Chapter 2: Measures of Central Tendency and Dispersion

Introduction

After collecting and organizing statistical data, we need simple numerical measures to describe the data and understand their main characteristics.

A large data set can contain many observations, making it difficult to understand the data by looking at the observations individually. **Measures of central tendency (mesures de tendance centrale)** help us identify a central or typical value around which the observations are concentrated.

However, knowing only the central value is not enough. Two data sets may have the same mean but may be distributed very differently. **Measures of dispersion (mesures de dispersion)** allow us to measure how much the observations are spread around the central value.

For example, consider the following two data sets:

Data set A:

9, 10, 10, 10, 11

Data set B:

2, 6, 10, 14, 18

Both data sets have the same arithmetic mean:

Mean = 10

However, the observations in Data set A are close to 10, whereas the observations in Data set B are much more spread out.

Therefore, to describe a statistical series properly, we need both:

- **Measures of central tendency (mesures de tendance centrale)** to describe the central or typical value;
- **Measures of dispersion (mesures de dispersion)** to describe the spread or variability of the observations.

These measures are particularly important in **physics**, where repeated measurements may have different values because of experimental variability and measurement uncertainty.

This chapter introduces the statistical measures used to **summarize a data set with a few meaningful numbers**. After learning how to organize and represent data in Chapter 1, students now need to answer questions such as:

- What is the **typical value** of the data?
- Where is the data **centered**?
- How much do the observations **vary** around this center?
- How can we compare two data sets with the same average but different variability?

I. Measures of Central Tendency

When we have a set of observations, it is often useful to find **one value that represents the data set**.

For example, suppose the marks obtained by five students are:

8,10,12,14,16.

Instead of describing all five values, we may want to give a single value that represents the general level of the marks.

This is the purpose of **measures of central tendency (mesures de tendance centrale)**.

The three most important measures are:

1. **Arithmetic mean (moyenne arithmétique)**
2. **Median (médiane)**
3. **Mode (mode)**

These measures do not give exactly the same information. The appropriate measure depends on the nature and distribution of the data.

1.1. Arithmetic Mean

The **arithmetic mean (moyenne arithmétique)** is obtained by adding all the observations and dividing the result by the number of observations.

Let $x_{1,}, x_2, x_3, x_4, x_5 \ldots., x_n$ be n observations. The arithmetic mean is denoted by $\bar{x}$and is defined by

$\bar{x} = \frac{1}{N}\sum_{i=1}^{i=n} x_i$ with N the number of value

Example 1

Consider the following examination marks:

8,10,12,14,16.

Thus, $\bar{x} = \frac{1}{5}(8+10+12+14+16)=12$

Therefore, the arithmetic mean is

The value 12 gives a measure of the central level of the marks.

Arithmetic Mean for a Frequency Table

In statistics, observations are often presented together with their **frequencies (effectifs)**.

Suppose that a variable x takes the values

Let $x_{1,}, x_2, x_3, x_4, x_5 \ldots, x_k$ with corresponding frequencies $n_{1,}, n_2, n_3, n_4, n_5 \ldots, n_k$

The total number of observations is $N = \sum_{i=1}^{i=k} n_i$

The arithmetic mean is then $\bar{x} = \frac{1}{N}\sum_{i=1}^{i=n} x_i\, n_i$ with $N = n_{1,}+n_2 + n_3+n_4+n_5$ ....

Example 2

Consider the following marks:

Mark xix_i Frequency nin_i

8	2
10	3
12	4
14	1

The total number of students is

N=2+3+4+1=10.N=2+3+4+1=10.

We calculate

Mark x_i Frequency n_i_i $n_i x_i$

8	2	16
10	3	30
12	4	48

Mark x_i	**Frequency n_i_i**	$n_i x_i$
14	1	14
Total	**10**	**108**

Therefore,

$\bar{x}=108/10=10.8$

Important remark

The arithmetic mean takes **all observations into account**. Consequently, it can be strongly affected by an extremely large or extremely small value.

For example,

10,11,12,13,50

has a mean of

$\bar{x}=(10+11+12+13+50)/=19.2$

Although most observations are between 10 and 13, the value 50significantly increases the mean.

This observation will help us understand why the **median** can sometimes be more appropriate than the mean.

1.2. Median

Definition

The **median (médiane)** is the value that divides an ordered data set into two parts:

- approximately 50% of the observations are below the median;
- approximately 50% are above the median.

The observations must first be arranged in **ascending order**.

Case 1: Odd number of observations

If the number of observations nn is odd, the median is the observation in position

n+12.\boxed{\frac{n+1}{2}}.

Example

Consider

7,3,10,5,8.7,\quad 3,\quad 10,\quad 5,\quad 8.

First, arrange the values:

3,5,7,8,10.3,\quad 5,\quad 7,\quad 8,\quad 10.

There are n=5n=5 observations.

The median position is

5+12=3.\frac{5+1}{2}=3.

Therefore,

Median=7.\boxed{\text{Median}=7}.

Case 2: Even number of observations

If nn is even, there are two central observations. The median is the arithmetic mean of these two values.

For nn observations, the two central positions are

n2andn2+1.\frac{n}{2} \quad\text{and}\quad \frac{n}{2}+1.

Example

Consider

4,8,10,12,15,20.4,\quad 8,\quad 10,\quad 12,\quad 15,\quad 20.

There are n=6n=6 observations.

The two central values are in positions

62=3\frac{6}{2}=3

and

62+1=4.\frac{6}{2}+1=4.

Therefore, the two central values are 1010 and 1212.

Hence,

Median=10+122=11.\text{Median} = \frac{10+12}{2} = 11.

Thus,

Median=11.\boxed{\text{Median}=11}.

Important remark

Unlike the mean, the median is **less affected by extreme values**.

For example:

10,11,12,13,50. The mean is

19.2,

whereas the median is

12.12.

The median therefore gives a better indication of the central position of these observations when the data contain an extreme value.

2.1.4. Mode

Definition

The **mode (mode)** is the value that occurs **most frequently** in a data set.

A data set may have:

- one mode: **unimodal (unimodale)**;
- two modes: **bimodal (bimodale)**;
- more than two modes: **multimodal (multimodale)**;
- no mode if all values occur with the same frequency.

Example

Consider

8,10,12,10,14,10,12. The frequencies are:

Value Frequency

8 1

10 3

12 2

14 1

The value 10 occurs three times, which is more than any other value.

Therefore,

Mode=10

Important remark

The mode is particularly useful when we want to know the **most common value**.

Unlike the mean and median, the mode can also be used for **qualitative variables**, such as:

- eye colour;
- type of material;
- brand of a product;
- experimental category.

2.1.5. Comparison between Mean, Median and Mode

The three measures answer slightly different questions:

Measure	Main question
Mean	What is the average value?
Median	What is the central value after ordering the data?
Mode	What is the most frequently occurring value?

For a roughly symmetric distribution, the three measures may be close to one another.

For a distribution containing extreme values, the mean can be considerably different from the median.

Thus, it is important not to calculate a measure mechanically. We should also **interpret what the measure tells us about the data**.

Exercise 1

The following marks were obtained by 10 students:

8,12,10,15,12,9,12,14,10,8.

Calculate the arithmetic mean.

1. Arrange the data in ascending order.

2. Determine the median.
3. Determine the mode.
4. Compare the three measures.

Solution

The sum of the observations is

8+12+10+15+12+9+12+14+10+8=110.8+12+10+15+12+9+12+14+10+8=110.

Since n=10n=10,

$\bar{x}$=110/10=11.

Therefore,

$\bar{x}$=11.

The ordered data are

8,8,9,10,10,12,12,12,14,15.

Since n=10, the median is the mean of the 5th and 6th observations:

Median=(10+12)/2=11.

The value 12 occurs three times, more than any other value. Hence,

Mode=12

Therefore,

Mean=11, Median=11, Mode=12This example shows that the three measures do not necessarily have the same value.

Exercise 2

The following table gives the number of hours spent studying by a group of students before an examination.

Number of hours x_i	**Number of students n_i**
1	2
2	5
3	8

Number of hours x_i	**Number of students n_i**
4	4
5	1

1. Determine the total number of students.
2. Calculate the arithmetic mean.
3. Determine the median.
4. Determine the mode.
5. Interpret the results.

