BIVARIATE STATISTICS

In univariate statistics, we analyze a single variable at a time. **Bivariate statistics** extends this analysis to explore the relationship between **two variables**. By examining pairs of data, we can investigate patterns, determine the nature and strength of their relationship, and use this relationship to make pictions. This chapter focuses on the relationship between two quantitative variables.

A BIVARIATE VARIABLES

Definition **Bivariate Data**

Bivariate data consists of pairs of values for two quantitative variables, recorded for each individual in a dataset. We typically denote these variables as (x, y), where:

- x is the **independent** (or explanatory) variable.
- y is the **dependent** (or response) variable.

Ex: A teacher records the hours each student studied (x) and their final exam score (y).

Hours Studied (x)	5	10	8	15
Exam Score (y)	50	85	75	95

Each pair of values, such as $(5, 50)$, is a single bivariate data point.

B SCATTER PLOTS

Definition **Scatter Plot**

A **scatter plot** is a graph that displays bivariate data as a collection of points in the Cartesian plane. The independent (explanatory) variable is plotted on the horizontal axis (x-axis), and the dependent (response) variable is plotted on the vertical axis (y-axis).

A **scatter plot** is the primary tool for visually identifying a potential relationship, or **correlation**, between two quantitative variables.

Method **Constructing a Scatter Plot**

1. **Identify Variables**: Determine which variable is independent (x) and which is dependent (y).
2. **Set Up the Axes**: Draw and label the horizontal axis for the x-variable and the vertical axis for the y-variable. Choose appropriate scales for both axes that cover the range of the data.
3. **Plot the Points**: For each pair of (x, y) values in your dataset, plot a single point on the graph at the corresponding coordinates.

Ex: A teacher recorded the number of hours students studied and their corresponding exam scores. The data is shown below:

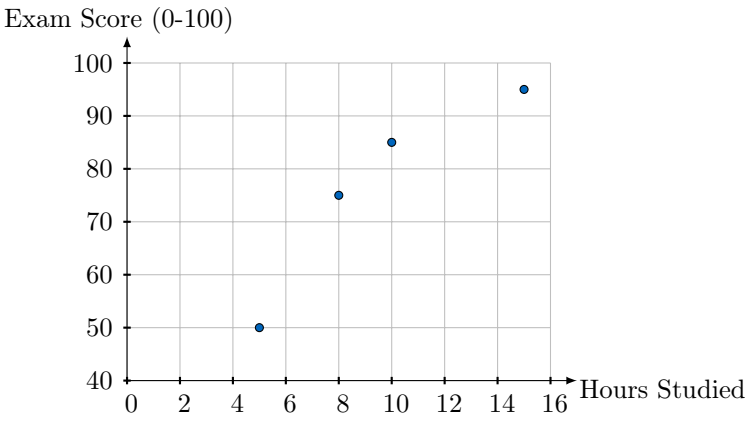
Hours Studied (x)	5	10	8	15
Exam Score (y)	50	85	75	95

Construct a scatter plot to visualize this data.

Answer:

1. **Variables:** "Hours Studied" is the independent variable (x) and "Exam Score" is the dependent variable (y).
2. **Axes:** The x-axis will be labeled "Hours Studied" and the y-axis will be "Exam Score". The scales must accommodate the data ranges.
3. **Plot Points:** We plot the four coordinate pairs: (5, 50), (10, 85), (8, 75), and (15, 95).

The resulting scatter plot is shown below:



C CORRELATION

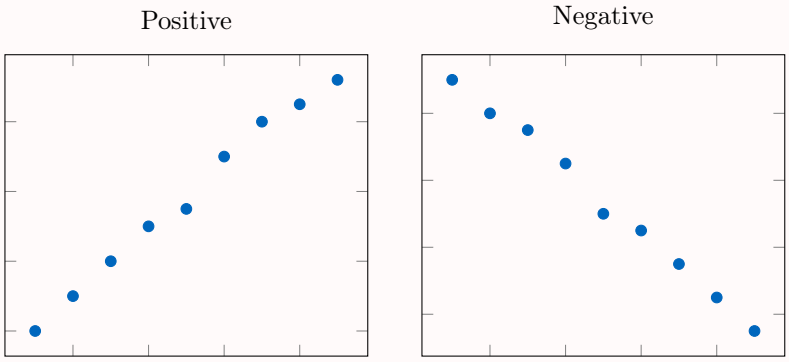
Definition Correlation

Correlation describes the nature of the relationship between two quantitative variables.

