Introduction

Statistics is the branch of mathematics devoted to the systematic collection, organization, analysis, interpretation, and presentation of data. Its primary goal is to extract meaningful information from numerical observations, to describe phenomena, and to support decision-making under uncertainty. Statistics provides methods to summarize large datasets, identify patterns and relationships, and infer conclusions about populations based on sample data. It is widely used across disciplines such as economics, engineering, psychology, and the natural sciences, where data-driven reasoning is essential.

In essence, statistical methods can be divided into two main categories:

- **Descriptive Statistics:** concerned with summarizing and presenting data in an informative way using tables, graphs, and summary measures (such as mean, median, variance, and standard deviation).
- **Inferential Statistics:** focused on drawing conclusions or generalizations about a population based on data from a sample. This includes estimation, hypothesis testing, correlation, and regression analysis, all of which rely on probability theory to measure and control uncertainty.

Biostatistics — also known as *biological statistics* — is a specialized field that applies statistical principles and methods to the study of biological, medical, and health-related phenomena. It plays a vital role in experimental design, data analysis, and interpretation within the life sciences. Biostatistics provides scientists with quantitative tools to:

- Design biological and medical experiments and surveys;
- Summarize and visualize experimental data clearly;
- Evaluate hypotheses and measure the strength of relationships between biological variables;
- Quantify the uncertainty associated with observations and conclusions;
- Support evidence-based decision-making in medicine, epidemiology, and public health.

Because biological data often exhibit natural variability and randomness, statistical reasoning becomes indispensable in separating true biological effects from random fluctuations. Thus, biostatistics forms a critical link between theoretical biology and empirical observation — transforming data into scientific understanding and guiding progress in research.

In summary, statistics provides the theoretical and methodological foundation for data analysis, while biostatistics extends this foundation to address the specific challenges of

biological systems, medical research, and public health investigations. Together, they constitute essential tools for modern scientific inquiry, ensuring that conclusions drawn from data are valid, reproducible, and meaningful.

Chapter 1

Reminders of Descriptive Statistics in One and Two Dimensions

1.1 One-Dimensional Descriptive Statistics (Univariate Statistics)

This branch of statistics studies a single variable within a dataset. It aims to summarize and describe data using tables, graphs, and numerical measures.

1.1.1 General Concepts

Population: The set of individuals defined by a given common property. It is generally a very large set.
Example: all lakes within a region, all trees in a forest, or all air samples collected in an industrial area.

Sample: A representative subset of the population used for analysis when studying the entire population is impractical.
Example: measuring the pH level of water in 20 randomly selected lakes to represent the overall water quality of the region.

Statistical Unit: The basic element of the population on which an observation or measurement is made.
Example: each tree measured for height, each water sample analyzed for nitrate concentration, or each bird counted in a biodiversity survey.

Statistical Variable: A characteristic observed or measured for each statistical unit.
Example: temperature of seawater, salinity level, air pollutant concentration, or the number of fish species per site.

Types of Variables

1. **Qualitative Variables:** Describe non-numerical or categorical characteristics.

   - **Ordinal (semi-quantitative):** Can be arranged in a meaningful order.
     Example: water quality classified as "poor", "moderate", or "good"; or soil erosion rated as "low", "medium", or "high".

   - **Nominal (pure qualitative):** Categories without a specific order.
     Example: type of ecosystem (forest, wetland, desert) or dominant vegetation species.

2. **Quantitative Variables:** Represent measurable numerical values.

- **Discrete:** Take countable values.
  Example: number of bird nests in a park, number of polluted sites in a coastal zone, or number of trees per plot.

- **Continuous:** Take any value within a range.
  Example: air temperature, rainfall amount, salinity, or dissolved oxygen concentration in water.

1.1.2 Statistical Tables

A **statistical table** is a structured and numerical representation of collected data that summarizes one or more variables. Each variable is organized into rows and columns, showing its different categories or measured values. Statistical tables are widely used in environmental and ecological studies to organize information and highlight relationships between variables such as temperature, pollution level, or biodiversity index.