2. Measures of Dispersion

2.1. Introduction

In the previous section, we studied measures of central tendency, such as the mean, median, and mode. These measures describe the central or typical value of a statistical series.

However, two statistical series may have the same mean but very different distributions.

Group A	Group B
9	2
10	6
10	10
10	14
11	18
Mean = 10	Mean = 10

Both groups have the same mean, equal to 10. However, the marks in Group A are close to the mean, whereas the marks in Group B are more spread out.

Measures of dispersion are statistical measures used to describe how spread out the values of a statistical series are around their central value.

The main measures of dispersion are:

1. Range
2. Variance
3. Standard deviation
4. Coefficient of variation

2.2. Range (Étendue)

Definition

The range is the difference between the maximum value and the minimum value of a statistical series.

Formula

Range = Maximum value − Minimum value

Example 1

Consider the following examination marks obtained by five students:

8, 10, 12, 14, 16

Step 1: Identify the maximum and minimum values.

Maximum value = 16

Minimum value = 8

Step 2: Calculate the range.

Range = 16 − 8 = 8

Interpretation

The difference between the highest and lowest marks is 8 points. The range gives a simple indication of the spread of the data.

Example 2: Comparison of Two Groups

Consider the following two groups:

Group A	Group B
8	2
9	6
10	10
11	14
12	18

For Group A: Range = 12 − 8 = 4

For Group B: Range = 18 − 2 = 16

Interpretation: Group B has a larger range, meaning that its examination marks cover a wider interval.

Remark

The range is easy to calculate, but it depends only on the two extreme values. It does not take into account how the other observations are distributed.

Exercise 1

The following data represent the number of hours studied by six students before an examination:

3, 5, 4, 8, 6, 2

5. Identify the maximum value.
6. Identify the minimum value.
7. Calculate the range.
8. Interpret the result.

2.3. Variance

1. Introduction

The range depends only on the maximum and minimum values. Variance provides a measure that takes every observation into account by measuring the squared deviations from the mean.

2. Definition

The variance is the arithmetic mean of the squared differences between each observation and the arithmetic mean.

To calculate it: find the mean; subtract the mean from each observation; square each difference; then calculate the mean of the squared differences.

3. Formula

Variance = Sum of squared deviations from the mean / Number of observations

For n observations: $V(X) = \Sigma(x_i - \bar{x})^2 / N$

For values with frequencies: $V(X) = \Sigma\, n_i(x_i - \bar{x})^2 / N$

Here, x_i is an observation, $\bar{x}$ is the mean, n is the number of observations, n_i is the frequency of a value, and N is the total number of observations.

4. Worked Example

Consider the marks: 2, 4, 6, 8, 10.

Step 1: Calculate the mean.

Mean = (2 + 4 + 6 + 8 + 10) / 5 = 6

Step 2: Calculate the deviations and their squares.

Observation x	Deviation x − 6	Squared deviation $(x - 6)^2$	
2	-4	16	
4	-2	4	
6	0	0	
8	2	4	
10	4	16	
Total	0	40	

Step 3: Divide the sum of squared deviations by the number of observations.

Variance = 40 / 5 = 8

Interpretation: The variance is 8 squared mark-units. It summarizes the dispersion of all five marks around their mean, 6.

5. Important Remark

Variance is expressed in squared units. For example, if the data are examination marks, the variance is expressed in marks squared. The standard deviation, studied next, is expressed in the original unit.

Exercise 2

Consider the observations: 3, 5, 5, 7.

9. Calculate the arithmetic mean.
10. Complete a table containing x, x − mean, and (x − mean)².
11. Calculate the variance.
12. Explain what the variance describes.

The second formula of the variance is the computational formula, also called the shortcut formula.

Formula for n observations

$$V(X) = [\Sigma x_i^2 / n] - \bar{x}^2$$

Formula for a frequency distribution

$$V(X) = [\Sigma n_i x_i^2 / N] - \bar{x}^2$$

Where:

- x_i: each observed value.
- n_i: frequency of the value x_i.
- N: total number of observations.
- $\bar{x}$: arithmetic mean.

Worked Example

Consider the marks: 2, 4, 6, 8, 10.

Step 1: Calculate the mean.

$$\bar{x} = (2 + 4 + 6 + 8 + 10) / 5 = 6$$

Step 2: Calculate the mean of the squared values.

$$(2^2 + 4^2 + 6^2 + 8^2 + 10^2) / 5 = 220 / 5 = 44$$

Step 3: Apply the computational formula.

$$V(X) = 44 - 6^2 = 44 - 36 = 8$$

Conclusion: The variance is 8. The computational formula gives the same result as the formula based on squared deviations from the mean.

Thank you. Here is the complete material, ready to copy into Word.

Chapter 2: Measures of Central Tendency and Dispersion

3. Standard Deviation

3.1. Introduction

In the previous section, we studied variance as a measure of dispersion. Variance indicates how far the observations are spread out around their arithmetic mean.

However, variance is expressed in squared units. For example, if the observations represent examination marks, the variance is expressed in squared marks.

This makes its interpretation less direct.

To overcome this difficulty, we introduce another measure of dispersion called the standard deviation.

The standard deviation is widely used in statistics because it measures the dispersion of observations around their arithmetic mean and is expressed in the same unit as the original data.

3.2. Definition

The standard deviation is the positive square root of the variance.

It measures the degree of dispersion of observations around their arithmetic mean.

- A small standard deviation indicates that the observations are generally close to the mean.
- A large standard deviation indicates that the observations are more widely spread out around the mean.
- A standard deviation equal to zero means that all observations are identical.

3.3. Formula

The standard deviation is denoted by the Greek letter σ (sigma).

Standard Deviation = √Variance

For a statistical series containing n observations:

$\sigma = \sqrt{\Sigma(xi - x^{-})2/n\Sigma(x_i - \bar{x})^2 / n}$

Where:

- x_i: an observation.
- $\bar{x}$: the arithmetic mean.
- n: the number of observations.
- Σ: the sum of all terms.

When the statistical series is given with frequencies:

$\sigma = \sqrt{\Sigma ni(xi - x^{-})2/N\Sigma n_i(x_i - \bar{x})^2 / N}$

Where:

- n_i: frequency of the value x_i.
- N: total number of observations.

In this chapter, we divide by n because we are describing the complete statistical series given in the exercise.

3.4. Method of Calculation

To calculate the standard deviation, follow these steps:

1. Calculate the arithmetic mean.
2. Calculate the deviation of each observation from the mean.
3. Square each deviation.
4. Calculate the sum of the squared deviations.
5. Divide this sum by the number of observations to obtain the variance.
6. Calculate the positive square root of the variance.

3.5. Detailed Example 1

Consider the following examination marks obtained by five students:

2, 4, 6, 8, 10

Step 1: Calculate the arithmetic mean

Mean = (2 + 4 + 6 + 8 + 10) / 5

Mean = 30 / 5

Mean = 6

Step 2: Calculate the deviations and their squares

Observation x_i	**Deviation $x_i - 6$**	**Squared deviation**
2	-4	16
4	-2	4
6	0	0

Observation x_i	Deviation $x_i - 6$	Squared deviation
8	2	4
10	4	16
Total	0	40

Step 3: Calculate the variance

Variance = 40 / 5

Variance = 8

Step 4: Calculate the standard deviation

Standard deviation = √8

Standard deviation ≈ 2.83

Conclusion:

The standard deviation of the examination marks is approximately 2.83 marks.

3.6. Computational Formula

For practical calculations, we can use another formula that is often faster.

Variance = (Sum of squared observations / n) − (Mean)²

Then:

Standard deviation = √Variance

Example

Consider again the data:

2, 4, 6, 8, 10

We have:

Mean = 6

Sum of squared observations:

$2^2 + 4^2 + 6^2 + 8^2 + 10^2$

= 4 + 16 + 36 + 64 + 100

= 220

Therefore:

Variance = 220 / 5 − 6^2

Variance = 44 − 36

Variance = 8

Standard deviation = √8 ≈ 2.83

We obtain the same result as in the previous method.