Definition Direction: Positive or Negative

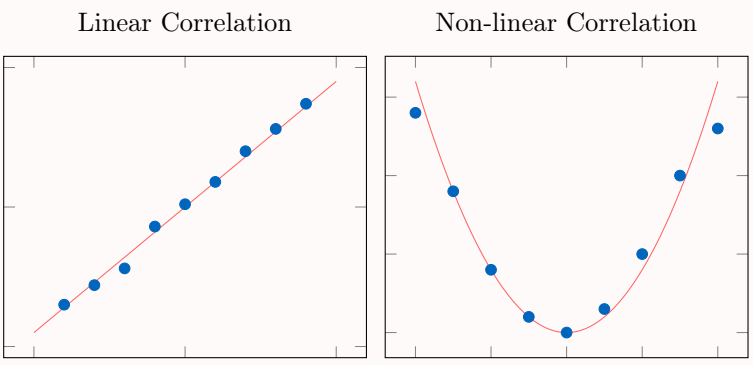
The **direction** describes the overall trend of the data.

- **Positive:** As the independent variable (x) increases, the dependent variable (y) tends to increase. The points trend upward.
- **Negative:** As the independent variable (x) increases, the dependent variable (y) tends to decrease. The points trend downward.



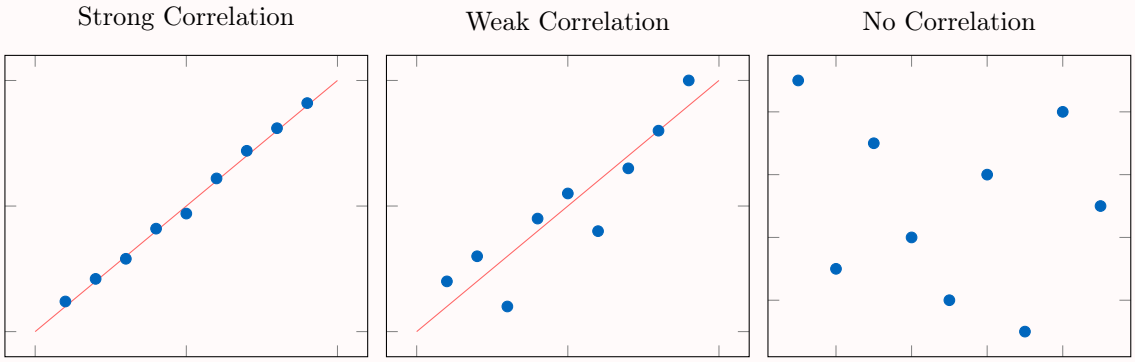
Definition Form: Linear or Non-linear

The **form** of the relationship is **linear** if the data points appear to follow a straight-line pattern. If they follow a curve other than a straight line, the form is **non-linear**.



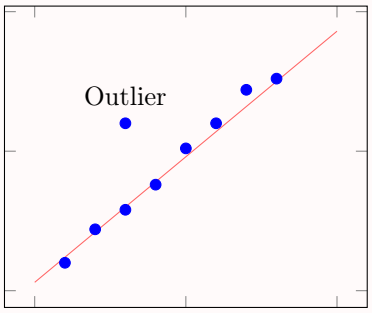
Definition Strength

The **strength** of a correlation describes how closely the data points adhere to the identified form.



Definition Outliers

An **outlier** is a data point that deviates significantly from the main pattern of the data.

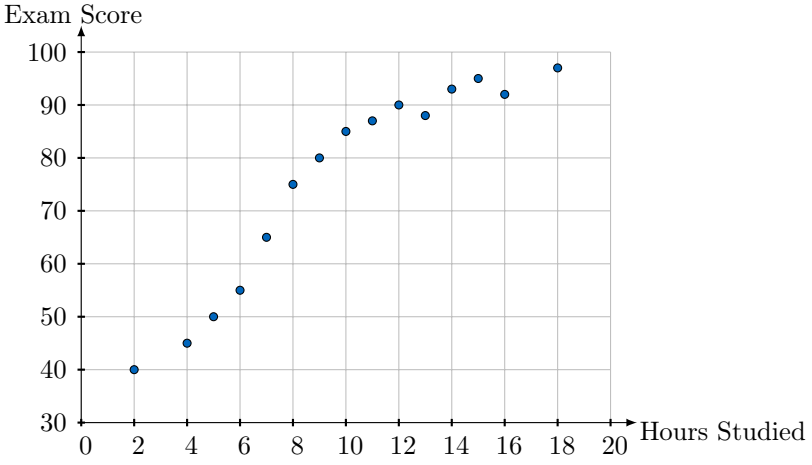


Method Describing a Correlation

When asked to describe the relationship shown in a scatter plot, you should always comment on all four features in a concise statement.