Example: Table showing the average concentration of dissolved oxygen in different river sites.

Site	Sampling Date	Water Temperature (°C)	Dissolved Oxygen (mg/L)
River Site A	March 2025	16.2	8.5
River Site B	March 2025	18.7	7.9
River Site C	March 2025	21.0	6.8
River Site D	March 2025	19.5	7.2

1.1.3 Graphical Representations

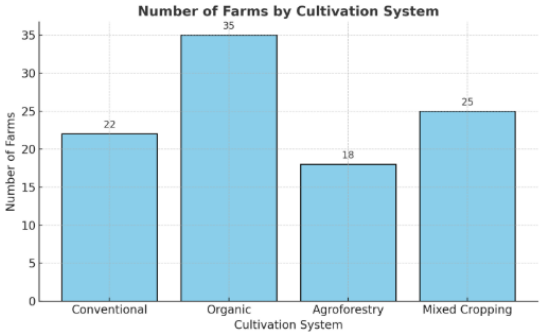
Graphical representations are essential tools in agroecology, as they provide a clear visual summary of data related to agricultural and environmental systems. They enable researchers to easily identify trends, seasonal variations, and patterns in variables such as crop yield, soil fertility, pest occurrence, or rainfall distribution.

a) Bar Chart

A **bar chart** is used to represent discrete variables that take distinct values. In agroecology, it can be employed to compare the number of farms adopting different agricultural practices or the average yield of various crop types.

Example: Number of farms by cultivation system.

Cultivation System	Number of Farms
Conventional	22
Organic	35
Agroforestry	18
Mixed Cropping	25



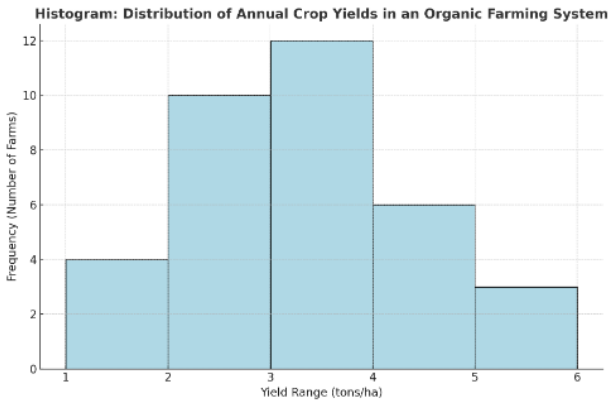
The chart shows that organic and mixed cropping systems are the most common approaches, reflecting a growing trend toward sustainable agricultural practices.

b) Histogram

A **histogram** is used to represent *continuous variables* such as soil pH, rainfall, or crop yield. It illustrates the distribution of these variables over a continuous range, helping to understand their variability in agroecosystems.

Example: Distribution of annual crop yields in an organic farming system.

Yield Range (tons/ha)	Frequency (Number of Farms)
1–2	4
2–3	10
3–4	12
4–5	6
5–6	3



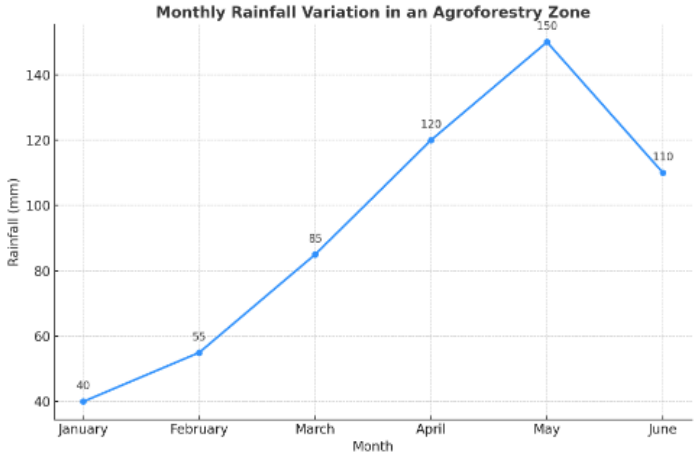
The histogram shows that most farms produce between 2 and 4 tons per hectare, indicating moderate productivity in the studied region.

