Statistic Assignment

2. I. Explain the different types of data (Qualitative and Quantitative) and provide examples of each. Discuss nominal, ordinal, interval, and ratio scales.

=> Qualitative data: (Categorical data)
Qualitative data refers to non-numerical information that describes characteristics or qualities. That means Qualitative data is a categorical data

→ Qualitative data is often categorical categorized into different groups or labels, which can be analyzed for patterns and insights.

Key feature of Qualitative data

1. Non-numerical nature: Qualitative data consists of descriptions or attributes, such as color, type or preference. For example, in a survey about favorite colors, response like "red", "blue", or "green" are qualitative

2. Categorical:- This data is usually grouped into categories (hence it's also known as "categorical data"). These categories can either be nominal (no inherent order, like gender or nationality) or

ordinal (where there's a natural order) like rating something as "poor", "good", or "excellent".

Example

① Nominal and ② ordinal

① Nominal

No order in categorical data naturally. Blood group, Gender, Color, etc.

② ordinal :-> order in data

Rank, good, better, best, grades, salary, high, medium, and low.

★ Statistical Analysis

- Frequency Counts
- Percentages
- Chi-Square Test

★ Visual Representation

- Bar charts
- Pie charts

2) Quantitative data (Numerical data):

Quantitative data is data that can be counted or measured and is represented as numbers.

→ data about numeric variables (e.g. how many, how much, or how often)

Quantitative = Quantity

allowing for various mathematical operations like (+, -, ×, ÷). Quantitative data provides concrete, measurable information that can be analyzed to identify trends, make predictions, or evaluate relationships between variables.

* Types of Quantitative data

Descrete data:

This type of consists of countable values, often integers. For example, the number of students in a classroom or the number of cars in parking lot.

→ Descrete data cannot be subdivided into finer increments within the range (e.g. you can't have half a student).

→ No of Children.

→ No of mobile Phones.

→ No of bank account.

→ total numbers of employees in team.

→ Bathroom, bedroom, Car number

* Continuous: Continuous data can take any value within a given range and can be subdivided.

→ real numbers.

Height, weight, speed, temperatures, car movie duration, length, Distance, salary, price, etc.

Example:- Discrete and Continuous.

Analysis of Quantitative data

1. Descriptive Statistics
   - measures of Central tendency
   - Mean
   - Median
   - Mode
2. Measure of dispersion
   - Range
   - Variance
   - Standard deviation
3. Inferential Statistics: Techniques used to make predictions or inferences about a population based on a sample data.
   - Hypothesis testing
   - Regression Analysis
   - Correlation

4. Interval data (measurement scale)
   Interval data is a type of quantitative data where the difference between values are meaningful, but there is no true zero point.
   → This type of data allows for the measurement of the difference between numbers, but ratio may not be meaningful.

Example:
- Temperature in celesius or fahrenheit.
- Dates on a calender.
- Time on a clock.

* ratio scale
The ratio scale is the highest level of measurement scale for quantitative data providing the most information and allowing for a wide range of mathematical operations. A ratio scale has all the properties of an interval scale but with an additional feature: a true zero point, meaning zero represents the complete absence of the quantity being measured.

1. True zero point: A ratio scale has an absolute zero which means that zero indicates a total absence of the quantity being measured. for example: In the context of weight, zero grams means no weight at all.

2. Equal intervals: Like an interval scale, the difference between values on a ratio scale is meaningful and consistent.
→ The diff between 10 and 20 kg is the same as between 20 and 30 kg.

3. Ratios are meaningful: there is a true zero point) you can meaningfully compare values in terms of ratios. for example: if one object weight 20 kg and another 10 weights 10 kg, you can say that the first object is twice as heavy as the second.

Example:
Height, weight, Time, Distance, income.

Q. 2 What are the measure of central tendency and when should you use each? Discuss the mean, median, and mode with examples and situations where each is appropriate.

⇒ Measure of central tendency are statistical tools used to identify the central value within a dataset.

→ These measures summarize a dataset with a single number that represents the center of the data distribution.

→ The three main measures of central tendency are the mean, median, and mode.

