

c) Predict the crop yield for the years 2009, and 2010 using the trend line equation from (a).

Solution:

$Year$	$Yield\ (Y)$	$X = 2(t - \bar{t})$	X^2	XY
2000	50	-7	49	-350
2001	52	-5	25	-260
2002	54	-3	9	-162
2003	56	-1	1	-56
2004	56	1	1	56
2005	60	3	9	180
2006	62	5	25	310
2007	64	7	49	448
Total	**454**	**0**	**168**	**168**

a)

$$\hat{\beta} = \frac{\sum XY}{\sum X^2} = 0.9881$$

$$\hat{\alpha} = \bar{Y} = 56.75$$

$$\therefore The\ trend\ line, \hat{Y} = 56.75 + 0.9881X$$

- **When the number of years is odd:**

$$\boldsymbol{X = (t - \bar{t})}$$

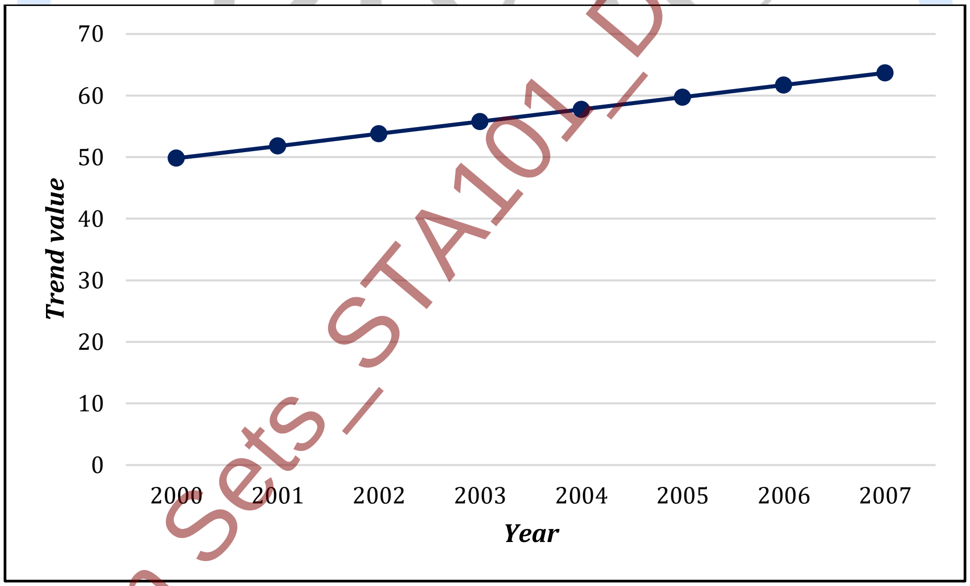
- **When the number of years is even:**

$$\boldsymbol{X = 2 \times (t - \bar{t})}$$

b)

Year	*Yield* (*Y*)	*X*	*Trend value*
2000	50	-7	49.8333
2001	52	-5	51.8095
2002	54	-3	53.7857
2003	56	-1	55.7619
2004	56	1	57.7381
2005	60	3	59.7143
2006	62	5	61.6905
2007	64	7	63.6667

Visualization:



c)

For 2009, X = 11, $\hat{Y} = 56.75 + (0.9881 \times 11) = 67.6191$

For 2010, X = 13, $\hat{Y} = 56.75 + (0.9881 \times 13) = 69.5953$

4. A researcher is analyzing the annual energy consumption (in million kWh) of a city over the past 7 years. The data collected is:

Year	2011	2012	2013	2014	2015	2016	2017
Energy Consumption	150	155	160	165	170	175	180

a) Find the linear equation that describes the trend in the annual energy consumption (in million kWh). [Ans: $\hat{Y} = 165 + 5X$]

b) Determine and visualize the trend value. **[Try yourself]**

5. A hospital is tracking the number of admissions (in hundreds) for a specific condition over the past 8 years. The data is as follows:

Year	2000	2001	2002	2003	2004	2005	2006	2007
Admissions (Hundreds)	20	22	24	26	28	30	32	34

a) Use least squares estimate method to discover and illustrate the trend line. [Ans: $\hat{Y} = 27 + 1X$]

b) Determine the trend value. **[Try yourself]**

Index number

1. An economist is analyzing the price trends of three commodities over two years: 2012 (base year) and 2016 (current year). The prices and quantities consumed of these commodities are as follows:

Commodity	2012		2016	
	Price	Quantity	Price	Quantity
A	20	50	25	60
B	15	40	18	45
C	10	30	12	35

a) Calculate the Laspeyres price index for current year. [Ans: 122.63]

b) Calculate the Paasche price index for current year. [Ans: 122.7]

c) Calculate the Fisher’s ideal price index for the current year and interpret the result. [Ans: 122.66]

2. From 2013 to 2019, this table tracks the prices and quantities of different items. It offers a snapshot of how markets change over time, reflecting shifts in demand and economic conditions.