c) Frequency Polygon

A **frequency polygon** connects the midpoints of histogram bars, providing a clear view of trends or comparisons between datasets. In agroecology, it can be used to represent the seasonal evolution of variables such as pest attacks, rainfall, or soil moisture.

Example: Monthly rainfall variation in an agroforestry zone.

Month	Rainfall (mm)
January	40
February	55
March	85
April	120
May	150
June	110



The frequency polygon highlights the progressive increase in rainfall from January to May, followed by a decrease in June — a typical pattern observed in tropical agroecological zones.

1.1.4 Data Reduction

The process of **data reduction** aims to condense large datasets into a small number of representative numerical values that capture the essential characteristics of the data. These numerical indicators are generally classified into two main categories:

- **Measures of central tendency (parameters of position)**
- **Measures of dispersion (parameters of variability)**

1. Parameters of Position

Definition: Measures of position, also known as *central tendency measures*, describe the central or typical value around which observations are distributed.

1. Arithmetic mean
2. Geometric mean
3. Harmonic mean

4. Quadratic (root mean square) mean
5. Median
6. Mode

a) Arithmetic Mean:

- For **discrete data:**

$$\bar{y} = \frac{\sum f_i y_i}{\sum f_i}$$

  where f_i is the frequency of each observation y_i.

- For **continuous data (grouped data):**

$$\bar{y} = \frac{\sum f_i x_i}{\sum f_i}$$

  where x_i is the class midpoint of each interval.

—

b) Geometric Mean:

- For **discrete or continuous data:**

$$y_g = \left(\prod y_i^{f_i}\right)^{1/\sum f_i}, \quad y_i > 0$$

—

c) Harmonic Mean:

- For **discrete or continuous data:**

$$y_h = \frac{\sum f_i}{\sum \frac{f_i}{y_i}}, \quad y_i \neq 0$$

—

d) Quadratic Mean (Root Mean Square):

- For **discrete or continuous data:**

$$y_q = \sqrt{\frac{\sum f_i y_i^2}{\sum f_i}}$$

—

e) Median:

- For **discrete data:** The median is the value that divides the dataset into two equal parts, such that 50% of the observations are below it and 50% are above it. To determine the median:

  1. Arrange the data in ascending order.
  2. Compute the cumulative frequencies.

3. Identify the value corresponding to the position:

$$\text{Median position} = \frac{n+1}{2}$$

The observation at this position is the median.

- For **continuous (grouped) data:**

$$Me = L + \left(\frac{\frac{n}{2} - F_{<}}{f_m} \right) h$$

where:

- L = lower boundary of the median class,
- $F_{<}$ = cumulative frequency before the median class,
- f_m = frequency of the median class,
- h = class width (amplitude of the class interval),
- n = total number of observations.

—

f) Mode:

- For **discrete data:** The mode is the most frequent value in the dataset.
- For **continuous (grouped) data:**

$$Mo = L + \left(\frac{f_m - f_{m-1}}{(f_m - f_{m-1}) + (f_m - f_{m+1})} \right) h$$

where: L = lower boundary of the modal class, f_m = frequency of the modal class, f_{m-1} = frequency of the preceding class, f_{m+1} = frequency of the following class, h = class width.

—

2. Parameters of Dispersion

Definition: Measures of dispersion quantify the degree of variability or spread of data around the central value.

