Ques 01). In a partially destroyed laboratory, the legible record of analysis of correlation of data, is as follows :

(i) 8x - 10y + 66 = 0

(ii) (ii) 4x - 18y -214 = 0.

What were (a) means of x and y, (b) the coefficient of correlation between x and y and (c) the standard deviation of y?

Soln:

Given the equations of the regression lines:

$$8x - 10y + 66 = 0$$

$$4x - 18y - 214 = 0$$

These can be interpreted as the lines of regression of y on x and x on y respectively.

1. **Rewriting the first equation:**

$$8x - 10y + 66 = 0$$

$$10y = 8x + 66$$

$$y = \frac{8}{10}x + \frac{66}{10}$$

$$y = 0.8x + 6.6$$

2. **Rewriting the second equation:**

$$4x - 18y - 214 = 0$$

$$18y = 4x - 214$$

$$y = \frac{4}{18}x - \frac{214}{18}$$

$$y = \frac{2}{9}x - \frac{107}{9}$$

Next, we determine the means of x and y by finding the point of intersection of these two lines.

We equate the two expressions for y:

$$0.8x + 6.6 = \frac{2}{9}x - \frac{107}{9}$$

Multiplying the entire equation by 90 to eliminate the fractions:

$$72x + 594 = 20x - 1070$$

Bringing like terms together:

$$52x = -1664$$

Thus:

$$x = \frac{-1664}{52} = -32$$

Substituting $x = -32$ into either equation to find y:

$$y = 0.8(-32) + 6.6 = -25.6 + 6.6 = -19$$

So, the means are $\bar{x} = -32$ and $\bar{y} = -19$.

Next, the coefficient of correlation r between x and y is calculated using the slopes of the regression lines:

$$r = \sqrt{b_{yx} \times b_{xy}}$$

From the equations:

$$b_{yx} = 0.8, \quad b_{xy} = \frac{2}{9}$$

Thus:

$$r = \sqrt{0.8 \times \frac{2}{9}} = \sqrt{\frac{16}{90}} = \sqrt{\frac{8}{45}} \approx 0.422$$

In conclusion:

- The means are $\bar{x} = -32$ and $\bar{y} = -19$.
- The coefficient of correlation r is approximately 0.422.

Ques 02).

A) A random sample of size 64 has been drawn from a population with standard deviation 20. The mean of the sample is 80.

(i) Calculate 95% confidence limits for the population means.

Soln:

(i) Calculation of 95% Confidence Limits for the Population Mean:

Given:

- Sample size, $n = 64$
- Population standard deviation, $\sigma = 20$
- Sample mean, $\bar{x} = 80$
- Confidence level = 95%

For a 95% confidence interval, the critical value from the standard normal distribution (Z) is $Z_{\alpha/2} = 1.96$.

The formula for the confidence interval for the population mean μ is:

$$\text{Confidence Interval} = \bar{x} \pm Z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

Substituting the values:

$$\text{Standard Error (SE)} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{64}} = \frac{20}{8} = 2.5$$

The confidence interval is:

$$80 \pm 1.96 \times 2.5$$

$$80 \pm 4.9$$

So, the 95% confidence interval for the population mean is:

$$(75.1, 84.9)$$

(ii) How does the width of the confidence interval change if the sample size is 256 instead?

Soln:

If the sample size increases to 256, let's recalculate the confidence interval.

Given:

- New sample size, $n = 256$

The standard error with the new sample size is:

$$\text{Standard Error (SE)} = \frac{\sigma}{\sqrt{n}} = \frac{20}{\sqrt{256}} = \frac{20}{16} = 1.25$$

The new confidence interval is:

$$80 \pm 1.96 \times 1.25$$

$$80 \pm 2.45$$

So, the 95% confidence interval with the larger sample size is:

$$(77.55, 82.45)$$

B) A population consists of the numbers 2,5,7,8 and 10. Write all possible simple random samples of size 3 (without replacement). Verify that the sample mean is an unbiased estimator of the population mean.