1. **Direction:** Is it positive or negative?
2. **Form:** Is it linear or non-linear?
3. **Strength:** Is it strong, moderate, or weak?
4. **Outliers:** Are there any notable outliers?

Ex: Describe the correlation between hours studied and exam scores shown in this scatter plot.



Answer: There appears to be a **strong, positive, linear correlation** between hours studied and exam scores. As the number of hours studied increases, the exam score tends to increase in a straight-line pattern. There are no obvious outliers.

D CORRELATION VS. CAUSATION

Correlation Does Not Imply Causation Observing a statistical relationship (correlation) between two variables, x and y, is not sufficient evidence to conclude that a change in x *causes* a change in y.

Definition Causation

Causation exists only if a change in the independent variable is shown to *directly cause* a change in the dependent variable. Proving causation requires a carefully designed controlled experiment, not just observational data.

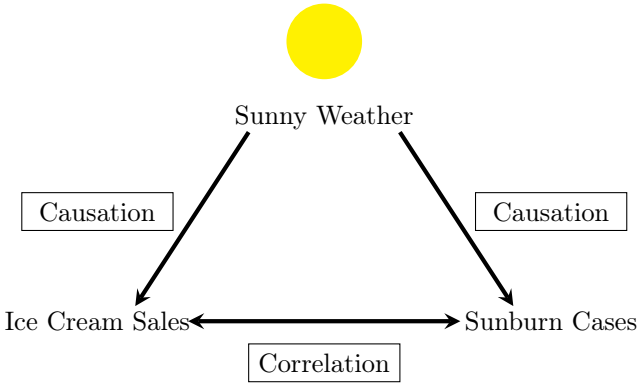
Definition Confounding Variable

Often, a correlation between two variables (x and y) is actually caused by a third, unobserved factor known as a **confounding variable** (z). This variable influences both x and y, creating an apparent but misleading relationship between them.

Ex: Data shows a strong positive correlation between ice cream sales and the number of people who get sunburned.
Does this mean eating ice cream causes sunburn? If not, identify the relationships and the likely confounding variable.

Answer: No, eating ice cream does not cause sunburn.

- The relationship between ice cream sales and sunburn cases is a **correlation**, not causation.
- The likely confounding variable is **Sunny Weather**. Hot and sunny days *cause* an increase in ice cream sales and also *cause* an increase in people getting sunburned.



E MEASURING LINEAR CORRELATION

While scatter plots allow us to visually describe a correlation, this assessment is subjective. To provide a precise and objective measure of the strength and direction of a **linear** relationship, we use numerical coefficients.

Definition Pearson's Correlation Coefficient (r)

The correlation coefficient is defined as the ratio of the covariance to the product of the standard deviations:

$$r = \frac{S_{xy}}{S_x S_y}$$

Where the components are calculated as follows:

- **Covariance (S_{xy}):**

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y})$$

- **Standard Deviation (S_x and S_y):**

$$S_x = \sqrt{\sum (x - \bar{x})^2} \quad \text{and} \quad S_y = \sqrt{\sum (y - \bar{y})^2}$$

where $\bar{x}$ and $\bar{y}$ are the means of the x and y data respectively, and $\sum$ means the sum over all the data values.

Ex: Calculate Pearson's Correlation Coefficient for the following data set :

x	0	1	2	3	4
y	0.8	3	4.8	7.1	9.2

Answer:

- **Step 1: Calculate the means**

$$\begin{aligned}\bar{x} &= \frac{\sum x}{n} \\ &= \frac{0+1+2+3+4}{5} \\ &= 2\end{aligned}$$

$$\begin{aligned}\bar{y} &= \frac{\sum y}{n} \\ &= \frac{0.8+3+4.8+7.1+9.2}{5} \\ &= 4.98\end{aligned}$$

- **Step 2: Calculate covariance (S_{xy})**

$$\begin{aligned}S_{xy} &= \sum (x-\bar{x})(y-\bar{y}) \\ &= (0-2)(0.8-4.98) + (1-2)(3-4.98) + \cdots + (4-2)(9.2-4.98) \\ &= 20.9\end{aligned}$$