3.7. Standard Deviation for a Series with Frequencies

When observations are presented in a frequency table, the standard deviation is calculated using the frequencies.

The formulas are:

Mean = $\Sigma n_i x_i$ / N

Variance = $\Sigma n_i (x_i - \bar{x})^2$ / N

Standard deviation = √Variance

Example

Consider the following statistical series:

Value x_i	**Frequency n_i**
2	1
4	2
6	3
8	2
10	2
Total	10

Step 1: Calculate the mean.

Mean = (2×1 + 4×2 + 6×3 + 8×2 + 10×2) / 10

Mean = (2 + 8 + 18 + 16 + 20) / 10

Mean = 64 / 10 = 6.4

Step 2: Calculate the squared deviations and multiply by the frequencies.

x_i	n_i	$x_i - 6.4$	$(x_i - 6.4)^2$	$n_i(x_i - 6.4)^2$
2	1	-4.4	19.36	19.36
4	2	-2.4	5.76	11.52
6	3	-0.4	0.16	0.48
8	2	1.6	2.56	5.12
10	2	3.6	12.96	25.92
Total	10			62.40

Step 3: Calculate the variance.

Variance = 62.40 / 10

Variance = 6.24

Step 4: Calculate the standard deviation.

Standard deviation = √6.24

Standard deviation ≈ 2.50

3.8. Comparison of Two Groups

Consider the following examination marks:

Group A: 8, 9, 10, 11, 12

Group B: 2, 6, 10, 14, 18

Both groups have the same arithmetic mean:

Mean A = Mean B = 10

For Group A:

Variance = 2

Standard deviation = √2 ≈ 1.41

For Group B:

Variance = 32

Standard deviation = √32 ≈ 5.66

Group	Mean	Variance	Standard Deviation
A	10	2	1.41
B	10	32	5.66

Interpretation

Group B has a larger standard deviation than Group A.

Therefore, the marks in Group B are more dispersed around their mean.

This example shows that two groups may have the same mean but different levels of dispersion.

3.9. Important Properties

1. The standard deviation is always non-negative.
2. It is equal to zero if all observations are identical.
3. It is expressed in the same unit as the original observations.
4. A larger standard deviation indicates greater dispersion.
5. It is sensitive to extreme values because deviations are squared.

3.10. Exercises

Exercise 1

The number of hours studied by five students is:

2, 4, 5, 5, 9

1. Calculate the arithmetic mean.
2. Calculate the variance.
3. Calculate the standard deviation.
4. Interpret the result.

Exercise 2

Two laboratory series have the same mean temperature of 20°C.

- Series A: standard deviation = 0.4°C.
- Series B: standard deviation = 1.2°C.

1. Which series has greater dispersion?
2. Explain your answer.

4. Coefficient of Variation

4.1. Introduction

The standard deviation measures the absolute dispersion of a statistical series.

However, when comparing two series with different means, the standard deviation alone may not be sufficient.

For example, a standard deviation of 5 may represent a large dispersion for a series with a mean of 10, but a smaller relative dispersion for a series with a mean of 100.

To solve this problem, we introduce another measure called the coefficient of variation.

The coefficient of variation allows us to compare the relative dispersion of different statistical series.

4.2. Definition

The coefficient of variation (CV) is a relative measure of dispersion.

It represents the standard deviation as a percentage of the arithmetic mean.

Unlike the standard deviation, which measures absolute dispersion, the coefficient of variation measures dispersion relative to the mean.

4.3. Formula

CV = (Standard Deviation / Arithmetic Mean) × 100%

Or:

$CV = (\sigma / \bar{x}) \times 100\%$

Where:

- CV: coefficient of variation.
- σ: standard deviation.

- $\bar{x}$: arithmetic mean.

The coefficient of variation is generally expressed as a percentage.

It is most appropriate when the mean is positive and nonzero and the data have a meaningful zero.

4.4. Interpretation

The coefficient of variation allows us to compare the relative variability of different statistical series.

- A small coefficient of variation indicates low relative dispersion.
- A large coefficient of variation indicates high relative dispersion.

When comparing two series, the series with the smaller coefficient of variation has lower relative dispersion.

4.5. Detailed Example 1

Consider two groups of students with the following statistical characteristics:

Group Mean Standard Deviation

A 10 2

B 20 5

Step 1: Calculate the coefficient of variation of Group A

CV(A) = (2 / 10) × 100

CV(A) = 20%

Step 2: Calculate the coefficient of variation of Group B

CV(B) = (5 / 20) × 100

CV(B) = 25%

Step 3: Compare the results

Group Mean Standard Deviation CV

A 10 2 20%

B 20 5 25%

Interpretation

The coefficient of variation of Group B is greater than that of Group A.

Therefore, Group B has greater relative dispersion.

4.6. Detailed Example 2: Comparison of Two Statistical Series

Consider the following two series:

Series A: 8, 9, 10, 11, 12

Series B: 10, 15, 20, 25, 30

For Series A

Mean = (8 + 9 + 10 + 11 + 12) / 5

Mean = 50 / 5 = 10

Variance = 2

Standard deviation = √2 ≈ 1.414

Coefficient of variation:

CV(A) = (1.414 / 10) × 100

CV(A) ≈ 14.14%

For Series B

Mean = (10 + 15 + 20 + 25 + 30) / 5

Mean = 100 / 5 = 20

Variance = 50

Standard deviation = √50 ≈ 7.071

Coefficient of variation:

CV(B) = (7.071 / 20) × 100

CV(B) ≈ 35.36%

Comparison

Series Mean Standard Deviation Coefficient of Variation

Series	Mean	Standard Deviation	Coefficient of Variation
A	10	1.414	14.14%
B	20	7.071	35.36%

Interpretation

The coefficient of variation of Series B is greater than that of Series A.

Therefore, Series B has greater relative dispersion.

This example illustrates why the coefficient of variation is useful when comparing series with different means.

4.7. Applications

The coefficient of variation is useful in several situations:

1. Comparing the relative variability of examination marks in two classes.
2. Comparing the consistency of measurements obtained from two experiments.
3. Comparing the relative variability of physical measurements with different scales, when the data satisfy the assumptions for using CV.
4. Studying the consistency of measurements or production processes.

4.8. Limitations

The coefficient of variation should be used carefully.

- It is undefined when the mean is zero.
- It can become misleading when the mean is very close to zero.
- It is generally not suitable for data with a negative mean.
- It should be used for quantities with a meaningful zero.

4.9. Exercises

Exercise 1

Two groups of students obtained the following statistical results:

Group	Mean	Standard Deviation
A	12	2

Group Mean Standard Deviation

B 20 5

Questions:

1. Calculate the coefficient of variation for Group A.
2. Calculate the coefficient of variation for Group B.
3. Compare the relative dispersion of the two groups.
4. Which group has the smaller relative variability?

Exercise 2

Two machines produce rods with the following characteristics:

Machine Mean Length Standard Deviation

A 50 cm 1 cm

B 100 cm 1.5 cm

Questions:

1. Calculate the coefficient of variation for each machine.
2. Compare their relative variability.
3. Explain why comparing only the standard deviations may give a different perspective.

Summary of the Two Sections

Measure	**What does it measure?**	**Unit**
Variance	Absolute dispersion around the mean	Squared unit
Standard Deviation	Absolute dispersion around the mean	Same unit as data
Coefficient of Variation	Relative dispersion compared with the mean	Percentage

Important distinction:

The standard deviation is used to describe the absolute dispersion of a statistical series, while the coefficient of variation is useful for comparing relative dispersion between series with different means.

Chapter 2: Measures of Central Tendency and Dispersion

Part II: Analysis of Two Statistical Series

Section 5: Introduction to Two Statistical Series

5.1. Introduction

In the previous part, we studied statistical series involving one variable, such as examination marks, students' heights, or the number of hours spent studying.

However, in many practical situations, we are interested in studying the relationship between two variables.

For example:

- The relationship between the number of hours studied and examination marks.
- The relationship between the temperature of a substance and its electrical resistance.
- The relationship between the mass of an object and its volume.
- The relationship between the time spent studying and the final examination result.

In these situations, each observation consists of two values measured on the same individual or experimental unit.

This type of data is called a bivariate statistical series (série statistique à deux variables).