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

a) Variance:

- For **discrete or continuous data:**

$$S^2 = \frac{\sum f_i(y_i - \bar{y})^2}{\sum f_i} \quad \text{or equivalently} \quad S^2 = \frac{\sum f_i y_i^2}{\sum f_i} - \left(\frac{\sum f_i y_i}{\sum f_i}\right)^2$$

—

b) Standard Deviation:

$$S = \sqrt{S^2}$$

—

c) Coefficient of Variation:

$$CV = \frac{S}{\bar{y}} \times 100\%$$

This allows comparison of the degree of variability between datasets with different units or scales.

—

d) Mean Absolute Deviation:

- For **discrete or continuous data:**

$$e_m = \frac{\sum f_i|y_i - \bar{y}|}{\sum f_i}$$

—

e) Range:

- For **discrete data:**

$$w = y_{\max} - y_{\min}$$

- For **continuous data:**

$$w = \text{upper limit of last class} - \text{lower limit of first class}$$

1.2 Regression: Two-dimensional descriptive statistics or bivariate statistics

When two statistical variables are studied simultaneously, we deal with **bivariate descriptive statistics**. The purpose of this analysis is to identify and describe the type and strength of the relationship that may exist between the two variables.

1.2.1 Construction of Statistical Tables

A **statistical table** allows for the organization of data describing two variables X and Y observed for the same individuals. Each cell in the table represents the number of observations corresponding to a specific combination of X and Y values.

	y_1	y_2	$\dots$	y_q	**Total** $(n_{i.})$
x_1	n_{11}	n_{12}	$\dots$	n_{1q}	$n_{1.}$
x_2	n_{21}	n_{22}	$\dots$	n_{2q}	$n_{2.}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
x_p	n_{p1}	n_{p2}	$\dots$	n_{pq}	$n_{p.}$
Total $(n_{.j})$	$n_{.1}$	$n_{.2}$	$\dots$	$n_{.q}$	n

where:

$$n_{ij} = \text{number of individuals for which } X = x_i \text{ and } Y = y_j,$$

$$n_{i.} = \sum_{j=1}^{q} n_{ij}, \quad n_{.j} = \sum_{i=1}^{p} n_{ij}, \quad n = \sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}.$$

Example: The following table shows the distribution of 100 individuals according to their education level (X) and income class (Y):

Education Level (X)	**Low Income**	**Medium Income**	**High Income**	**Total**
Primary	20	10	2	32
Secondary	5	20	10	35
University	2	11	20	33
Total	27	41	32	100

This table provides a first descriptive overview of how income varies with education level: individuals with higher education tend to have higher income.

—

1.2.2 Graphical Representations

Graphical representation is a useful method for visualizing the relationship between two variables. It provides an intuitive view of how the data behave together and whether the relationship appears strong, weak, or nonexistent.

The type of graph depends on the nature of the variables involved:

- **Both variables are qualitative:** Bar charts or contingency plots are used to compare category combinations. *Example:* Relationship between farming system type (organic, conventional, mixed) and soil quality class (low, medium, high).
- **One variable qualitative and the other quantitative:** Bar charts or boxplots are used to compare numerical measurements across categories. *Example:* Relationship between land use type and average yield (tons/ha).
- **Both variables quantitative:** A scatter plot is used, where each observation (x, y) represents a pair of values for the same unit. *Example:* Relationship between rainfall (mm) and crop yield (tons/ha).

By interpreting these graphical summaries, we can determine:

- whether the relationship is **positive** (both variables increase together),
- **negative** (one increases while the other decreases),
- or **absent** (no visible association).

These descriptive analyses provide the basis for deeper statistical methods such as correlation and regression, which quantify and model the strength of the relationship between two variables.

1.2.3 Data Reduction

In bivariate descriptive statistics, **data reduction** refers to the process of summarizing the relationship between two quantitative variables in a compact and meaningful way. Instead of analyzing all individual observations, the aim is to extract a few key indicators that describe the joint behavior of the two variables.