1. mean (Arithmetic Average)
The mean is the sum of all data values divided by the number of values in the dataset. it is the most common measures of central tendency.

$$\text{Mean} = \frac{\sum_{i=1}^{N} x_i}{N}$$

∴ When to use the mean:

1. Symmetrical data
2. Continuous and Discrete data
- Mathematical Analysis

Example:
If the test scores of five students are 70, 75, 80, 85 and 90 the mean is:

$$\text{Mean} = \frac{70+75+80+85+90}{5} = 80$$

2. Median (middle value)

The median is the middle value in a dataset when the values are arranged in ascending or descending order. If there is an even number of values, the median is the average of the two middle numbers and number is odd then most middle element is a median.

∴ When to use median

- Skewed data
- Ordinal data
- Non-normal Distributions

Example

4, 5, 2, 3, 2, 2

Sort 2, 2, 2, 3, 4, 5

→ Count the number of Element = 6

if Count is even

2 2 3 4 5

Median: average of two middle most element

$$= \frac{2+3}{2} = 2.5$$

4, 5, 2, 3, 2

Sort 2, 2, 3, 4, 5

Count = 5

↓ odd

median: the middle most element

= 3

* Mode: (most frequent value)

mode The mode is the value that appears most frequently in the dataset. A dataset may have no mode, one mode, or multiple modes (bimodal, multimodal)

→ When to use

- Categorical data
- Multimodal distributions
- non-numeric data

Example: 2, 2, 3, 3, 4, 4, 4, 4, 3, 4, 2

mode = 4

Q. 3. Explain the concept of dispersion. How do variance and standard deviation measure the spread of data?

⇒ The measure of dispersion help to interpret the variability of data. i.e. to know how much homogenous or heterogeneous the data is in simple terms, it shows how squeezed or scattered the variable is.

Common measures of dispersion

- Range
- interquartile Range (IQR)
- variance
- Standard perition.

* Variance

Variance measures the average squared deviations from the mean. it gives us an idea of how much the data points are spread out from the mean in a dataset. it quantifies the overall variability of the data

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Where,

- σ^2 is the variance
- x_i is each individual data point
- μ is the mean of the datasets
- N is the number of data points

Interpretation

- Large Variance: Indicates that the data points are widely spread out around the mean.
- Small Variance: suggests that the data points are closely clustered around the mean.

Standard deviation

The standard deviation is the square root of the variance, providing a measure of dispersion in the same units as the data.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

where:

- σ is the standard deviation.
- x_i is each data point.
- μ is the mean.
- N is the number of data points.

→ Both variance and standard deviation quantify how much individual data points deviate from the mean.

1. Variance: gives the average squared deviation from the mean which helps in understanding the overall dispersion but can be less intuitive due to its squared units.
2. Standard deviation directly shows the average distance of data points

from the Mean in the same units as the data making it more interpretable for practical applications.

Examples

Dataset 1: 5, 6, 7, 8, 9

Dataset 2: 1, 5, 7, 9, 13

Both dataset have the same mean 7.

- Variance of Dataset 1 is small because the values are close to mean.
- Variance of dataset 2 is larger because the values are more spread out from the mean.

→ The standard deviation would reflects these differences, with dataset 2 having a larger standard deviation than dataset 1.

Q. 4. What is a box plot and what can it tell you about the distribution of data?

⇒ A box plot (also called a box-and whisker plot) is a graphical representation of a dataset that visually display its distribution, spread and variability. It is particularly

useful for identifying key features of the data such as the median, quartiles, ranges, and potential outliers. Box plots provide a summary of the distribution without making assumptions about the underlying statistical distribution of the data.

Key Components of a Box Plot

1. Median (Q2): The line inside the box represents the median (the middle value of the dataset when ordered).

2. Box: The box represents the interquartile range (IQR), which is the range between the first quartile (Q1) and the third quartile (Q3). The edges of the box are drawn at Q1 (25th percentile) and Q3 (75th percentile).

3. Whiskers: The "whiskers" extend from the edges of the box to the smallest and largest data points within a certain range. Usually, they are drawn up to 1.5 times the IQR from the quartiles.