Soln:

The population consists of the elements: $2, 5, 7, 8, 10$.

To find all possible simple random samples of size 3, we calculate combinations of 5 items taken 3 at a time. The number of such combinations is given by:

$$\text{Number of samples} = \binom{5}{3} = \frac{5!}{3!(5-3)!} = \frac{5 \times 4 \times 3!}{3! \times 2!} = \frac{5 \times 4}{2 \times 1} = 10$$

So, there are 10 possible samples, which are:

1. $\{2, 5, 7\}$
2. $\{2, 5, 8\}$
3. $\{2, 5, 10\}$
4. $\{2, 7, 8\}$
5. $\{2, 7, 10\}$
6. $\{2, 8, 10\}$
7. $\{5, 7, 8\}$
8. $\{5, 7, 10\}$
9. $\{5, 8, 10\}$
10. $\{7, 8, 10\}$

Now, we calculate the mean for each of these samples:

1. $\{2,5,7\} \rightarrow \bar{x}_1 = \frac{2+5+7}{3} = \frac{14}{3} = 4.67$
2. $\{2,5,8\} \rightarrow \bar{x}_2 = \frac{2+5+8}{3} = \frac{15}{3} = 5.00$
3. $\{2,5,10\} \rightarrow \bar{x}_3 = \frac{2+5+10}{3} = \frac{17}{3} = 5.67$
4. $\{2,7,8\} \rightarrow \bar{x}_4 = \frac{2+7+8}{3} = \frac{17}{3} = 5.67$
5. $\{2,7,10\} \rightarrow \bar{x}_5 = \frac{2+7+10}{3} = \frac{19}{3} = 6.33$
6. $\{2,8,10\} \rightarrow \bar{x}_6 = \frac{2+8+10}{3} = \frac{20}{3} = 6.67$
7. $\{5,7,8\} \rightarrow \bar{x}_7 = \frac{5+7+8}{3} = \frac{20}{3} = 6.67$
8. $\{5,7,10\} \rightarrow \bar{x}_8 = \frac{5+7+10}{3} = \frac{22}{3} = 7.33$
9. $\{5,8,10\} \rightarrow \bar{x}_9 = \frac{5+8+10}{3} = \frac{23}{3} = 7.67$
10. $\{7,8,10\} \rightarrow \bar{x}_{10} = \frac{7+8+10}{3} = \frac{25}{3} = 8.33$

The population mean μ is calculated as:

$$\mu = \frac{2+5+7+8+10}{5} = \frac{32}{5} = 6.4$$

Let's calculate the average of all sample means:

$$\text{Average of sample means} = \frac{\bar{x}_1 + \bar{x}_2 + \bar{x}_3 + \cdots + \bar{x}_{10}}{10}$$

Substituting the values:

$$\text{Average of sample means} =$$

$$\frac{4.67 + 5.00 + 5.67 + 5.67 + 6.33 + 6.67 + 6.67 + 7.33 + 7.67 + 8.33}{10}$$

$$\text{Average of sample means} = \frac{63.01}{10} = 6.401$$

Ques 03). A computer chip manufacturer claims that at most 2% of the chips it produces are defective. To check the claim of the manufacturer, a researcher selects a sample of 250 of these chips. If there are eight defective chips among these 250, test the null hypothesis is that more than 2% of the chips are defective at 5% level of significance. Does this disprove the manufactrurer's claim?(Given that $Z_{0.05} = 1.645$)

Soln:

- **Null Hypothesis (H_0):** The proportion of defective chips is at most 2%.

$$H_0 : p \leq 0.02$$

- **Alternative Hypothesis (H_1):** The proportion of defective chips is greater than 2%.

$$H_1 : p > 0.02$$

The sample proportion $\hat{p}$ represents the observed proportion of defective chips in the sample. It is calculated as:

$$\hat{p} = \frac{\text{Number of defective chips}}{\text{Sample size}} = \frac{8}{250} = 0.032$$