- **Step 3: Calculate standard deviations (S_x, S_y)**

$$\begin{aligned}S_x &= \sqrt{\sum (x-\bar{x})^2} \\ &= \sqrt{(0-2)^2 + (1-2)^2 + \cdots + (4-2)^2} \\ &= \sqrt{10} \approx 3.162\end{aligned}$$

$$\begin{aligned}S_y &= \sqrt{\sum (y-\bar{y})^2} \\ &= \sqrt{(0.8-4.98)^2 + (3-4.98)^2 + \cdots + (9.2-4.98)^2} \\ &= \sqrt{43.728} \approx 6.613\end{aligned}$$

- **Step 4: Calculate r**

$$\begin{aligned}r &= \frac{S_{xy}}{S_x S_y} \\ &= \frac{20.9}{\sqrt{10} \times \sqrt{43.728}} \\ &\approx 0.999\end{aligned}$$

There is a very strong positive correlation.

In practice, for large datasets, Pearson's coefficient is calculated using digital tools (calculators or software). However, it is essential to understand how to calculate it manually when provided with **summary statistics**.

Proposition **Properties of Pearson's Correlation Coefficient (r)**

The Pearson's correlation coefficient (r) is a value in the range $[-1, 1]$ that quantifies the direction and strength of a **linear** relationship between two quantitative variables.

- The **sign** of r indicates the **direction** (positive or negative).
- The **magnitude** (absolute value) of r indicates the **strength**. An $|r|$ value close to 1 implies a strong linear correlation, while a value close to 0 implies a weak or no linear correlation.

Value of $\lvert r \rvert$	Strength of Correlation
$\lvert r \rvert = 1$	Perfect
$0.9 \leq \lvert r \rvert < 1$	Very Strong
$0.7 \leq \lvert r \rvert < 0.9$	Strong
$0.5 \leq \lvert r \rvert < 0.7$	Moderate
$0.3 \leq \lvert r \rvert < 0.5$	Weak
$0 \leq \lvert r \rvert < 0.3$	Very Weak or None

Definition **Coefficient of Determination (r^2)**

The **coefficient of determination (r^2)** is the square of the correlation coefficient. It is a value in the range $[0, 1]$ and is typically expressed as a percentage.
The value of r^2 represents the **proportion of the variance** in the dependent variable (y) that is predictable from the independent variable (x). In simple terms, it tells us how well the linear model fits the data.

Ex: A study of hours spent studying and exam scores finds a correlation coefficient of $r = 0.9$.
Interpret both r and r^2.

Answer:

- **Interpretation of r:** Since $r = 0.9$, there is a **very strong, positive, linear correlation** between the hours spent studying and the exam scores.

- **Interpretation of r^2:** We calculate $r^2 = (0.9)^2 = 0.81$. This means that **81% of the variation** in the exam scores can be explained by the linear relationship with the number of hours spent studying. The remaining 19% is due to other factors (e.g., natural ability, quality of sleep, etc.).

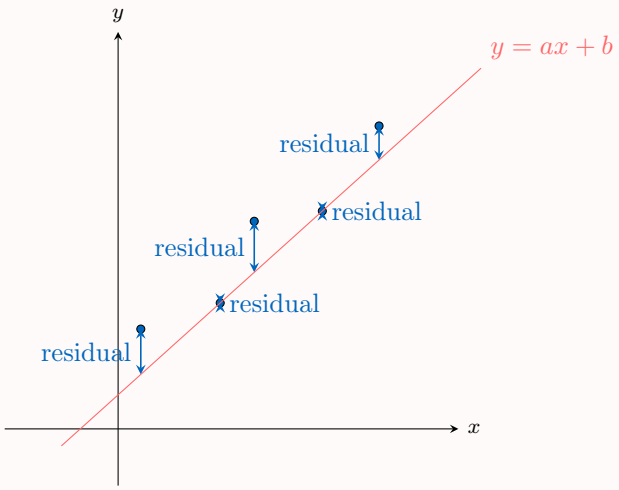
F LINEAR REGRESSION

When a scatter plot indicates a linear correlation between two variables, we can model this relationship using a straight line. This line, known as the **regression line**, can be used to make predictions. The reliability of this model is often assessed using the coefficient of determination (r^2). A high r^2 value indicates that a large proportion of the variance in the dependent variable is explained by the independent variable, suggesting the linear model is a good fit for the data.