5.2. Definition

A bivariate statistical series is a statistical series in which two variables are observed simultaneously for each individual or experimental unit.

The two variables are generally denoted by X and Y.

Each individual is represented by a pair of values:

(x_i, y_i)

Where:

- x_i is the value of variable X for individual i.
- y_i is the corresponding value of variable Y for the same individual.

The complete series is written as:

$(x_1, y_1), (x_2, y_2), ..., (x\square , y\square)$

The number of observations is n.

5.3. Example 1: Study Hours and Examination Marks

Consider the following data collected from five students:

Student	Study Hours (X)	Examination Mark (Y)
1	1	8
2	2	10
3	3	12
4	4	14
5	5	16

In this example:

- X represents the number of hours studied.
- Y represents the examination mark.

Each student is represented by a pair:

(1, 8), (2, 10), (3, 12), (4, 14), (5, 16)

For example, the pair (3, 12) means that the third student studied for 3 hours and obtained a mark of 12.

Important: The two values in each pair must belong to the same student. We cannot separate the study hours from their corresponding marks.

5.4. Marginal Statistical Series

From a bivariate statistical series, we can study each variable separately.

The series of values of X is called the marginal series of X.

The series of values of Y is called the marginal series of Y.

For the previous example:

Marginal series of X:

1, 2, 3, 4, 5

Marginal series of Y:

8, 10, 12, 14, 16

We can calculate the mean of each variable separately.

Mean of X:

$\bar{x} = (1 + 2 + 3 + 4 + 5) / 5$

$\bar{x} = 15 / 5 = 3$

Mean of Y:

$\bar{y} = (8 + 10 + 12 + 14 + 16) / 5$

$\bar{y} = 60 / 5 = 12$

Therefore:

- Mean study time = 3 hours.
- Mean examination mark = 12.

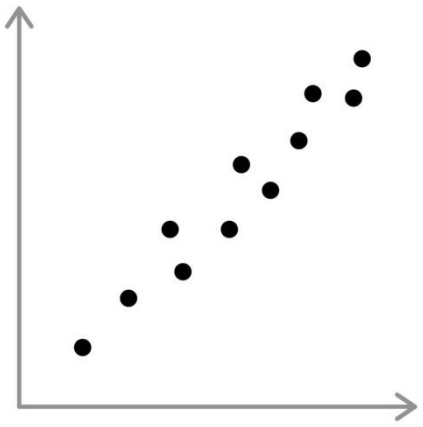
5.5. Scatter Diagram

A scatter diagram (diagramme de dispersion) is a graphical representation of a bivariate statistical series.

Each pair (x_i, y_i) is represented by a point in a coordinate system:

- The horizontal axis represents variable X.
- The vertical axis represents variable Y.

The scatter diagram helps us observe the possible relationship between the two variables.



For the study-hours example, the points are:

(1, 8), (2, 10), (3, 12), (4, 14), (5, 16)

The points form an increasing pattern: when study time increases, examination marks also increase in this small dataset.

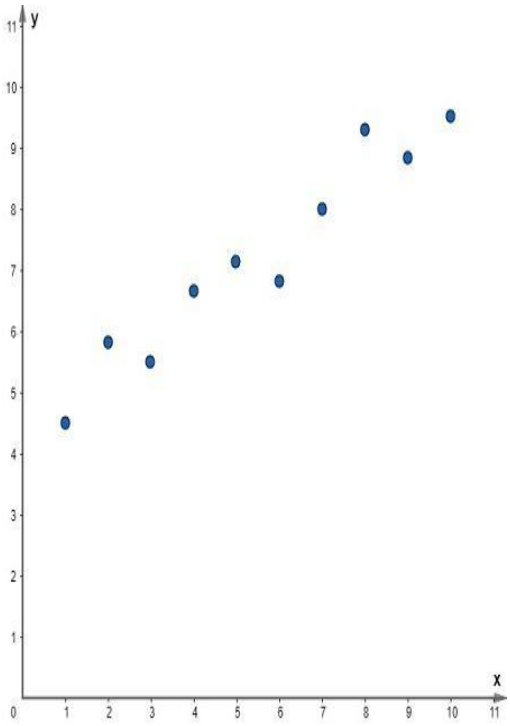
This is an observed association in the example; it does not by itself prove that study time causes the marks to increase.

5.6. Different Types of Relationships

A scatter diagram can show different patterns.

1. Positive relationship

When X increases, Y tends to increase.

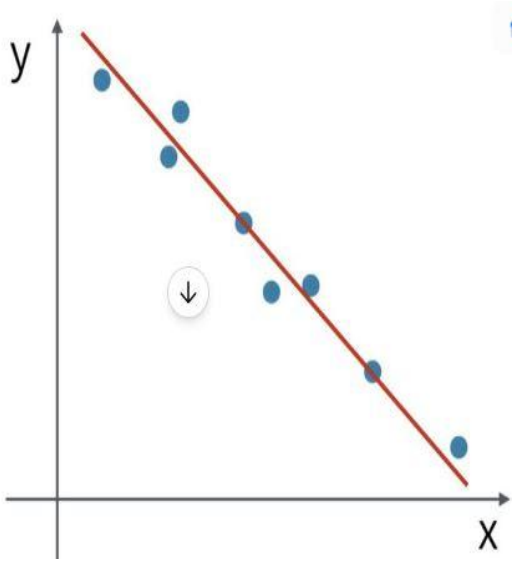


Example: study time and examination marks in a dataset where higher study times tend to accompany higher marks.

2. Negative relationship

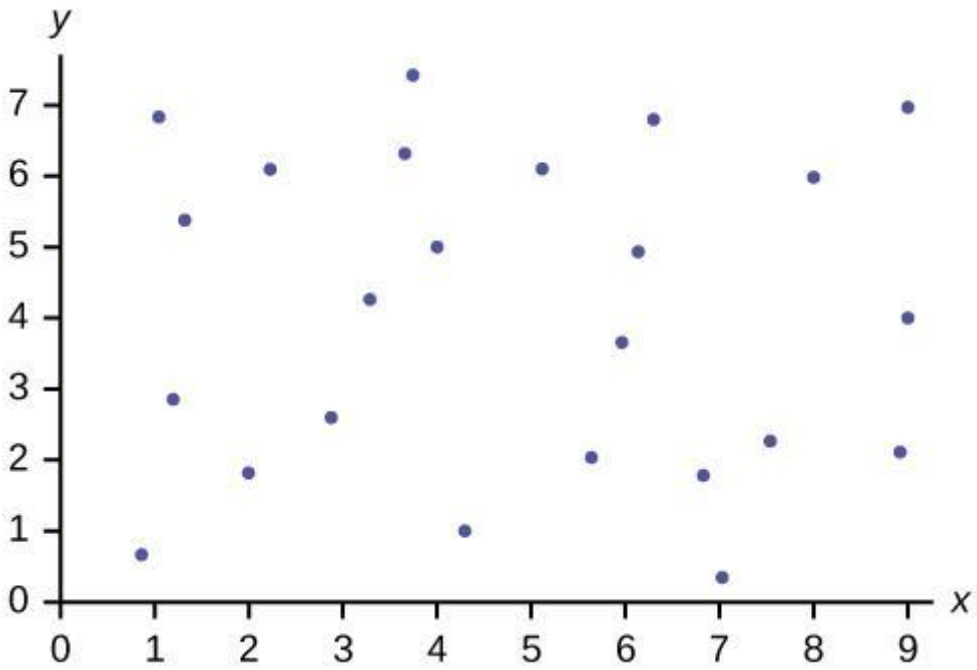
When X increases, Y tends to decrease.

Example: for a fixed distance, average travel time decreases as average speed increases.



3. No clear linear relationship

The points do not show a clear increasing or decreasing linear pattern.



This does not necessarily mean that the variables are unrelated; a nonlinear relationship may still exist.

5.8. Exercises

Exercise 1: Identification of Variables

A physics teacher records the study time and examination marks of six students.

