Scatter Plots and Correlation

Week 3
Lecture 1
Spring 2025
Rida Maryam

Introduction

- Scatter plots and correlation are fundamental concepts in data analysis and statistics.
- They help us understand relationships between two numerical variables.

Why scatter plot is important ?

1. Identifying Trends & Relationships

- Helps determine if two variables have a **positive, negative, or no correlation** (e.g., height vs. weight).
- Example: In **healthcare**, a scatter plot can show if higher sugar intake leads to higher blood sugar levels.

2. Spotting Outliers & Anomalies

- Identifies unusual data points that may indicate errors or special cases.
- Example: In **fraud detection**, a scatter plot of transactions vs. time can highlight suspicious purchases.

3. Making Predictions

- Helps in forecasting trends based on past data.
- Example: In **finance**, a scatter plot of past stock prices vs. time can help predict future trends.

4. Comparing Variables

- Shows how two factors interact, helping in decision-making.
- Example: In **marketing**, a scatter plot can compare advertising budget vs. sales revenue to determine ROI.

Scatter Plot: Definition & Concept

- A **scatter plot** is a type of graph that shows the relationship between two variables.
- Each dot on the graph represents a data point, with one variable on the x-axis and the other on the y-axis.
- Scatter plots are a most effective tool for **geographic data** and **2D data** in general.

Example:
- Data on students' **study hours** and their **exam scores**. A scatter plot can show if students who study more tend to score higher.

How to Create a Scatter Plot

- X-axis: Independent variable (e.g., Study hours)
- Y-axis: Dependent variable (e.g., Exam score)
- Each point represents one observation.

Example Data:

Study Hours (X)	Exam Score (Y)
1	45
2	50
3	55
4	65
5	70

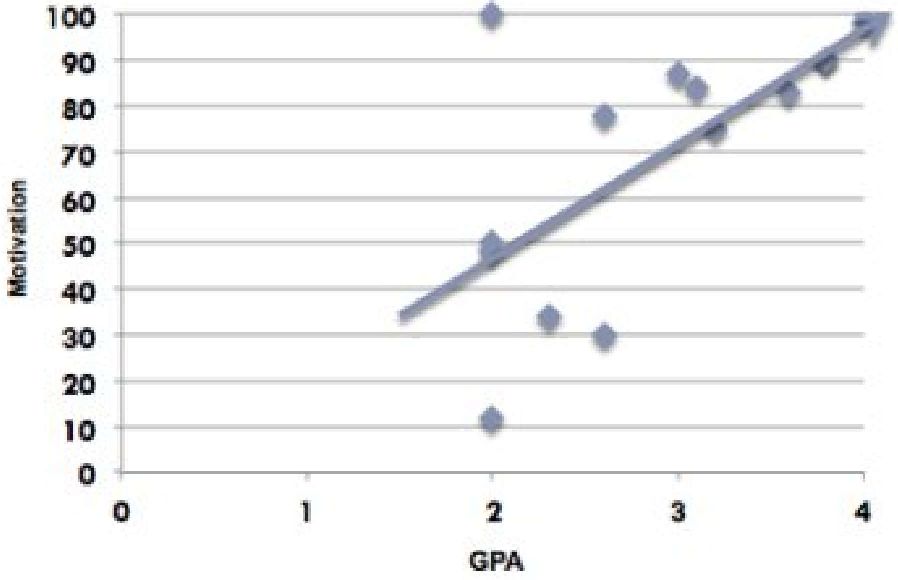
Correlation: Definition & Concept

- **Correlation** is a statistical measure that describes the **strength** , **form**(shape) and **direction** of a relationship between two variables.