	Year	2013	2014	2015	2016	2017	2018	2019
Item 1	**Price**	56	58	60	65	68	81	83
	Quantity	25	30	20	35	28	32	22
Item 2	**Price**	49	54	58	66	68	72	75
	Quantity	18	23	15	29	21	27	16
Item 3	**Price**	70	72	74	73	75	78	80
	Quantity	22	27	19	33	25	31	20

a) Use the three-year moving average method to discover and illustrate the trend in the price of "Item 1".

b) Fit linear trend model to the data (price of item 2) using least squares estimation.

c) Apply the best price index number method, and determine the index number for the year 2019 with the base year set as 2018.

Sampling techniques

1. A research team is conduction a study on the employment status of individuals in a large metropolitan city. The city is divided into 10 administrative zones, each with a mix of urban and suburban areas. The population of each zone varies, and the team wants to ensure their study results accurately represent the entire city. Due to budget constraints, they cannot survey the entire population and must use sampling techniques. They are interested studying the following groups:

i. Unemployed individuals aged 18-35.
ii. Part time workers in suburban areas.
iii. Full time workers in urban areas.

Each administrative zone has a registry of residents with their employment status, age, and address.

a) If the team wants to ensure equal representative from each administration zone, which sampling method should they use? Explain why.

Solution: Stratified Random Sampling. This method ensures that each administrative zone (stratum) is equally represented in the sample. Within each zone, individuals can be randomly selected, maintaining balance across all zones regardless of their population sizes.

b) To focus on specific groups (e.g., Unemployed individuals aged 18-35), what sampling technique should they use? Justify your choice.

Solution: Purposive Sampling. Purposive sampling (or judgmental sampling) allows the researchers to target specific groups directly based on the study's focus. Using the registry, they can identify and select only unemployed individuals aged 18–35, saving resources and time.

c) What sampling method would you recommend if the team has limited resources and wants to survey only a few zones but ensure diversity?

Solution: Cluster Sampling. Cluster sampling divides the city into groups (clusters) based on geographic location (zones). A few clusters (zones) are randomly selected, and all or a subset of individuals within those clusters are surveyed. This method is cost-effective while maintaining diversity.

Example: Select 3 zones randomly and survey all individuals or a random sample within those zones.

2. retail company wants to evaluate customer satisfaction with their service across 15 branches in a city. Each branch has a different number of customers visiting daily. The company is limited by budget and cannot survey all customers. They want to gather feedback effectively using a combination of probability and non-probability sampling techniques.

The customer groups of interest are:

i. Regular customers who shop at least once a week.
ii. First-time customers.
iii. Customers visiting during peak hours.

The company also wants to compare feedback from smaller and larger branches to understand how branch size affects satisfaction.

a) If the company wants to ensure that customers from all 15 branches are represented proportionally to their daily customer count, which sampling method should they use?

   Solution: Proportional Stratified Sampling. This method divides the population into strata (branches) and selects a sample proportional to each branch's size (customer count).

b) To collect feedback from first-time customers only, which sampling method would be most appropriate?

   Solution: Purposive Sampling. This non-probability method allows targeting specific groups based on their characteristics (e.g., first-time customers). The company could identify these customers through billing data or questionnaires at entry.

c) If the company randomly selects three branches and surveys all customers in those branches, which sampling method is being used?

   Solution: Cluster Sampling. In cluster sampling, the population is divided into clusters (branches), and a random selection of clusters is surveyed in full.

3. A university wants to conduct a survey to understand the study habits of undergraduate students across all departments. The university has 5,000 undergraduate students enrolled, and the administration decides to survey 200 students. To ensure fairness, they want each student to have an equal chance of being selected. How can the university implement simple random sampling to select 200 students?

Solution: The university can list all 5,000 undergraduate students and assign each a unique identification number. Using a random number generator or lottery method, 200 student IDs are selected without replacement. The selected students are then contacted for the survey.

4. As a researcher, you aim to explore the average usage of campus facilities among university students. Describe how you would employ cluster sampling techniques to achieve this objective. **[Try yourself]**

5. A factory wants to assess the quality of widgets produced on an assembly line. Each day, the line produces 5,000 widgets, and the quality assurance team needs to inspect a sample of 100 widgets. Rather than randomly selecting samples, they decide to use systematic sampling for simplicity and efficiency. How can the team implement systematic sampling to select the 100 widgets from the production line?

Solution: Assign a unique number (1 to 5,000) to each widget produced that day. Calculate the sampling interval k (e.g., sampling interval $k = 50$ for this case). Randomly select the starting point (r) between 1 and k. Select every kth widget starting from r. For example, if $k = 50$ and $r = 23$, the selected widgets would be 23, 73, 123, 173, and so on.