Two main categories of parameters are typically distinguished:

- **Univariate parameters:** describe each variable separately (mean, variance, standard deviation);
- **Bivariate parameters:** describe the relationship between the two variables (covariance, correlation, regression).

These parameters provide a bridge between raw data and interpretation — allowing the analyst to determine whether two phenomena evolve together, in opposite directions, or independently.

—

Covariance

The **covariance** measures the joint variability between two quantitative variables. It indicates whether the variables tend to vary in the same direction or in opposite directions.

- For **discrete data:**

$$\mathrm{Cov}(X,Y) = \frac{\sum f_i (x_i - \bar{x})(y_i - \bar{y})}{\sum f_i}$$

  where:

  - x_i, y_i are the corresponding values of the two variables,
  - f_i is the frequency of each pair (x_i, y_i),
  - $\bar{x}$ and $\bar{y}$ are the means of X and Y respectively.

- For **continuous (grouped) data:**

$$\mathrm{Cov}(X,Y) = \frac{\sum f_i (x_i - \bar{x})(y_i - \bar{y})}{\sum f_i}$$

  where:

  - x_i and y_i are the midpoints (class centers) of the corresponding intervals of X and Y,
  - f_i represents the frequency of each class pair,
  - $\bar{x}$ and $\bar{y}$ are the mean values of X and Y calculated from the grouped data.

Interpretation:

- If $\mathrm{Cov}(X,Y) > 0$: when X increases, Y also tends to increase (positive association). *Example:* In agroecology, an increase in soil nitrogen content is often associated with higher crop yields.

- If $\text{Cov}(X, Y) < 0$: when X increases, Y tends to decrease (negative association). *Example:* An increase in air temperature may correspond to a decrease in soil moisture.
- If $\text{Cov}(X, Y) = 0$: there is no linear relationship between the two variables. *Example:* The number of pollinator species may not be directly related to daily rainfall in certain regions.

—

Correlation Coefficient (r)

The **correlation coefficient** r standardizes the covariance to express the *strength and direction* of the linear relationship between two quantitative variables:

$$r = \frac{\text{Cov}(X, Y)}{S_X S_Y} \quad \text{with} \quad -1 \leq r \leq 1$$

where S_X and S_Y are the standard deviations of X and Y respectively.

Interpretation of the Correlation Coefficient

- If $r > 0$: there is a **positive correlation** — as X increases, Y tends to increase as well. *Example:* Higher rainfall tends to correspond to higher plant biomass.
- If $r < 0$: there is a **negative correlation** — when X increases, Y decreases. *Example:* Increasing pesticide concentration may be associated with lower insect biodiversity.
- If $r \approx 0$: there is **no linear correlation** — no apparent relationship between the two variables. *Example:* The number of farms in a region may not be linearly related to the average wind speed.

$r \approx +1$	$r \approx -1$	$r \approx 0$
Strong positive correlation	Strong negative correlation	No linear correlation

The correlation coefficient provides a dimensionless measure (between -1 and $+1$), which makes it suitable for comparing relationships across different datasets or units of measurement. A value of r close to $+1$ or -1 indicates a strong linear relationship, while values near 0 suggest a weak or nonexistent one.

—

Additional Remarks

While correlation and covariance help describe the *existence* and *direction* of relationships, they do not imply causality. A strong correlation between two variables does not mean that one causes the other — they may both be influenced by a third factor.

Example: In an agricultural context, crop yield and pest abundance may be correlated, but both could depend primarily on weather conditions or pesticide use.

Therefore, correlation analysis should often be followed by **regression analysis**, which aims to model and predict one variable based on another while quantifying the degree of uncertainty in the prediction.