Correlation: Direction

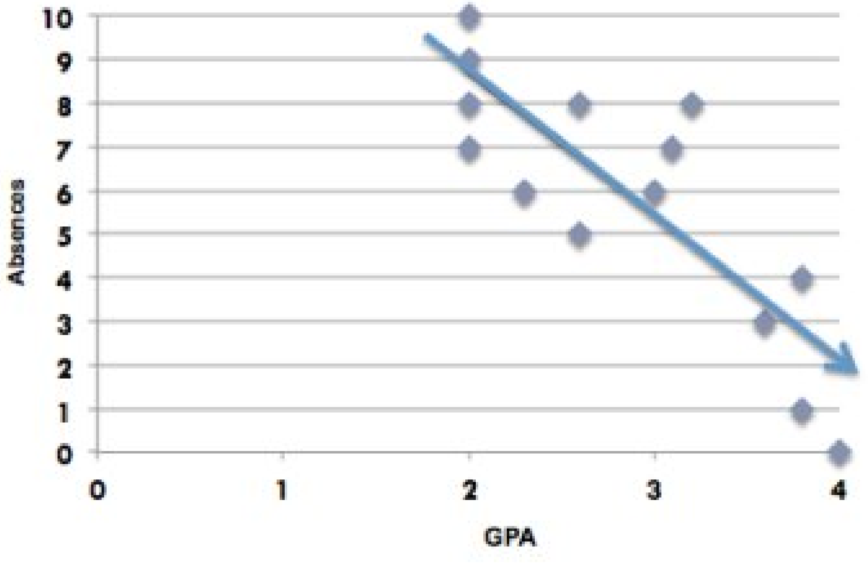
- **Positive Correlation:**
  - As X increases, Y increases (e.g., Study time vs. Test scores).
  - Points trend upwards.
- **Negative Correlation:**
  - As X increases, Y decreases (e.g., price of a product vs. demand).
  - Points trend downwards.
- **No Correlation:**
  - No clear pattern (e.g., Shoe size vs. Intelligence).
  - Randomly scattered points.

Positive Association



This example compares students' **achievement motivation** and their **GPA**. These two variables have a positive association because as GPA increases, so does motivation

Negative Association



This example compares students' **GPA** and their number of **absences**. These two variables have a negative association because, in general, as a student's number of absences decreases, their GPA increases.

Correlation: Strength(Positive)

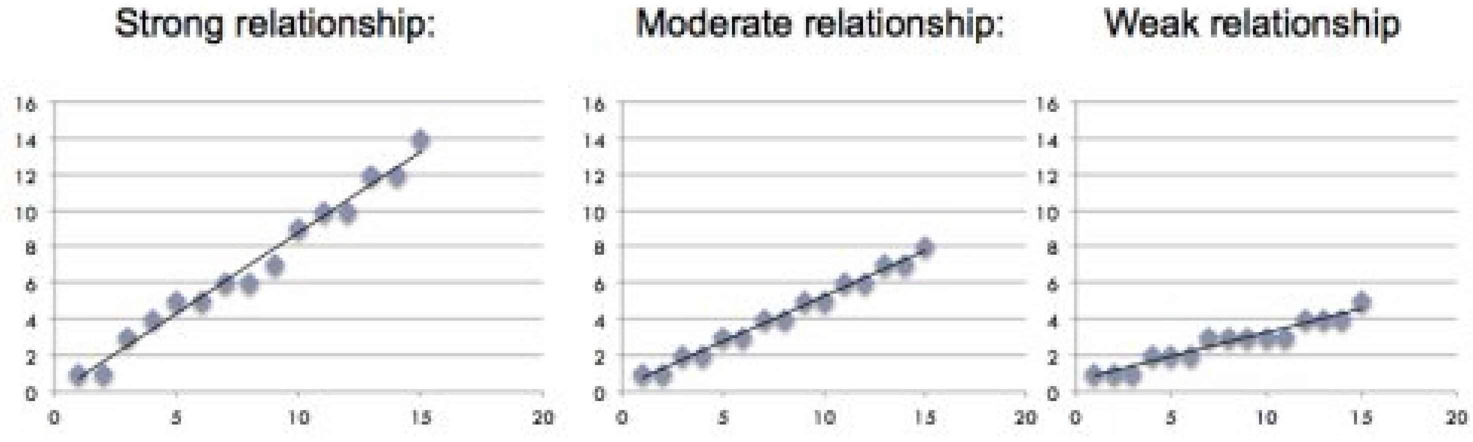
- **Positive Correlation:**
  - **Perfect Positive Correlation** (+1.0): All points lie on a straight upward line.
  - **Strong Positive Correlation** (Close to +1): Data points are close to an upward line.
  - **Weak Positive Correlation** (Near 0 but Positive): A slight upward trend.