4. Outliers: These are individual data points that fall outside the whiskers

Breakdown of Components:

- Minimum: The smallest data point within 1.5 times the IQR of Q1.
- First Quartile (Q1): The 25 the Percentile, where 25% of the data lies below this value.
- Median (Q2): The 50th percentile, dividing the dataset in half.
- Third Quartile (Q3): The 75th percentiles where 75% of the data lies below this value.
- Maximum: The largest data point within 1.5 times the IQR of Q3
- Outliers: Data points that fall beyond the whiskers, indicating unusual or extreme values in the dataset.

Example of a box plot:
dataset of exam scores:
45, 50, 53, 55, 58, 60, 63, 65, 70, 75, 80, 85, 90, 95, 100

$$Q_1 = \frac{n+1}{4}\text{th} = \frac{15+1}{4}\text{th} = 4\text{th no.} = Q_1 = 55$$

$$Q_3 = \frac{3(n+1)}{4}\text{th} = \frac{3(15+1)}{4} = 12\text{th no.} = Q_3 = 85$$

$$Q_2 = \frac{(n+1)}{2}\text{th} = \frac{(15+1)}{2} = 8\text{th} = 65$$

- Whiskers: Extend from Q1 to 150 value.
- No obvious outliers in this case.

* What a box plot tells you about the data

1. Center (median): The line inside the box shows the median, which is a measure of central tendency. It gives an idea of where the middle of the data lies.

2. Spread (IQR): The width of the box represents the interquartile range (IQR), which indicates the spread of the middle 50% of the data.

3. Symmetry / Skewness.
- if the median is roughly centered inside the box, the data is symmetrically distributed.

- Right-skewed (positively skewed): the median is closer to Q1, and the right whisker is longer, indicating a longer tail on the right.
- Left-skewed (negatively skewed): The median is closer to Q3, and the left whisker is longer, indicating a longer tail on the left.

4. Range: The distance between the whiskers gives an idea of the

total range of the data. A large range indicates more variability in the dataset.

5. Outliers. Outliers are shown as individual points the whiskers.

Q. 5. Discuss the role of random sampling in making inferences about populations.

⇒ Random sampling is fundamental to making accurate inferences about populations in statistics.

1. Creating a Representative Sample
Random sampling ensures that every individual in a population has an equal chance of being selected. This results in a sample that accurately reflects the population's characteristics.

2. Facilitating Generalization
The primary goal of random sampling is to allow researchers to generalize their findings from the sample to the entire population.

3. Minimizing Bias
Random sampling reduces the risk of selection Bias which occurs when the sample does not accurately represent the population.

4. Supporting Statistical validity
Most statistical methods, such as confidence intervals and hypothesis testing, assume that data is collected through random sampling.

5- Error Estimation
With Random sampling, it's set possible to estimate sampling error- the difference between the sample statistic and the true population parameter.

6. Cost - Efficiency
Random sampling allows researchers to collect data from a subset of the population rather than the entire population, which can be expensive and time-consuming.

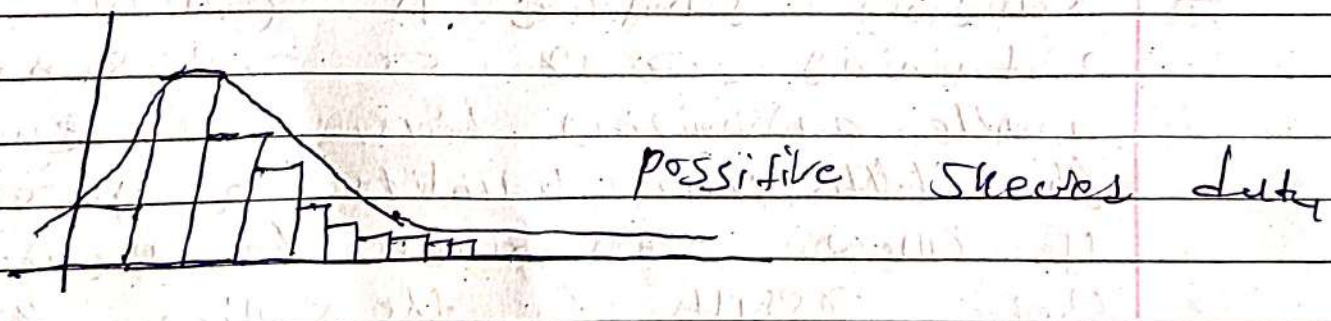
→ Random sampling is crucial for obtaining a representative sample, minimizing bias, and ensuring the statistical validity of inferences. It allows researchers to generalize their results to the entire population while accounting for errors and ensuring accuracy.