Student	Study Hours (X)	Mark (Y)
1	1	7
2	2	9
3	3	11
4	4	13
5	5	14
6	6	17

Questions:

1. Identify the two variables X and Y.
2. What is the number of observations?
3. Write the six pairs (x_i, y_i).
4. Write the marginal series of X.
5. Write the marginal series of Y.
6. Calculate the arithmetic mean of X.
7. Calculate the arithmetic mean of Y.
8. Draw the scatter diagram.
9. Describe the general pattern observed.

Exercise 2: Physical Measurements

A student measures the mass and volume of five objects made of the same material.

Object	Volume (cm^3)	Mass (g)
1	2	16
2	4	32
3	6	48
4	8	64
5	10	80

Questions:

1. Identify X and Y.
2. Write the corresponding pairs.
3. Calculate the mean of X and the mean of Y.

4. Draw the scatter diagram.
5. Describe the relationship between volume and mass.

End of Section 5.

Next, we will study Section 6: Covariance, including its definition, formulas, detailed calculations, and interpretation.

You are absolutely right to point this out. The explanation may confuse students if we do not clearly distinguish between the student's identification number and the number of study hours.

Let us clarify the difference.

1. Student number is only an identification

In our table:

Student ID	Study hours (X)	Examination mark (Y)
1	1	8
2	2	10
3	3	12
4	4	14
5	5	16

The first column identifies the students. It is not a variable in our statistical study.

The variables we study are:

- X: Number of study hours.
- Y: Examination mark.

2. What does the point (3, 12) mean?

It means that a student studied 3 hours and obtained 12 marks.

It does not mean that student number 3 obtained 12 marks.

This distinction is essential.

3. How should we present the example to students?

I suggest removing the student identification column from the graph and presenting the data directly as pairs:

Study hours (X)	**Examination mark (Y)**
1	8
2	10
3	12
4	14
5	16

Then explain:

Each row gives one observation. The first value is the number of study hours, and the second value is the corresponding examination mark.

The points to plot are:

(1, 8), (2, 10), (3, 12), (4, 14), (5, 16).

For teaching purposes, I agree that this presentation is clearer and avoids confusing the student identification number with the statistical variable.

The scatter diagram (scatter plot) is useful because it allows us to study the relationship between two statistical variables.

Instead of looking only at numbers in a table, we can see how the two variables change together.

Let us explain this using examples relevant to physics students.

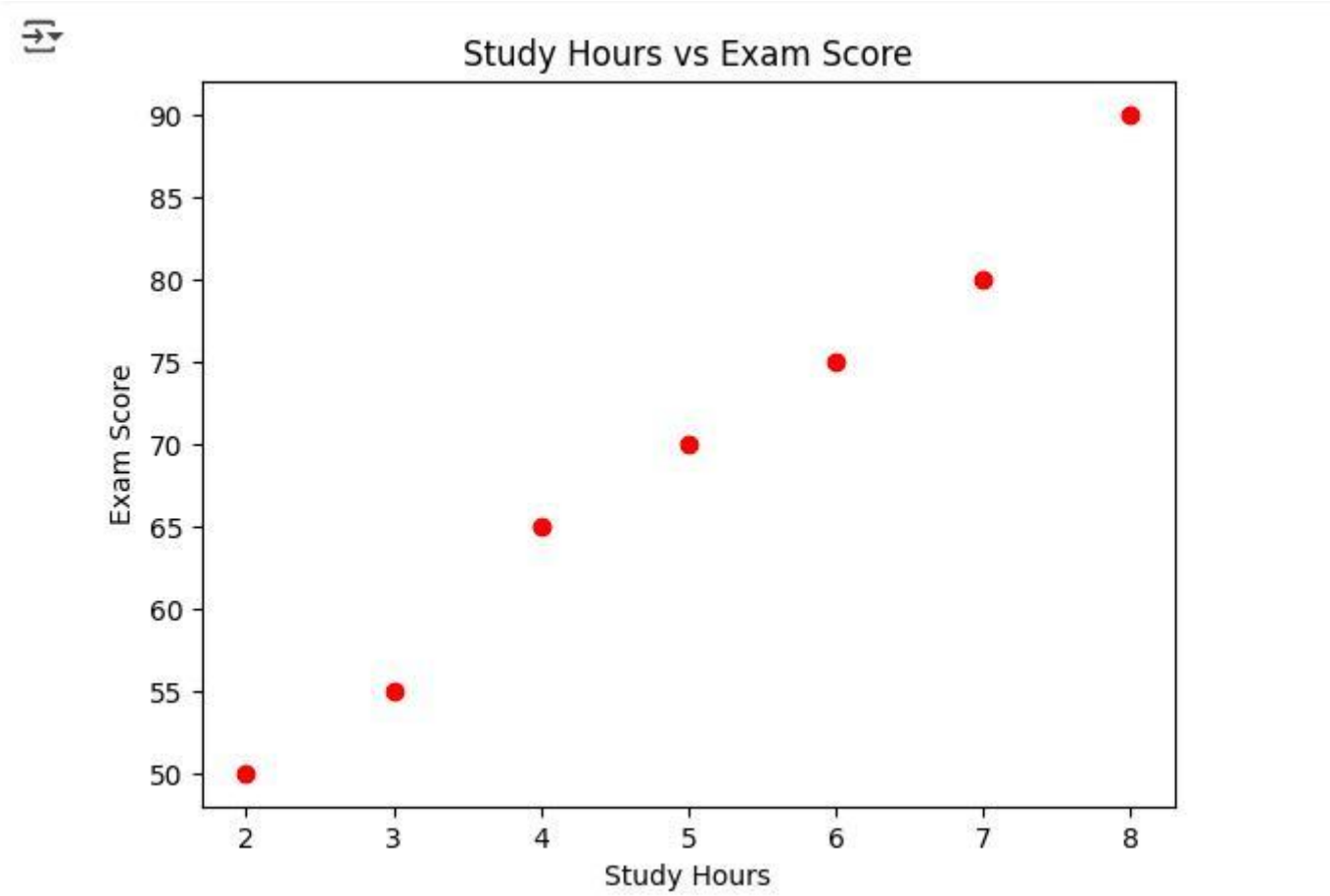
1. Example: Study hours and examination marks

Suppose we have the following data:

Study hours (X)	**Examination mark (Y)**
1	8
2	10

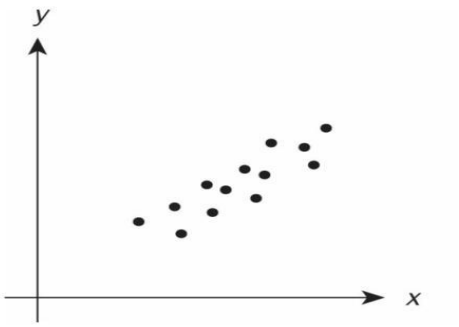
Study hours (X)	Examination mark (Y)
3	12
4	14
5	16

When we plot these points, we observe that the marks increase as study hours increase.



The diagram helps us observe a positive relationship between the two variables.

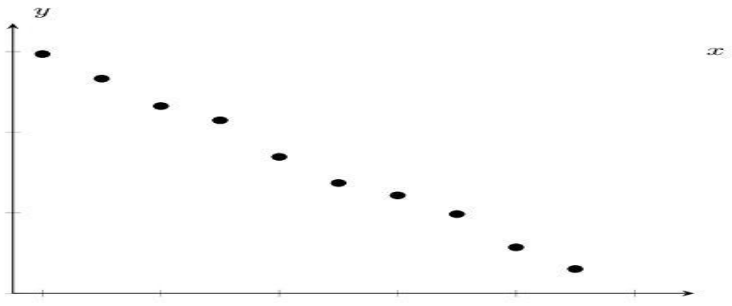
A. Positive relationship



When X increases, Y tends to increase.

Example: Study hours and examination marks.