Correlation: Strength(Negative)

➢ **Negative Correlation :**

- **Perfect Negative Correlation (-1.0):** All points lie on a straight downward line..
- **Strong Negative Correlation (Close to -1):** Data points are close to a downward line.
- **Weak Negative Correlation (Near 0 but Negative):** A slight downward trend.

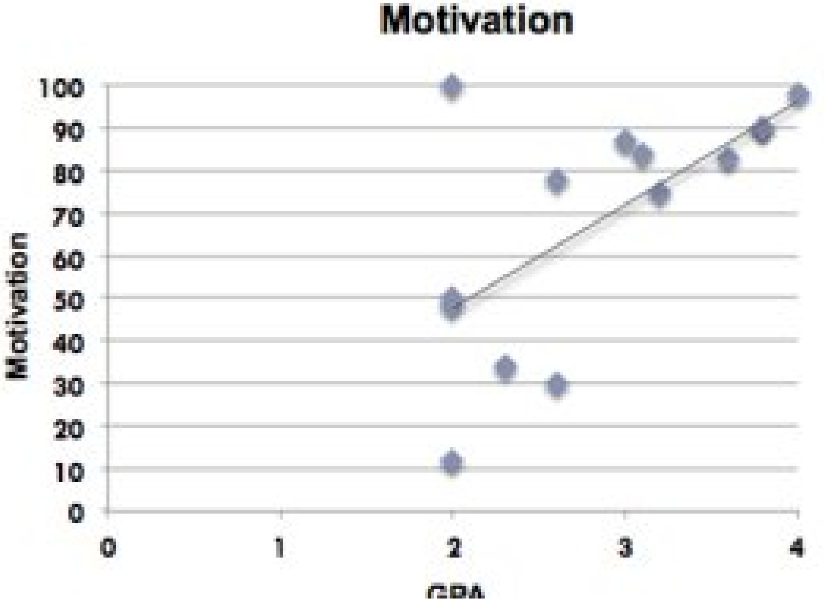
Strong, Moderate and Weak



Correlation: Form

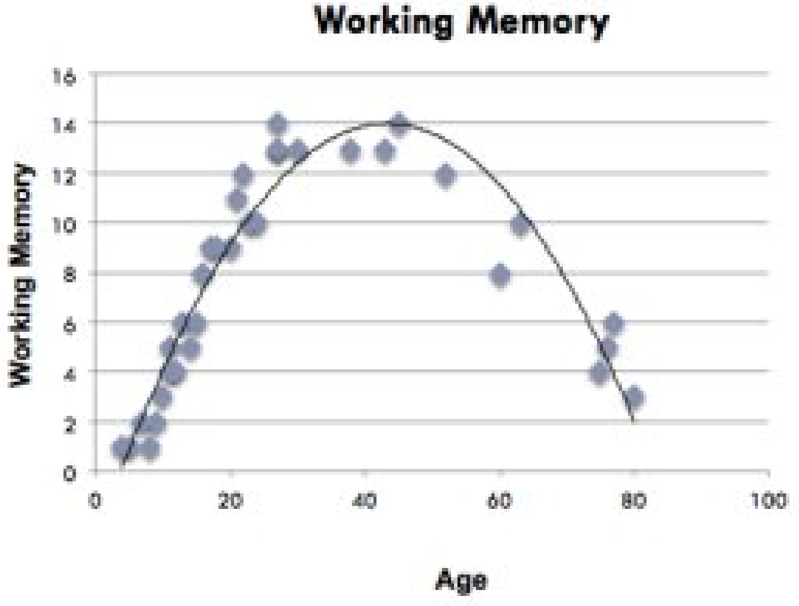
- **Linear Relationship:** One variable increases by approximately the same rate as the other variables changes by one unit.
- **Curvilinear relationship:** One variable does not increase at a constant rate and may even start decreasing after a certain point.