Q. 6: Explain the skewness and its types. How does skewness affect the interpretation of data?

Q) Skewness is a statistical measure that describes the asymmetry of a distribution of values around its mean. It indicates the direction and extent to which the values deviate from a normal distribution. Skewness can be positive, negative or zero, depending on the shape of the data distribution.

Types of Skewness

1. Positive skewness (Right-skewed)
   - In a positively skewed distribution, the tail on the right side (higher values) is longer or fatter than the left side.
   - The mean is usually greater than the median, and most of the data values cluster toward the lower end.



Possitive skewed data

2. Negative skewness (left-side skewed)
   - In a negatively skewed distribution, the tail on the left side (lower values) is longer or fatter than the right side.
   - The mean is usually less than the median, and most of the data values

cluster toward the higher end.

3. Zero skewness (Symmetric)
- In a symmetric distribution the data is evenly distributed around the mean with no skewness.
- the mean, median and mode are all equal.

Skewness interpretation and effects on data

1. mean, median, and mode
- in positively skewed data the mean is greater than the median, and the median is greater than the mode (Mean > Median > Mode).
- in negatively skewed data the mean is less than the median, and the median is less than the mode (Mean < Median < Mode).

2. outliers
- skewed distribution are often driven by outliers. A right skewed distribution may be influenced by unusually high values, while a left skewed distribution may contain unusually low values. These outliers can distort the interpretation of summary statistics, particularly the mean.

4. Data Transformation
When data is highly skewed transformations (like logarithmic or square root transformation) can reduce skewness and make the data more symmetrical.

5. What is the interquartile range (IQR), and how is it used to detect outliers?
The interquartile Range (IQR) is a measure of statistical dispersion that describes the spread of the middle 50% of a dataset. It is difference between the third quartile (Q3) and the first quartile (Q1), which are the 75th and 25th percentiles of the data, respectively.

$$IQR = Q3 - Q1$$

- Q1 (First quartile): The value below which 25% of the data fall.
- Q3 (Third quartile): The value below which 75% of the data fall.

Purpose of IQR
IQR is used to measure the spread of the middle portion of data, filtering out extreme values, which makes it less sensitive to outliers compared to other measures like range or standard deviation.

How IQR is used to Detect outliers

Outliers are data points that are significantly different from the rest of the data. The IQR is often used to identify potential outliers through the following

Calculate the IQR

- Subtract the first quartile from the third quartile to find the IQR.

2. A common rule of thumb is to consider any data points that fall below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$

- Lower fence : $Q_1 - 1.5 * IQR$
- Upper fence : $Q_3 + 1.5 \times IQR$
- Any data points below the lower fence or above the upper fence are considered potential outliers.

Example

10, 12, 13, 15, 16, 18, 19, 21, 24, 29, 45

$n = 11$

number is odd

then $Q_1 = \frac{n+1}{4} th = \frac{11+1}{4} th = \frac{12}{4} = 3rd\ no$

$Q_1 = 13$

$Q_3 = 3\frac{(n+1)}{4} th = 3\frac{(11+1)}{4} = \frac{3 \times 12^3}{4} = 9th$ no $= 24$

$IQR = Q_3 - Q_1 = 24 - 13 = 11$

- Lower fence: $Q_1 - 1.5 \times IQR = 13 - 1.5 \times 11$
  $= -3.5$
- Upper fence: $Q_3 + 1.5 \times IQR = 24 + 1.5 \times 11$
  $= 40.5$
- Any value below -3.5 or above 40.5 are considered outliers
- In this case, 45 is an outlier because it is greater than 40.5

Q. 8. Discuss the conditions under which the binomial distribution is used.

⇒ The binomial distribution is used to model the number of successes in a fixed number of independent trial of a binary (yes/no or success/failure) experiment. For the binomial distribution to apply, certain conditions must be met.

Conditions for using the binomial distribution

1. Fixed number of trials
- The experiment or process must be repeated a fixed of times, Each repetition is called a trial.
- For example, flipping a coin 10 times or rolling a die 15 times.

2. Two possible outcomes per trial
- Each trial must result in

exactly two possible outcomes, commonly referred to as "success" and "failure".