The standard error of the proportion under the null hypothesis is given by:

$$SE = \sqrt{\frac{p_0(1 - p_0)}{n}}$$

Substituting the values $p_0 = 0.02$ (claimed proportion) and $n = 250$ (sample size):

$$SE = \sqrt{\frac{0.02 \times (1 - 0.02)}{250}} = \sqrt{\frac{0.02 \times 0.98}{250}} = \sqrt{\frac{0.0196}{250}} = \sqrt{7.84 \times 10^{-5}} \approx 0.00886$$

The test statistic Z is calculated to determine how far the sample proportion deviates from the hypothesized population proportion:

$$Z = \frac{\hat{p} - p_0}{SE} = \frac{0.032 - 0.02}{0.00886} \approx 1.354$$

Given the significance level $\alpha = 0.05$, the critical value for a one-tailed test is $Z_{0.05} = 1.645$.

- If $Z \geq 1.645$, we reject the null hypothesis H_0.
- If $Z < 1.645$, we fail to reject the null hypothesis H_0.

In this case, $Z = 1.354$ is **less** than 1.645.

Since the test statistic $Z = 1.354$ is less than the critical value 1.645, we **fail to reject the null hypothesis**. This means there is not enough evidence at the 5% significance level to conclude that more than 2% of the chips are defective.

Ques 04). A) A problem of statistics is given of three students A,B and C whose chances of solving it are 0.3,0.5 and 0.6 respectively. What is the probability that the problem will be solved?

Soln:

Three students A, B, and C attempt to solve a problem. The probabilities of solving the problem are given as:

- $P(A \text{ solves the problem}) = 0.3$
- $P(B \text{ solves the problem}) = 0.5$
- $P(C \text{ solves the problem}) = 0.6$

First, we need to calculate the probability that each student fails to solve the problem:

- Probability that A does **not** solve the problem: $P(A') = 1 - 0.3 = 0.7$
- Probability that B does **not** solve the problem: $P(B') = 1 - 0.5 = 0.5$
- Probability that C does **not** solve the problem: $P(C') = 1 - 0.6 = 0.4$

The probability that none of the students solves the problem is the product of these individual probabilities (assuming independence):

$$P(\text{None of them solves the problem}) =$$

$$P(A') \times P(B') \times P(C') = 0.7 \times 0.5 \times 0.4 = 0.14$$

The probability that at least one student solves the problem is the complement of the probability that none of them solves it:

$$P(\text{At least one solves the problem}) = 1 - P(\text{None solves the problem})$$

Substituting the value:

$$P(\text{At least one solves the problem}) = 1 - 0.14 = 0.86$$

The probability that the problem will be solved by at least one of the students is 0.86.

B) Suppose 2% of the items made in a factory are defective. Find the probability that there are:
(i) 3 defectives in a sample of 100

Soln:

Given:

- Probability of a defective item, $p = 0.02$
- Probability of a non-defective item, $q = 1 - p = 0.98$

Here:

- $n = 100$
- $k = 3$
- $p = 0.02$
- $q = 0.98$

Using the binomial formula:

$$P(X = 3) = \binom{100}{3} (0.02)^3 (0.98)^{97}$$

Calculate the binomial coefficient $\binom{100}{3}$:

$$\binom{100}{3} = \frac{100!}{3!(100-3)!} = \frac{100 \times 99 \times 98}{3 \times 2 \times 1} = 161700$$

Substitute the values:

$$P(X = 3) = 161700 \times (0.02)^3 \times (0.98)^{97}$$

Let's compute this:

$$P(X = 3) = 161700 \times (0.000008) \times 0.1315$$

$$P(X = 3) \approx 0.170$$

(ii) no defectives in a sample of 50

Soln:

Here:

- $n = 50$
- $k = 0$
- $p = 0.02$
- $q = 0.98$

Using the binomial formula:

$$P(X = 0) = \binom{50}{0}(0.02)^0(0.98)^{50}$$

Since $\binom{50}{0} = 1$ and $(0.02)^0 = 1$:

$$P(X = 0) = 1 \times 1 \times (0.98)^{50}$$

Calculate $(0.98)^{50}$:

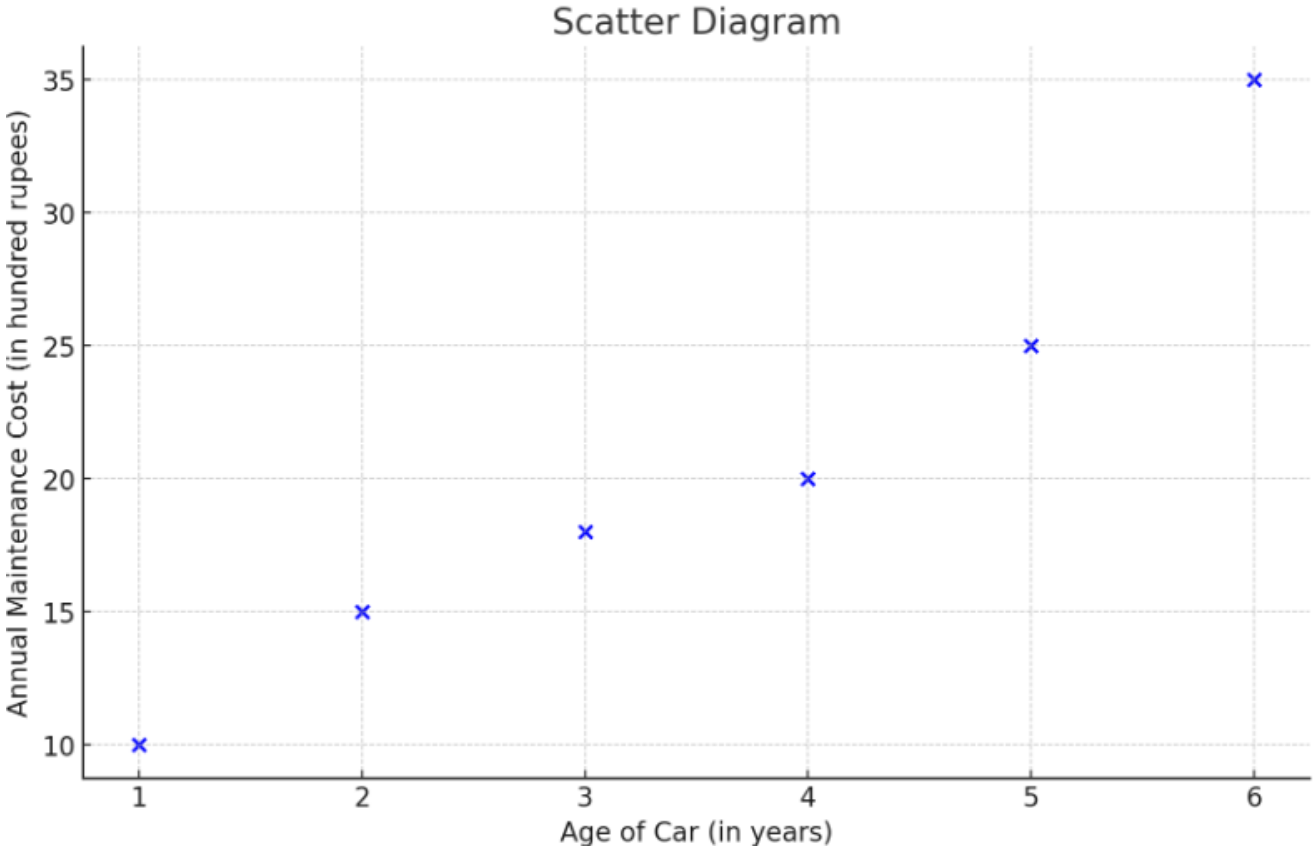
$$P(X = 0) \approx 0.364$$

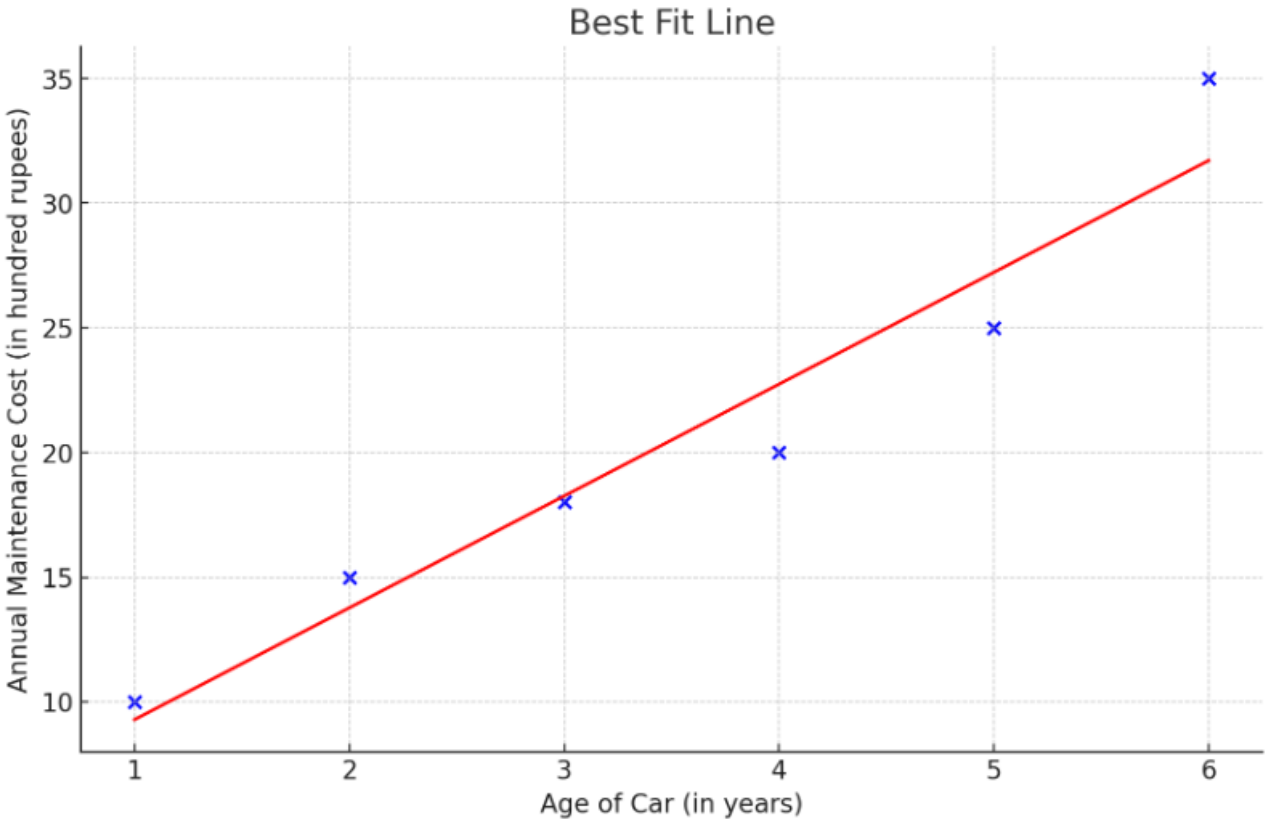
Ques 05). A Manager of a car company wants to estimate the relationship between age of cars and annual maintenance cost. The Following data from six cars of same model are obtained as:

Age of Car (in years)	Annual Maintenance Cost (In hundred rupees)
1	10
2	15
3	18
4	20
5	25
6	35

(a) construct a scatter diagram for the data given above

Soln:





(b) Fit a best linear regression line, by considering annual maintenance cost as the dependent variable and the age of the car as the independent variable.