Linear relationship



This example illustrates a linear relationship. This means that the points on the scatterplot closely resemble a **straight line.**

Curvilinear relationship



This example describes a curvilinear relationship between
the variable “age” and the variable “working memory.” In this example, working memory increases throughout childhood,
remains steady in adulthood, and begins decreasing around age 50.

Real-Life Examples of Scatter Plots

- Study time vs. Test scores (Positive)
- Temperature vs. Ice cream sales (Positive)
- Exercise vs. Weight (Negative)
- Amount of Coffee Consumed vs. Sleep Duration (Negative)
- Hours of sleep vs. Productivity (Positive up to a point)
- Shoe size vs. Intelligence (No correlation)

Class Activity: Identify Correlation

1. Collect data from the class (e.g., hours studied vs. grades in score).
2. Plot a scatter diagram.
3. Identify if the correlation is positive, negative, or none.
4. Discuss possible reasons behind the trend.

The Correlation Coefficient and Its Interpretation

Week 3
Lecture 2
Spring 2025
Rida Maryam

Why it is required

- The strength of the relationship between two variables is a crucial piece of information.
- Relying on the interpretation of a scatterplot is too subjective.
- More precise evidence is needed, and this evidence is obtained by computing a coefficient that measures the strength of the relationship under investigation.

What is the Correlation Coefficient?

- The correlation coefficient (r) measures the strength, form(shape) and direction of a relationship between two variables.

Formula:

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

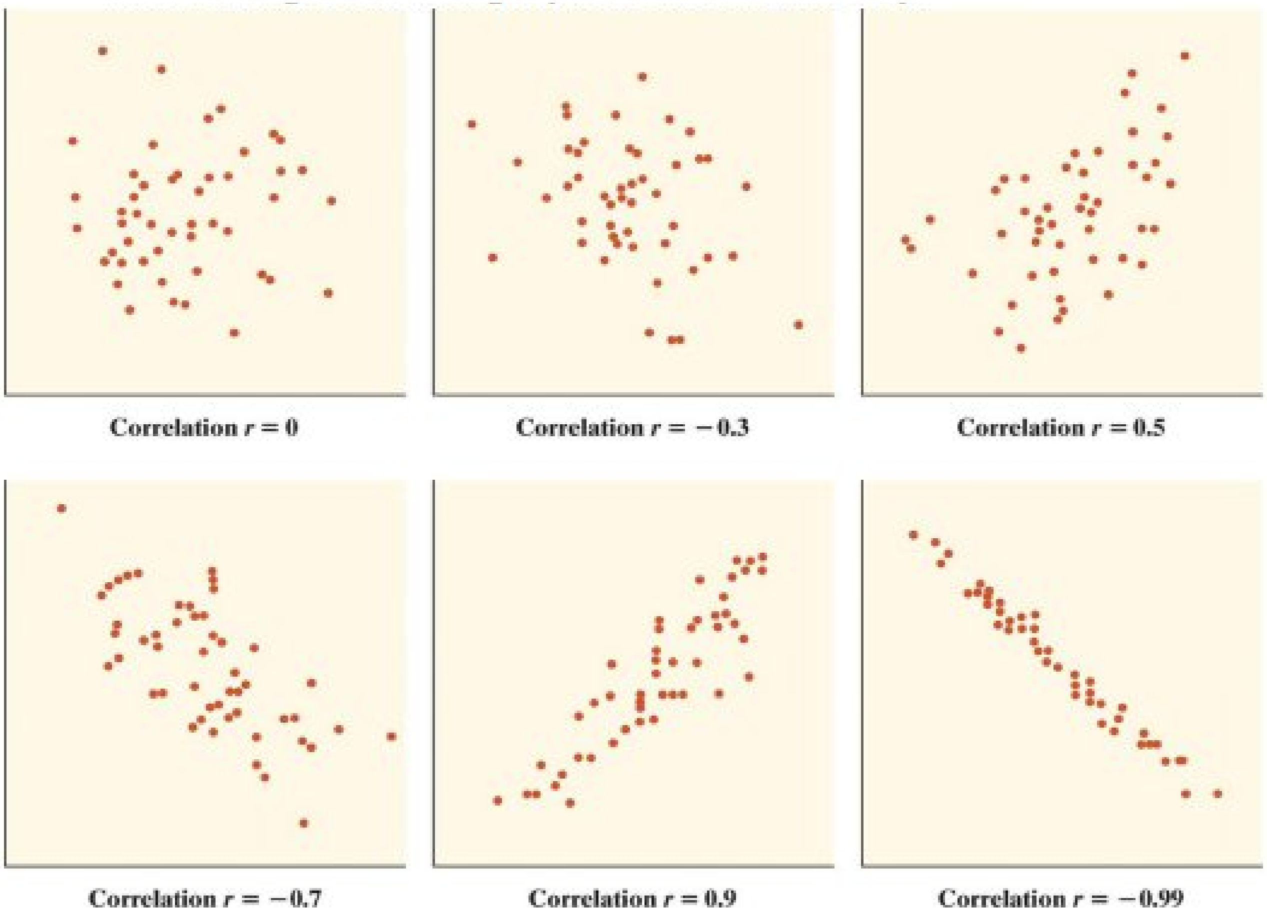
Values of r range between -1 and +1.