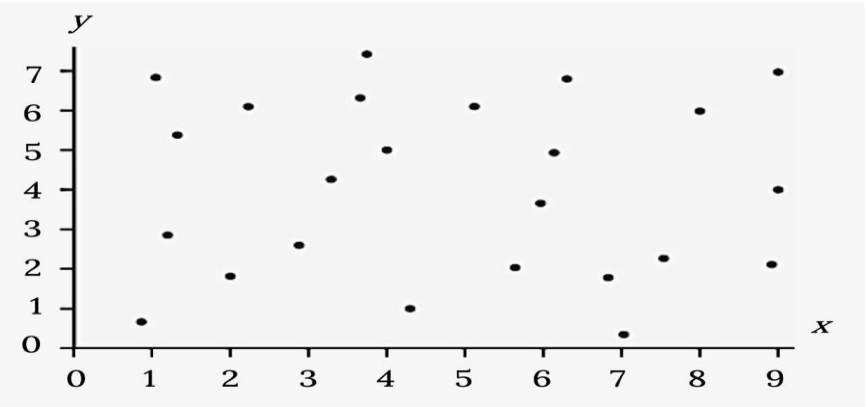
B. Negative relationship



When X increases, Y tends to decrease.

Example: For a fixed distance, the time needed to travel it decreases as speed increases.

C. No clear relationship



The points do not show a clear increasing or decreasing pattern.

In this case, knowing X does not help us identify a clear trend in Y.

3. Why is this diagram important in statistics?

4. Application in physics

Consider the relationship between the mass of a substance and its volume.

For a given material, increasing the volume generally increases the mass.

The scatter diagram can help us visualize this relationship and investigate whether the points follow an approximately straight line.

A scatter diagram helps us:

1. Visualize the relationship between two variables.
2. Identify whether the relationship is positive, negative, or unclear.
3. Observe whether the points are close to a straight line or widely scattered.
4. Detect unusual observations that are far from the other points.
5. Prepare for the study of covariance and correlation.

   Remark: A scatter diagram shows an association between two variables, but it does not by itself prove that one variable causes the other.

This leads naturally to the next statistical concepts:

- Covariance: measures how two variables vary together.
- Correlation coefficient: measures the strength and direction of a linear relationship.
- Linear regression: allows us to describe the relationship using a mathematical equation.

We can now begin the next section of Chapter 2, following the scatter diagram.

I suggest presenting the three concepts progressively, with simple definitions, practical examples, and detailed calculations.

Section 6: Covariance, Correlation and Linear Regression

Introduction

In the previous section, we learned how to represent two statistical variables using a scatter diagram.

However, a scatter diagram gives us mainly a visual description of the relationship between two variables.

We may want to answer more precise questions:

- Do the two variables tend to increase or decrease together?
- How strong is their linear relationship?
- Can we describe this relationship using a mathematical equation?

To answer these questions, we introduce three statistical tools:

1. Covariance
2. Correlation coefficient
3. Linear regression

6.1 Covariance

1. Definition

Covariance (covariance) is a statistical measure that indicates how two variables vary together.

It helps us identify the direction of their relationship.

- Positive covariance: the two variables tend to increase or decrease together.
- Negative covariance: when one variable increases, the other tends to decrease.
- Zero covariance: there is no linear co-variation indicated by the covariance.

2. Formula

For n pairs of observations (x_i, y_i), the covariance is:

$Cov(X,Y) = \Sigma(xi-x^-)(yi-y^-)\Sigma(x_i - \bar{x})(y_i - \bar{y}) / n$

Where:

- $\bar{x}$ is the mean of variable X.
- $\bar{y}$ is the mean of variable Y.
- n is the number of observations.

3. Simple example

Consider the following data:

Study hours (X)	Examination mark (Y)
1	8
2	10
3	12
4	14

Study hours (X)	Examination mark (Y)
5	16

First, calculate the means:

Mean of X = (1 + 2 + 3 + 4 + 5) / 5 = 3

Mean of Y = (8 + 10 + 12 + 14 + 16) / 5 = 12

Now construct the calculation table:

X	Y	X – 3	Y – 12	(X – 3)(Y – 12)
1	8	-2	-4	8
2	10	-1	-2	2
3	12	0	0	0
4	14	1	2	2
5	16	2	4	8
Total				20

Therefore:

Cov(X,Y) = 20 / 5 = 4

The covariance is positive, indicating that the two variables tend to increase together.

Remark: The numerical value of covariance depends on the units of measurement. This is why we introduce the correlation coefficient.

6.2 Correlation Coefficient

1. Definition

The correlation coefficient (coefficient de corrélation) is a statistical measure used to determine the direction and strength of the linear relationship between two quantitative variables.

It is denoted by the letter r.

The correlation coefficient helps us answer two questions:

1. Do the two variables tend to increase or decrease together?
2. How strong is their linear relationship?

The value of the correlation coefficient always belongs to the interval:

$-1 \leq r \leq +1$

2. Formula

The correlation coefficient is calculated using the following formula:

$r = Cov(X,Y) / (\sigma_x \times \sigma_\gamma)$

Where:

- $Cov(X,Y)$ is the covariance between X and Y.
- σ_x is the standard deviation of X.
- σ_γ is the standard deviation of Y.

Unlike covariance, the correlation coefficient has no unit of measurement. This makes it possible to compare the strength of linear relationships between variables measured in different units.

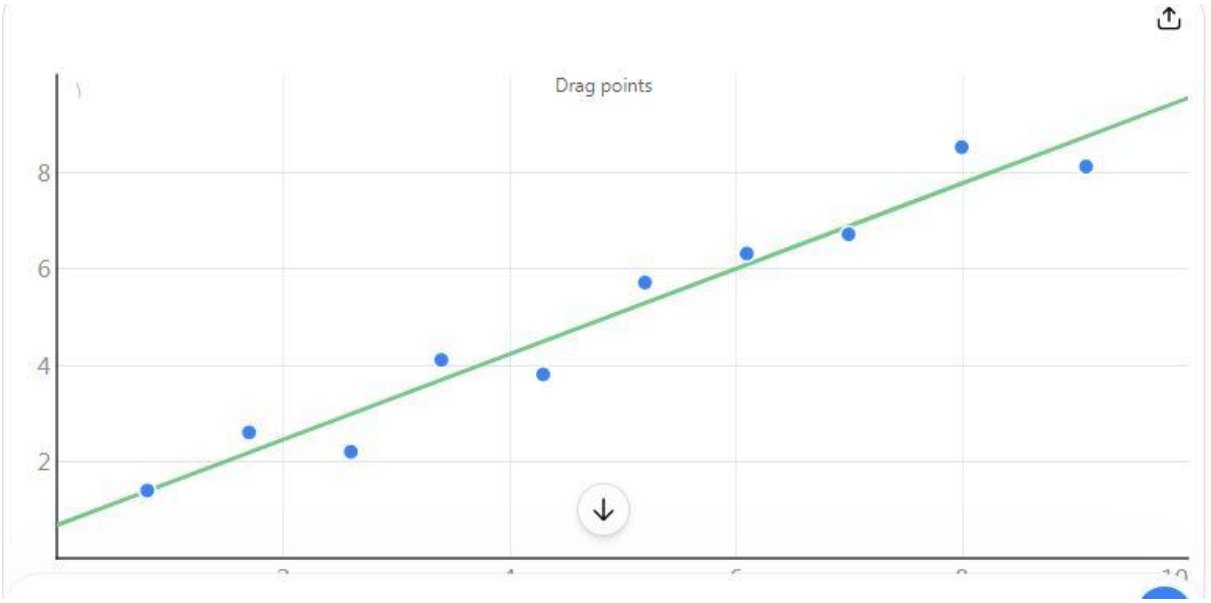
3. Interpretation of the Correlation Coefficient

The sign of r indicates the direction of the relationship, while its absolute value indicates the strength of the linear relationship.

A. Positive correlation (r > 0)

When r is positive, the two variables tend to vary in the same direction.

- When X increases, Y tends to increase.
- When X decreases, Y tends to decrease.



Linear correlation is positive (r = 0.98).

Example: Study hours and examination marks may have a positive correlation because students who study more hours tend, on average, to obtain higher marks.

B. Negative correlation (r < 0)

When r is negative, the two variables tend to vary in opposite directions.