Soln:

- The best-fit linear regression line is found using the method of least squares. The equation of the line is:

$$\text{Annual Maintenance Cost} = 4.49 \times (\text{Age of Car}) + 4.80$$

- This line has been plotted on the graph along with the scatter points. The line shows the trend that the annual maintenance cost increases as the car's age increases.

(c) use this regression line to predict the annual maintenance cost for the car of age 8 years.

Soln:

- Using the regression line, we predict the annual maintenance cost for a car that is 8 years old:

$$\text{Predicted Annual Maintenance Cost} \approx 40.69 \text{ hundred rupees} = 4069 \text{ rupees}$$

Ques 06). What do you understand by the term forecasting? With the help of the a suitable example discuss the relation between forecasting and future planning. Briefly discuss both forecasting model.

Soln : **Forecasting** is the process of making predictions about future events or trends based on historical data, current conditions, and analysis. It is widely used in various fields such as economics, business, finance, weather prediction, and more. Forecasting helps organizations and individuals to anticipate future needs, challenges, or opportunities, allowing them to make informed decisions and plan accordingly.

Example: Let's consider a retail business that experiences seasonal fluctuations in sales. By analyzing past sales data, the company can forecast future sales for the upcoming seasons. For instance, if historical data shows a consistent increase in sales during the holiday season, the company can forecast a similar trend for the coming year. This forecast allows the business to plan its inventory, marketing strategies, and staffing levels in advance, ensuring that it is well-prepared to meet customer demand.

Forecasting and future planning are closely intertwined. Forecasting provides the necessary insights and predictions about future events or trends, which serve as the foundation for effective future planning. In essence, forecasting informs the planning process by providing data-driven expectations, enabling organizations or individuals to create strategies that align with anticipated future scenarios.

For example, in the context of the retail business mentioned earlier, the forecasted sales data for the holiday season would be used to plan inventory levels, marketing campaigns, and staffing requirements. Without accurate forecasting, the company might either overstock or understock its products, leading to either excess inventory costs or missed sales opportunities. Thus, forecasting ensures that future planning is grounded in realistic expectations, enhancing the likelihood of achieving desired outcomes.

Types of Forecasting Models:

There are two primary types of forecasting models:

1. **Qualitative Forecasting Models:** These models rely on expert judgment, intuition, and subjective opinions rather than numerical data. They are often used when historical data is not available or when the situation involves new products, technologies, or markets.
Examples:
Delphi Method: This involves gathering insights from a panel of experts who provide their forecasts independently, and then revising their opinions based on the group's feedback until a consensus is reached.

Market Research: Surveys, focus groups, and interviews are used to gather qualitative data about consumer preferences and potential demand.

2. **Quantitative Forecasting Models:** These models use historical numerical data and mathematical techniques to predict future trends. They are most effective when there is a significant amount of reliable data available.
Examples:
Time Series Analysis: This involves analyzing historical data points collected over time to identify trends, seasonal patterns, and cyclical movements, and then projecting these patterns into the future. Methods like moving averages and exponential smoothing fall under this category.

Causal Models: These models assume that the variable to be forecasted has a cause-and-effect relationship with one or more other variables. For example, sales might be forecasted based on advertising spend using regression analysis.

Forecasting is a crucial tool that provides the foresight necessary for effective future planning. By choosing the appropriate forecasting model—whether qualitative or quantitative—organizations can better prepare for future challenges and opportunities, thereby enhancing their ability to achieve long-term goals.

Ques 07). Using the Regression line y=90+50x, fill up the values in the table below:

SAMPLE No. (i)	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	138	-	-	-	-
$\hat{e}_i$	0	-	-	-	-

After filling the table,compute the parameters of Goodness to fit i.e. R and R^2. Based on the result of R and R^2, interpret the correlation between variable x and y.