Interpreting the Correlation Coefficient

Value of r	Interpretation
$r = +1$	Perfect Positive Correlation
$r > 0.7$	Strong Positive Correlation
$0.3 < r < 0.7$	Moderate Positive Correlation
$0 < r < 0.3$	Weak Positive Correlation
$r = 0$	No Correlation
$-0.3 < r < 0$	Weak Negative Correlation
$-0.7 < r < -0.3$	Moderate Negative Correlation
$r < -0.7$	Strong Negative Correlation
$r = -1$	Perfect Negative Correlation

Correlations

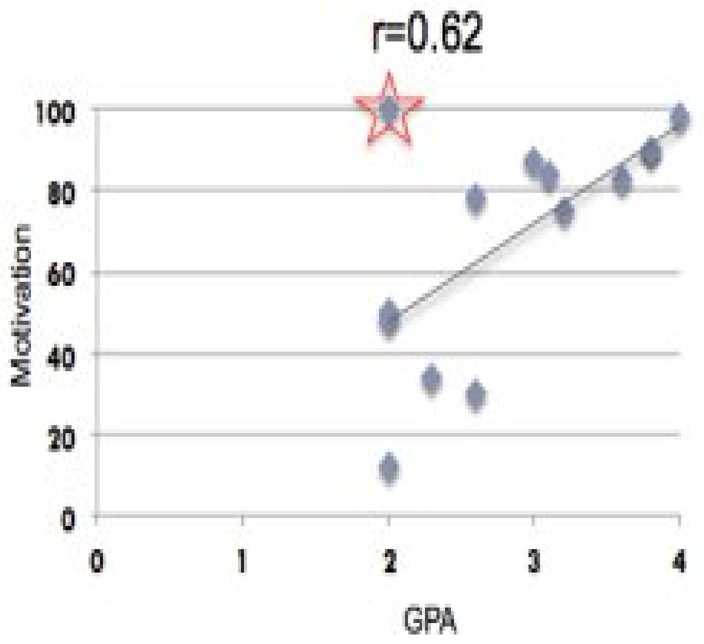
The images below illustrate what the relationships might look like at different degrees of strength (for different values of r)