1.2.4 Regression Line

The concept of the regression line is to represent a scatter of data points by a straight line that best fits their overall distribution. This line should pass as close as possible to all points; therefore, the goal is to minimize the deviations between the observed values and the estimated values on the line. In essence, the regression line provides a simplified model that explains how the variable Y depends on the variable X (see Figure 1.1).

$$Y = aX + b \quad \text{with} \quad a = \frac{\text{Cov}(X,Y)}{\text{Var}(X)}, \quad b = \bar{Y} - a\bar{X}$$

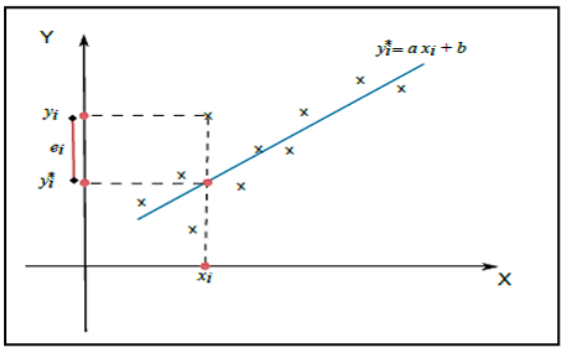


Figure 1.1: The regression line representing the best fit to the data points.

Linear Adjustment Linear adjustment consists of approximating the cloud of points by a straight line derived from the regression equation. This adjustment allows for easier visualization and interpretation of the general relationship between the two variables.

—

Remark

The correlation coefficient r is used to evaluate the validity of the linear adjustment. It measures the degree of alignment between the observed points and the regression line. We apply the following decision criteria:

- If $|r| < 0.7$, the linear relationship is weak — the adjustment is not acceptable (*line rejected*).
- If $|r| \geq 0.7$, the relationship is sufficiently strong — the adjustment is acceptable (*line accepted*).

Application Example: In an agroecological experiment, researchers studied the relationship between the weight of tomato plants (X, in kilograms) and the number of fruits produced (Y). The purpose was to determine whether heavier plants tend to produce a larger number of fruits.

Plant Weight X (kg)	14	17	24	25	27	33	34	37	40	41	42
Number of Fruits Y	38	45	60	65	70	80	85	90	95	100	102

Questions:

1. Compute the correlation coefficient r. What can you conclude?
2. Determine the regression line equation Y as a function of X.
3. Estimate the number of fruits expected for a plant weighing 50 kg.

Solution: Solution:

$$\bar{X} = \frac{1}{11}(14 + 17 + 24 + 25 + 27 + 33 + 34 + 37 + 40 + 41 + 42) = 30.36, \quad S_x = 9.66$$

$$\bar{Y} = \frac{1}{11}(38 + 45 + 60 + 65 + 70 + 80 + 85 + 90 + 95 + 100 + 102) = 75.45, \quad S_y = 20.12$$

$$\overline{XY} = \frac{1}{11}(14{\times}38{+}17{\times}45{+}24{\times}60{+}25{\times}65{+}27{\times}70{+}33{\times}80{+}34{\times}85{+}37{\times}90{+}40{\times}95{+}41{\times}100{+}42{\times}102)$$

$$\mathrm{Cov}(X, Y) = \overline{XY} - \bar{X}\bar{Y} = 2475.45 - (30.36 \times 75.45) = 188.27$$

$$r = \frac{\mathrm{Cov}(X, Y)}{S_x S_y} = \frac{188.27}{9.66 \times 20.12} = 0.97$$

Interpretation: There is a very strong positive linear relationship between plant weight and the number of fruits produced.

$$Y = aX + b$$
$$a = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)} = \frac{188.27}{(9.66)^2} = 2.01$$
$$b = \bar{Y} - a\bar{X} = 75.45 - 2.01(30.36) = 14.51$$
$$\boxed{Y = 2.01X + 14.51}$$

For $X = 50$:

$$Y = 2.01(50) + 14.51 = 115.51 \approx 116 \text{ fruits.}$$

Conclusion: The analysis shows that tomato plants with greater biomass tend to produce more fruits. For each additional kilogram of plant weight, the number of fruits increases by about **2 fruits on average**.