Soln:

1. Fill in the Missing Values:

SAMPLE No. (i)	x_i	y_i	$\hat{y}_i$	e_i
12	0.96	138	138.0	0.0
21	1.28	160	154.0	6.0
15	1.65	178	172.5	5.5
1	1.84	190	182.0	8.0
24	2.35	210	207.5	2.5

2. Parameters of Goodness-of-Fit:

- **Correlation Coefficient R:** $R = 0.994$
- **Coefficient of Determination R^2:** $R^2 = 0.988$

3. Interpretation:

The values $R = 0.994$ and $R^2 = 0.988$ indicate a very strong positive correlation between the variables x and y. The high R^2 value suggests that 98.8% of the variation in y is explained by the variation in x, meaning the regression model fits the data very well.

Ques 08). (i) Explain Linear and circular systematic sampling with example

Soln:

Systematic sampling is a statistical method where samples are selected at regular intervals from an ordered list. There are two main types: **Linear Systematic Sampling** and **Circular Systematic Sampling**.

1. Linear Systematic Sampling:

In linear systematic sampling, the sample is chosen by selecting a random starting point and then picking every k-th item from a list. The process stops when the end of the list is reached.

Example: Suppose you have a list of 50 employees and you need to select a sample of 5 employees.

- **Step 1:** Calculate the sampling interval k:

$$k = \frac{\text{Population Size}}{\text{Sample Size}} = \frac{50}{5} = 10$$

- **Step 2:** Choose a random starting point, say the 3rd employee.
- **Step 3:** Select every 10th employee after the 3rd one, so your sample will include the 3rd, 13th, 23rd, 33rd, and 43rd employees.

This method ensures the sample is spread out evenly across the entire list.

2. Circular Systematic Sampling:

Circular systematic sampling is similar to linear systematic sampling but with a key difference: when the end of the list is reached, the selection continues from the beginning, forming a loop or circle.

Example: Imagine you have a list of 12 products and want to sample 4 of them.

- **Step 1:** Calculate the sampling interval k:

$$k = \frac{\text{Population Size}}{\text{Sample Size}} = \frac{12}{4} = 3$$

- **Step 2:** Start at a random point, say the 2nd product.
- **Step 3:** Select every 3rd product: the 2nd, 5th, 8th, and 11th products. After the 12th product, the selection continues from the start, so you would include the 2nd product again if more samples were needed.

This method is particularly useful when the list is cyclical, and there is no logical beginning or end.

(ii) Explain Z-test and t-test with example.

Soln: **Z-Test**

A Z-test is a statistical test used to determine if there is a significant difference between sample and population means or between the means

of two samples. It is typically used when the sample size is large (n > 30) and the population variance is known or the sample size is large enough to estimate it accurately.

Formula: $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ where:

- $\bar{X}$ is the sample mean
- μ is the population mean
- σ is the population standard deviation
- n is the sample size

Example: Suppose you have a sample of 100 students' test scores from a school where the average test score for all students (population mean) is known to be 75 with a population standard deviation of 10. You want to know if the average score of your sample is significantly different from the population mean.

Let's say the sample mean score is 78.

$$Z = \frac{78 - 75}{10/\sqrt{100}} = \frac{3}{1} = 3.0$$

You would then compare this Z-value to a critical value from the Z-distribution table to determine significance (e.g., at a 5% significance level, the critical Z-value is approximately ±1.96).

T-Test

A t-test is used to determine if there is a significant difference between the means of two groups or between the sample mean and a known value when the sample size is small (n ≤ 30) and the population variance is unknown. It is used when the data follows a normal distribution and is appropriate for smaller sample sizes.

Formula: $t = \frac{X-\mu}{s/\sqrt{n}}$ where:

- $\bar{X}$ is the sample mean
- μ is the population mean
- s is the sample standard deviation
- n is the sample size

Example: Suppose you have a smaller sample of 20 students' test scores from the same school. The sample mean is 78, but you don't know the population standard deviation, only the sample standard deviation, which is 12.

$$t = \frac{78-75}{12/\sqrt{20}} = \frac{3}{2.68} \approx 1.12$$

You would then compare this t-value to a critical value from the t-distribution table based on the degrees of freedom (n-1 = 19) to determine significance.