- Example: flipping a coin results in either heads or tails, or a medical test is either positive or negative.

3- Independence of trials.

- The outcome of each trial must be independent of the outcomes of other trials. This means that the result of one trial should not affect the result of another.

4. Constant probability of success (p)

- The probability of success, denoted as P, must remain the same for each trial. Likewise, the probability of failure, 1-P must also remain constant.

Binomial Distribution formula

The

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

- $P(X = k)$ is the probability of k successes.
- $\binom{n}{k}$ is the binomial coefficient $\frac{n!}{k!(n-k)!}$, representing the number of ways to choose k successes from

n trials.
- p^k represents the probability of k successes.
- $(1-p)^{n-k}$ represents the probability of n-k failures.

Examples of Binomial distribution in use

1- Coin tossing
- if a fair coin is tossed 5 times, the binomial distribution can be used to calculate the probability of getting exactly 3 heads out of 5 trials.

2. Quality control
- in manufacturing process, if the probability of a defective product is 0.02, and 100 products are tested the binomial distribution can be used to determine the probability of finding exactly 2 defective products.

3. medical testing
- if a diagnostic test has a 90% probability of accurately detecting a disease and 20 patients are tested, the binomial distribution can be used to find the probability of correctly diagnosing exactly 15 patients.

Q. 9. Explain the properties of the normal distribution and the empirical rule (68-95-99.7 rule).

⇒ The normal distribution, also known as the Gaussian distribution, is a continuous probability distribution characterized by a symmetrical bell-shaped curve. It is one of the most important distributions in statistics due to its wide applicability in real world scenarios, such as heights, test scores, measurements errors, etc.

Properties of the Normal distribution

1. Symmetry:
- The normal distribution is symmetric around its mean (μ). This means that the left and right sides of the curve are mirror images.
- The mean, median and mode of the distribution are equal and located at the center of the curve.

2. Bell-shaped Curve
- The distribution has a characteristic bell-shaped curve that peaks at the mean and tapers off symmetrically in both directions.
- Most of the data points are clustered around the mean, and the probability decreases as you move further from

from the mean.

3. Mean and Standard Deviation
- Mean (μ): Determines the location or the center of the distribution.
- Standard deviation (σ): Determines the spread or width of the distribution. A larger standard deviation means the data is more spread out, while a smaller one means the data is closely clustered around the mean.

4. Asymptotic:
- The tails of the normal distribution extend indefinitely approaching but never touching the horizontal axis.

5. Total area under the curve:
- The total area under the normal distribution curve equals 1, representing the entire probability space (or 100% of the probability).

6. Empirical Rule (68-95-99.7 Rule)
- The Empirical Rule, also known as the 68-95-99.7 rule, applies to normally distributed data.

1. The Empirical Rule (68 - 95 - 99.7 Rule)

→ The Empirical rule provides insight into how data is distributed around the mean in a normal distribution. It helps to quickly estimate probabilities and understand the spread of the data.

1. 68% of the data lies within one standard deviation ($\mu \pm 1\sigma$) of the mean:
   - Approximately 68% of the observations fall within one standard deviation (between $\mu - \sigma$ and $\mu + \sigma$).
   - This means that if the mean of dataset is 100, and the standard deviation is 20, then about 68% of the data will fall between 80 and 120.

2. 95% of the data lies within two standard deviations ($\mu \pm 2\sigma$) of the mean:
   - About 95% of the data falls within two standard deviations (between $\mu - 2\sigma$ and $\mu + 2\sigma$).
   - Using the same example, about 95% of the data will fall between 80 and 120.

3. 99.7% of the data within three standard deviations ($\mu \pm 3\sigma$) of the mean:
   - Nearly all the data (99.7%) falls within three standard deviations (between $\mu - 3\sigma$ and $\mu + 3\sigma$).
   - In this case, 99.7% of the data falls between 70 and 130.

Q. 10- Provide a real life example of a poisson process and calculate the probability for a specific event.

⇒ Customer arrivals at a bank

Suppose a bank receives, on average, 5 customer every 10 minutes. This scenario can be modeled as a Poisson process because:

1. Customer arrivals are independent of each other.
2. The rate of arrival (5 customer per 10 minutes) is constant over time.
3. Two customers cannot arrive at the exact same time.