Definition **Least Squares Regression Line**

The **least squares regression line**, written as $y = ax + b$, is the unique line of best fit that models the linear relationship between x and y. It is calculated by minimizing the sum of the squares of the **residuals**.
A **residual** is the vertical distance between an observed data point (x_i, y_i) and the predicted point on the regression line $(x_i, \hat{y}_i)$.

$$\text{Residual} = \text{observed } y - \text{predicted } y = y_i - \hat{y}_i$$



In practice, specifically for large datasets, the equation of the regression line is determined using the statistical functions of a calculator (GDC) or software. However, it is possible to calculate the coefficients manually using summary statistics.

A fundamental property of the least squares regression line is its relationship with the arithmetic means of the data sets. This property is frequently used to find missing variables without needing the full dataset.

Proposition **Mean Point**

The least squares regression line ($y = ax + b$) passes through the point defined by the mean of the x-values and the mean of the y-values: $(\bar{x}, \bar{y})$.

$$\bar{y} = a\bar{x} + b$$

Definition **Interpolation and Extrapolation**

The regression line can be used to make predictions:

- **Interpolation** is the process of predicting a y-value for an x-value that is **within** the range of the original data. If the correlation is strong, interpolation is generally considered reliable.
- **Extrapolation** is the process of predicting a y-value for an x-value that is **outside** the range of the original data. Extrapolation is generally considered unreliable, as we cannot assume the linear trend continues indefinitely.

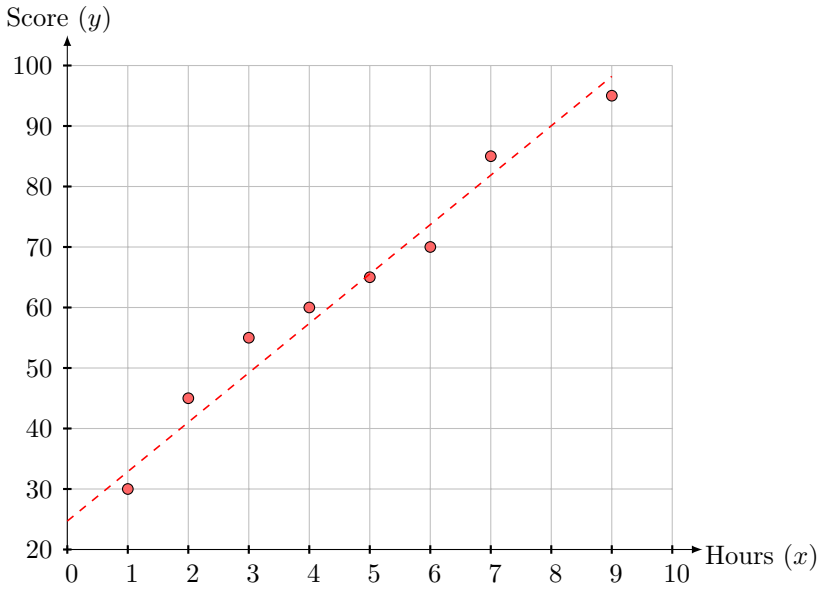
Ex: The average weekly study time and the final mathematics test score for a selection of students are shown below:

Study Time (x hours)	2	5	1	7	4	9	6	3
Test Score (y marks)	45	65	30	85	60	95	70	55

1. Construct a scatter diagram to illustrate the data.
2. Find the equation of the regression line y on x. State and interpret its gradient in the context of the problem.
3. Estimate the test score for a student who studies for 8 hours per week.
4. Estimate the weekly study time for a student who scores 40 marks.
5. Comment on the reliability of your estimates in 2 and 3.

Answer:

1. **Scatter Diagram:**



2. **Regression Line:**
Using a calculator, the equation is approximately:

$$y = 8.17x + 24.7$$

Interpretation:
The gradient is 8.17. This means that for every additional hour spent studying per week, the test score increases by approximately 8.17 marks on average.

3. **Estimation (Score for 8 hours):**
Substitute $x = 8$ into the equation:

$$y = 8.17(8) + 24.7 = 90.06$$

The estimated score is roughly **90 marks**.

4. **Estimation (Time for 40 marks):**
Substitute $y = 40$ into the equation:

$$40 = 8.17x + 24.7$$

$$15.3 = 8.17x$$

$$x \approx 1.87$$

The estimated study time is roughly **1.9 hours**.

5. **Reliability:**

   - The estimate in **3** ($x = 8$) is **reliable** because 8 hours is within the range of the given data ($1 \leq x \leq 9$). This is *interpolation*.
   - The estimate in **4** ($y = 40$) is also likely to be **reliable** because the corresponding x value (≈ 1.9) falls within the data range. However, caution should be used when predicting x from y using a y-on-x regression line mathematically.