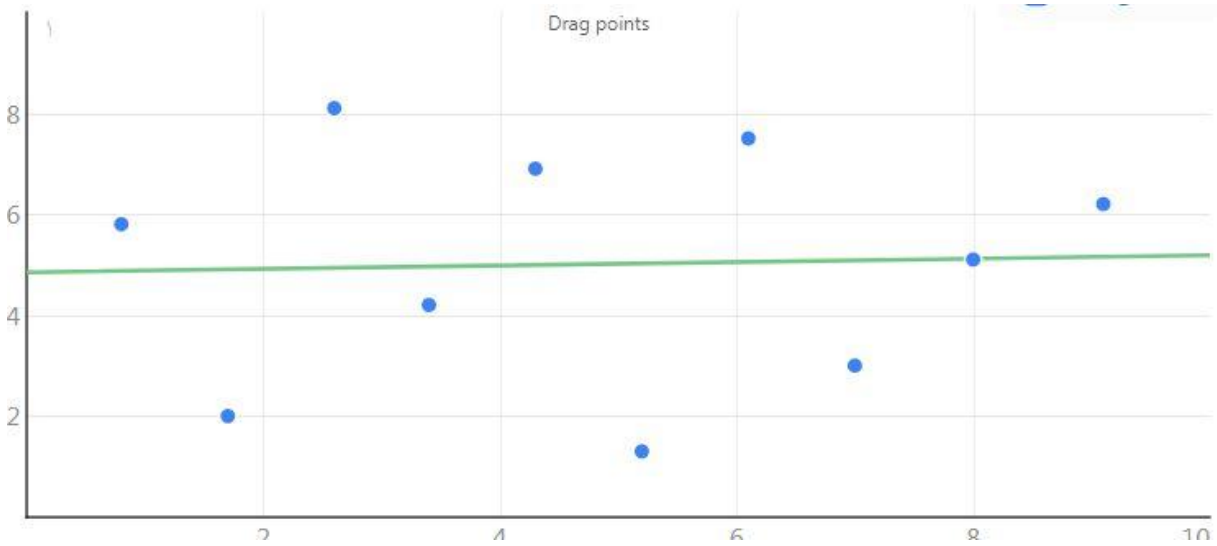
- When X increases, Y tends to decrease.
- When X decreases, Y tends to increase.



Example: For a fixed distance, when the speed of a vehicle increases, the travel time decreases.

C. No linear correlation (r ≈ 0)

When r is close to zero, there is little or no linear relationship between the two variables.



This does not necessarily mean that the variables are completely unrelated. A nonlinear relationship may exist.

4. Strength of the Linear Relationship

The absolute value of r indicates the strength of the linear relationship.

Correlation coefficient	**Interpretation**
r = +1	Perfect positive linear correlation
0.7 < r < 1	Strong positive linear correlation
0.3 < r ≤ 0.7	Moderate positive linear correlation
0 < r ≤ 0.3	Weak positive linear correlation
r = 0	No linear correlation
−0.3 ≤ r < 0	Weak negative linear correlation
−0.7 ≤ r < −0.3	Moderate negative linear correlation

Correlation coefficient	Interpretation
$-1 < r < -0.7$	Strong negative linear correlation
$r = -1$	Perfect negative linear correlation

Note: These strength categories are approximate conventions used for teaching. Their interpretation depends on the context and the type of data.

5. Worked Example: Study Hours and Examination Marks

Consider the following data:

Study hours (X)	Examination mark (Y)
1	8
2	10
3	12
4	14
5	16

We previously calculated:

- Mean of X: $\bar{X} = 3$
- Mean of Y: $\bar{Y} = 12$
- Covariance: $Cov(X,Y) = 4$
- Variance of X: $V(X) = 2$
- Variance of Y: $V(Y) = 8$

Step 1: Calculate the standard deviations

Standard deviation of X:

$\sigma_x = \sqrt{2} \approx 1.414$

Standard deviation of Y:

$\sigma_y = \sqrt{8} \approx 2.828$

Step 2: Calculate the correlation coefficient

Using the formula:

$r = Cov(X,Y) / (\sigma_x \times \sigma_y)$

Substitute the values:

$r = 4 / (\sqrt{2} \times \sqrt{8})$

Since $\sqrt{2} \times \sqrt{8} = \sqrt{16} = 4$:

$r = 4 / 4$

r = +1

Perfect positive linear correlation

6. Interpretation of the Result

We obtained r = +1.

This means that the two variables have a perfect positive linear relationship in this dataset.

In our example:

- As study hours increase, examination marks increase.
- All five observations lie exactly on a straight line.
- The relationship is positive and perfectly linear.

However, this result does not prove that studying more hours is the cause of higher marks. Correlation describes an association, not necessarily causation.

7. Important Remarks

1. The correlation coefficient is always between −1 and +1.
2. The sign indicates the direction of the linear relationship.
3. The absolute value indicates its strength.
4. A correlation coefficient close to zero indicates a weak linear relationship, but does not exclude a nonlinear relationship.
5. Correlation does not necessarily imply causation.

8. Summary

Value of r	Direction	Strength
Close to +1	Positive	Strong
Close to 0	No clear linear direction	Weak linear relationship
Close to −1	Negative	Strong

Value of r	Direction	Strength
+1 or −1	Positive or negative	Perfect linear relationship

Conclusion: The correlation coefficient provides a numerical measure of the direction and strength of the linear relationship between two quantitative variables. It complements the visual information obtained from the scatter diagram.

6.3 Linear Regression

1. Definition

Linear regression (régression linéaire) is a statistical method used to describe the linear relationship between two variables by an equation of a straight line.

The equation is:

$Y = aX + b$

Where:

- X is the explanatory variable.
- Y is the variable we want to explain or estimate.
- a is the slope of the line.
- b is the intercept with the Y-axis.

The important property of the regression line is that it passes through the point formed by the means of the two variables: ($\bar{X}$, $\bar{Y}$). This means that when $X = \bar{X}$, the corresponding value on the regression line is $Y = \bar{Y}$. Therefore, we substitute $\bar{X}$ and $\bar{Y}$ into the equation: $\bar{Y} = a\bar{X} + b$

2. Example

Using the same data:

Study hours (X)	Examination mark (Y)
1	8
2	10
3	12
4	14
5	16

1. Start with the general equation

The equation of a straight line is:

$Y = aX + b$

Where:

- a is the slope.
- b is the intercept with the vertical axis.

We want to determine the values of a and b.

The regression line always passes through the point formed by the two means:

$(\bar{X}, \bar{Y})$

In other words, when we replace X by its mean $\bar{X}$, the corresponding value on the regression line is $\bar{Y}$.

Therefore, we can write:

$\bar{Y} = a\bar{X} + b$

This is the key step.

3. Isolate b

Starting from:

$\bar{Y} = a\bar{X} + b$

Subtract $a\bar{X}$ from both sides:

$\bar{Y} - a\bar{X} = b$

Thus:

$b = \bar{Y} - a\bar{X}$

This is the formula used to calculate the intercept.

4. Apply it to our example

We previously calculated:

- $\bar{X} = 3$
- $\bar{Y} = 12$

- a = 2

Substituting:

$b = 12 - 2 \times 3$

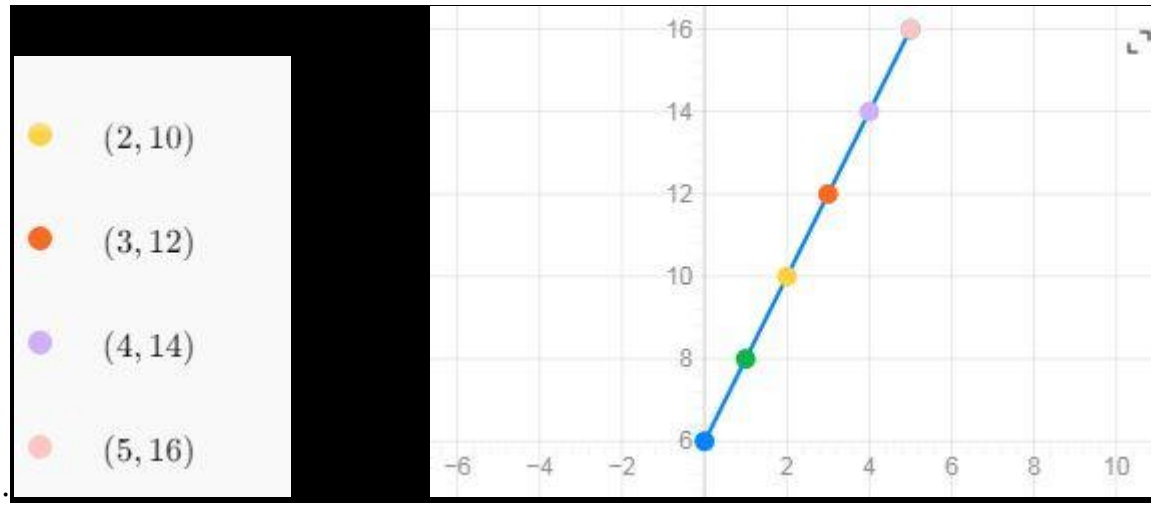
$b = 12 - 6$

$b = 6$

Therefore, the regression equation is:

$Y = 2X + 6$

5. Geometrical interpretation



In our example, we obtained:

Cov(X,Y) = 4

Interpretation

Since the covariance is positive, the two variables tend to vary in the same direction.