Now lets calculate the probability that exactly 7 customers will arrive in the next 10 minutes.

Poisson distribution formula

The probability of observing exactly k events (here 7 customers) in a fixed period, given the average rate of 5 customer per 10 minutes is given by the Poisson probability formula.

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

- $P(X = k)$ is the probability of observing k events (7 customers).
- λ is the average rate of occurrence (5 customers in 10 minutes).

- k is the number of events (customers) for which we are calculating the probability (7 customers).
- e is Euler's number ($e \approx 2.718$).

Calculation

- $\lambda = 5$, $k = 7$

$$P(X = 7) = \frac{5^7 e^{-5}}{7!}$$

Now, $5^7 = 78125$,

$e^{-5} = 0.006737$, $7! = 5040$

$$P(X = 7) = \frac{78125 \times 0.006737}{5040} = \frac{525.125}{5040}$$

$$= 0.1044$$

The Probability that exactly 7 customers will arrives at the bank in the next 10 minutes is approximately 0.1044.

2. 11. Explain what a random variable is and differentiate the discrete and continuous random variables.

⇒ A random variable is a numerical value that represents the outcomes of a random phenomenon or experiment. It assigns a real number to each possible outcome in a sample space, allowing random events to be analyzed mathematical

tically. Random variables are essential in probability and statistics as they allow us to quantify uncertainty.

types of random variables

1. Discrete Random Variables
2. Continuous Random variables

1. ~~Discrete Random~~ variable

	Discrete Random	Continuous Random
Values	Countable, finite or infinite set of values	→ uncountable, infinite number of values in a range.
Example	Number of heads in coin tosses, dice rolls.	Heights, weights, time distance
Probability distribution	Probability mass function (PMF)	Probability density function (PDF).
Probability	non zero probabilities for specific values	Probability of any specific value is 0.
Calculation of Probability	Sum of probabilities for individual outcomes	integral of the probability density over a range.
	$P(X = x)$	$P(a \le y \le b) = \int_a^b f(y)\, dy$

Q. 12. Provide an example dataset, Calculate both covariance and correlation and interpret the results.

→ Suppose we have the following data for two variable, x (Number of study hours) and y (Exam Scores), for 5 students.

Student	1	2	3	4	5
Study hours (x)	2	3	5	7	9
Exam score (y)	50	60	80	85	95

- Mean of $x = \bar{x} = \frac{2+3+5+7+9}{5} = \frac{26}{5} = 5.2$
- mean of $y = \bar{y} = \frac{50+60+80+85+95}{5} = \frac{370}{5} = 74$

Calculate Covariance

$$Cov(x,y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Student	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	2	50	-3.2	-24	76.8
2	3	60	-2.2	-14	30.8
3	5	80	-0.2	6	-1.2
4	7	85	1.8	11	19.8
5	9	95	3.8	21	79.8

$$Cov(x,y) = \frac{206}{5} = 41.2$$

Calculate Correlation

$$\rho(x,y) = \frac{Cov(x,y)}{\sigma_x \sigma_y}$$

$$\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

Standard deviation of x

Student	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
1	-3.2	10.24
2	-2.2	4.84
3	-0.2	0.04
4	-1.8	3.24
5	3.8	14.44

$$\sigma_x = \sqrt{\frac{32.8}{5}} = \sqrt{6.56} \approx 2.56$$

Standard deviation of y

Student	$y_i - \bar{y}$	$(y_i - \bar{y})^2$
1	-24	576
2	-14	196
3	6	36
4	11	121
5	21	441

$$\sigma_y = \sqrt{\frac{1370}{5}} = \sqrt{274} \approx 16.55$$

Correlation Calculation.

$$P(x,y) = \frac{41.2}{2.56 \times 16.55} = \frac{41.2}{42.38} \approx 0.97$$

Interpretation of Results.

1. Covariance: The covariance of 41.2 indicates that there is a positive relationship between the number of study hours and test exam scores (y. Since the covariance is positive, we know that as Study

hours increase, exam scores tend to increase as well.

2. Correlation: The correlation coefficient of 0.972 suggests a strong positive linear relationship between the two variables. This means that more study hours are strongly associated with higher exam scores.