(iii) Explain correlation and regression with example.

Soln: **Correlation**

Correlation measures the strength and direction of the linear relationship between two variables. It ranges from -1 to 1, where:

- **1** indicates a perfect positive linear relationship,
- **-1** indicates a perfect negative linear relationship,
- **0** indicates no linear relationship.

Common Measure: The Pearson correlation coefficient (\(r \)) is the most widely used measure.

Formula: $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}}$ where:

- x_i and y_i are the individual data points,
- $\bar{x}$ and $\bar{y}$ are the means of the x and y variables.

Example: Suppose you want to study the relationship between hours studied and exam scores. You collect data from 10 students and find that the correlation coefficient between hours studied and exam scores is r = 0.85. This indicates a strong positive linear relationship, meaning that as the number of hours studied increases, exam scores tend to increase as well.

Regression

Regression analysis estimates the relationship between a dependent variable and one or more independent variables. It helps in predicting the value of the dependent variable based on the values of the independent variables.

Simple Linear Regression: Involves one independent variable.

Formula: $Y = \beta_0 + \beta_1 X + \epsilon$ where:

- Y is the dependent variable,
- X is the independent variable,
- β_0 is the intercept,
- β_1 is the slope (coefficient) of the independent variable,
- ϵ is the error term.

Example: Continuing with the previous example, let's say you use simple linear regression to predict exam scores based on hours studied. Your regression equation might look like this:
Exam Score = 50 + 5 x (Hours Studied)

If a student studies for 4 hours, you can predict their exam score as:
Exam Score = 50 + 5 x 4 = 70

(iv) Explain Probability Distribution with example.

Soln: A probability distribution describes how the values of a random variable are distributed. It provides the probabilities of occurrence of different possible outcomes.

Types of Probability Distributions

1. **Discrete Probability Distribution**: Deals with discrete random variables, which can take on a finite or countably infinite number of values. Examples include the binomial distribution and the Poisson distribution.

2. **Continuous Probability Distribution**: Deals with continuous random variables, which can take on an infinite number of values within a range. Examples include the normal distribution and the exponential distribution.

Key Concepts

Probability Mass Function (PMF): For discrete distributions, the PMF provides the probability of each possible value.

Probability Density Function (PDF): For continuous distributions, the PDF describes the likelihood of the random variable taking on a particular value. The area under the PDF curve over an interval represents the probability of the random variable falling within that interval.

Cumulative Distribution Function (CDF): Describes the probability that the random variable will take a value less than or equal to a given point. It's applicable for both discrete and continuous distributions.

Examples

1. Discrete Probability Distribution:

Example: **Binomial Distribution**

Suppose you flip a fair coin 3 times. The random variable X represents the number of heads obtained.

The probability distribution of X (number of heads) can be calculated

using the binomial distribution formula:
$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$
where:
n is the number of trials (3 flips),
k is the number of successes (number of heads),
p is the probability of success (0.5 for a fair coin).

For example:
The probability of getting exactly 2 heads (*k* = *2*) is:
$P(X = 2) = \binom{3}{2} (0.5)^2 (0.5)^{3-2} = 3 \times 0.25 \times 0.5 = 0.375$

2. Continuous Probability Distribution:

Example: **Normal Distribution**

Suppose the heights of adult women in a certain country are normally distributed with a mean of 65 inches and a standard deviation of 3 inches. The random variable *X* represents the height.

The probability distribution of *X* can be described by the normal distribution with the parameters:

- Mean (μ) = 65 inches,
- Standard deviation (σ) = 3 inches.

The probability density function (PDF) of the normal distribution is:
$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

For example:
To find the probability that a randomly selected woman is between 62 and 68 inches tall, you would integrate the PDF from 62 to 68. This can be done using standard normal distribution tables or computational tools.