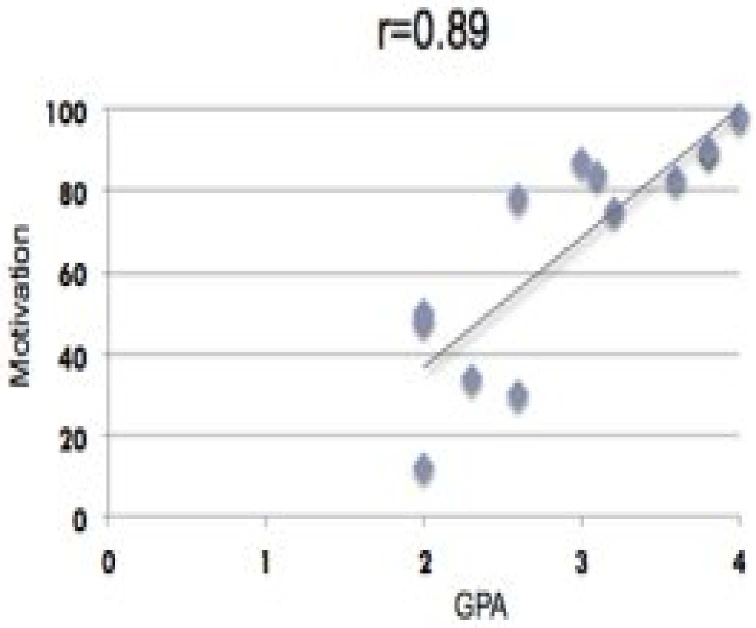


Pearson r: Assumptions

- Correlation requires that both variables be quantitative.
- Correlation describes linear relationships. Correlation does not describe curve relationships between variables, no matter how strong the relationship is. Cautions:
- Correlation is not resistant. r is strongly affected by outliers. Correlation is not a complete summary of two-variable data.

Example


r=0.62
Motivation
GPA
0
20
40
60
80
100
0
1
2
3
4


r=0.89
Motivation
GPA
0
20
40
60
80
100
0
1
2
3
4

Real-Life Examples of Correlation

- Study Time vs. Exam Scores ($r > 0.7$, Strong Positive)
- Temperature vs. Ice Cream Sales ($r > 0.7$, Strong Positive)
- Exercise vs. Weight ($r < -0.7$, Strong Negative)
- Coffee Consumption vs. Sleep Duration ($r < -0.7$, Strong Negative)
- Social Media Use vs. Study Performance (r between -0.3 and -0.7, Moderate Negative)

Class Activity: Calculate r

1. On the given dataset (e.g., hours studied vs. test scores), calculate the correlation coefficient.
2. Identify whether the correlation is positive, negative, or none.
3. Discuss findings with the class.

Study Hours (X)	Exam Score (Y)
1	45
2	50
3	55
4	65
5	70

Pitfalls in Correlation Analysis

Week 3
Lecture 3
Spring 2025
Rida Maryam

Objectives

- Correlation analysis has **pitfalls and limitations** that can lead to **misinterpretations** and **incorrect conclusions**.

Why is this important?

- Correlation does not imply causation.
- External factors can influence relationships.
- Misuse of correlation can lead to false conclusions.

Common Pitfalls in Correlation Analysis

1. Correlation Does Not Imply Causation
2. Confounding (Hidden) Variables
3. Sensitivity to Outliers
4. Non-Linear Relationships Can Be Misinterpreted
5. Small Sample Size Problems

1. Correlation Does Not Imply Causation

- - Just because two variables move together does not mean one causes the other.
- - Example: Ice cream sales and shark attacks are correlated, but hot weather is the real cause.

2. Confounding (Hidden) Variables

- - A third variable can explain the observed correlation.
- - Example: Bigger shoe sizes correlate with better reading skills, but age is the real factor.

3. Sensitivity to Outliers

- - Outliers Can Mislead Correlation
- Extreme data points can distort correlation values.
- - Example: If most students score 50-80, but one gets 0, it affects the study time vs. test score correlation.

4. Non-Linear Relationships

- - Correlation only measures linear relationships.
- - Example: Stress vs. performance follows a curve, not a straight line.

5. Small Sample Size Problems

- - Small datasets can give misleading correlation values.
- - Example: Studying the test scores of only 3 students might not reflect the true trend.