In our example:

- When study hours increase, examination marks also increase.
- When study hours decrease, examination marks tend to decrease.

Therefore, we observe a positive relationship between study hours and examination marks.

Important: The value of covariance alone does not tell us whether the relationship is weak or strong because its numerical value depends on the units of measurement.

6.3 Linear Regression: Interpretation of Results

We obtained the regression equation:

$Y = 2X + 6$

Interpretation of the equation

The equation contains two important coefficients.

1. Slope: a = 2

The slope indicates the change in the predicted value of Y when X increases by one unit.

In our example:

For each additional study hour, the predicted examination mark increases by 2 marks.

2. Intercept: b = 6

The intercept represents the predicted value of Y when X = 0.

In our example:

If a student studies 0 hours, the equation predicts a mark of 6.

However, this is a mathematical prediction from the model, not necessarily a realistic conclusion about students.

Application: Prediction

If a student studies 6 hours:

$Y = 2 \times 6 + 6 = 18$

Interpretation:

According to the regression equation, the predicted examination mark for 6 study hours is 18.

This is an estimate, not a guaranteed result.

Final comparison: What do the results tell us?

Concept	Numerical result	Interpretation
Covariance	4	The variables tend to vary in the same direction
Correlation	r = 1	Perfect positive linear relationship in this dataset

Concept	Numerical result	Interpretation
Regression	$Y = 2X + 6$	Each additional study hour corresponds to an increase of 2 in the predicted mark

Chapter 2: Measures of Central Tendency and Dispersion

Section 10: Exercises and Applications

The following exercises are designed to help students apply the statistical concepts studied in this chapter, including measures of dispersion, covariance, correlation, and linear regression.

Exercise 1: Range, Variance and Standard Deviation

The following data represent the examination marks obtained by 5 students:

12, 14, 10, 16, 8

1. Calculate the mean.
2. Calculate the range.
3. Calculate the variance.
4. Calculate the standard deviation.
5. Interpret the standard deviation.

Exercise 2: Comparison of Two Groups

Two groups of students obtained the following marks:

Group A: 8, 9, 10, 11, 12

Group B: 2, 6, 10, 14, 18

1. Calculate the mean of each group.
2. Calculate the range of each group.
3. Calculate the variance of each group.
4. Calculate the standard deviation of each group.
5. Compare the dispersion of the two groups.
6. Explain why two groups can have the same mean but different standard deviations.

Exercise 3: Coefficient of Variation

Two machines produce components whose lengths are measured in millimeters.

Machine	Mean length (mm)	Standard deviation (mm)
A	20	2
B	25	4

1. Calculate the coefficient of variation for each machine.
2. Express each result as a percentage.
3. Interpret the results.
4. Compare the relative dispersion of the two machines.

Exercise 4: Covariance

The following table gives the number of hours spent studying and the corresponding examination marks for five students.

Study hours (X)	Examination mark (Y)
1	8
2	10
3	12
4	14
5	16

1. Calculate the mean of X and the mean of Y.
2. Construct a calculation table containing $X - \bar{X}$, $Y - \bar{Y}$, and their product.
3. Calculate the covariance.
4. Interpret the sign of the covariance.

Exercise 5: Correlation Coefficient

For the data in Exercise 4, we have:

- Cov(X,Y) = 4
- Variance of X = 2
- Variance of Y = 8

1. Calculate the standard deviation of X and Y.
2. Calculate the correlation coefficient.
3. Interpret its sign and value.
4. What does the result tell us about the linear relationship between study hours and examination marks?

Exercise 6: Linear Regression

Consider the following data:

X	Y
1	5
2	7
3	9
4	11
5	13

1. Calculate the means $\bar{X}$ and $\bar{Y}$.
2. Calculate the covariance Cov(X,Y).
3. Calculate the variance V(X).
4. Determine the slope a using:

$a = Cov(X,Y) / V(X)$

5. Determine the intercept b using:

$b = \bar{Y} - a\bar{X}$

6. Write the regression equation $Y = aX + b$.
7. Interpret the slope and intercept.
8. Estimate the value of Y when X = 6.

Exercise 7: Application in Physics

An experiment is conducted to study the relationship between the volume and mass of a substance.

Volume X (cm^3)	Mass Y (g)
2	16
4	32
6	48
8	64

Volume X (cm^3)	Mass Y (g)
10	80

1. Represent the data using a scatter diagram.
2. Describe the relationship between volume and mass.
3. Calculate the mean of X and Y.
4. Calculate the covariance.
5. Calculate the correlation coefficient.
6. Determine the regression equation of Y on X.
7. Interpret the slope in terms of the physical quantities.
8. Estimate the mass corresponding to a volume of 12 cm^3.

Section 12: Final Review Exercises

These exercises combine the principal concepts studied throughout Chapter 2.

Exercise 1: Measures of Dispersion

The following data represent the temperatures, in °C, recorded during five consecutive measurements:

18, 20, 22, 24, 26

1. Calculate the arithmetic mean.
2. Calculate the range.
3. Calculate the variance.
4. Calculate the standard deviation.
5. Calculate the coefficient of variation.
6. Interpret the results.

Exercise 2: Comparison of Two Statistical Series

Two laboratories measure the same physical quantity. The results are:

Laboratory A: 48, 49, 50, 51, 52

Laboratory B: 40, 45, 50, 55, 60

1. Calculate the mean of each series.
2. Calculate the range of each series.
3. Calculate the variance and standard deviation of each series.
4. Calculate the coefficient of variation of each series.
5. Compare the dispersion of the two series.
6. Explain which series has measurements more concentrated around its mean, based on the calculated standard deviations.

Exercise 3: Covariance and Correlation

Consider the following observations:

X	Y
1	12
2	10
3	8
4	6
5	4

1. Calculate the means $\bar{X}$ and $\bar{Y}$.
2. Calculate the covariance Cov(X,Y).
3. Calculate the variances V(X) and V(Y).
4. Calculate the correlation coefficient.
5. Interpret the sign and value of the correlation coefficient.
6. Draw the scatter diagram and describe the observed relationship.

Exercise 4: Regression and Prediction

The following data represent the relationship between the temperature X (°C) and a measured physical quantity Y.

X	Y
0	3
1	5
2	7
3	9
4	11

1. Calculate the means of X and Y.
2. Calculate Cov(X,Y).
3. Calculate V(X).
4. Determine the slope a.
5. Determine the intercept b.
6. Write the regression equation.

7. Interpret the slope and intercept.
8. Estimate Y when X = 6.
9. Represent the observations and the regression line on the same graph.

Exercise 5: Comprehensive Application

A physics student investigates the relationship between the time of heating and the temperature of a substance.

The following measurements are obtained:

Time X (min)	Temperature Y (°C)
1	25
2	30
3	35
4	40
5	45

Answer the following questions:

1. Identify the two statistical variables.
2. Calculate their arithmetic means.
3. Calculate the range of each variable.
4. Calculate the variance and standard deviation of each variable.
5. Calculate the covariance.
6. Calculate the correlation coefficient.
7. Interpret the correlation coefficient.
8. Determine the regression equation of Y on X.
9. Interpret the slope and intercept.
10. Estimate the temperature after 6 minutes using the regression equation.
11. Draw the scatter diagram and the regression line.
12. Explain the relationship observed between heating time and temperature.