Elementary Statistics

**@source : gemini – 2.5 pro – 22 June 2025**

Elementary statistics is a field built on a foundation of core concepts and terminologies that allow us to collect, analyze, interpret, and present data. Understanding these terms is the first step toward statistical literacy. The main crux of these terminologies can be broken down into two major branches of statistics and their foundational concepts.

**The Two Branches of Statistics**

At the highest level, elementary statistics is divided into two main areas:

**1. Descriptive Statistics:** This branch focuses on summarizing and organizing data to describe its main features. It provides a straightforward summary of the data you have.

**2. Inferential Statistics:** This branch uses data from a sample to make generalizations, predictions, or inferences about a larger population. It allows us to draw conclusions beyond the immediate data.

**Foundational Concepts and Terminologies**

These are the absolute basics that underpin both descriptive and inferential statistics:

* **Population:** This refers to the entire group of individuals, items, or events that you want to study and draw conclusions about. For example, all university students in a country.
* **Sample:** A subset or a smaller, manageable group of individuals selected from the population. A well-chosen sample should be representative of the entire population. For example, 1,000 randomly selected university students from that country.
* **Variable:** A characteristic or attribute that can be measured and can vary among individuals in a sample or population.
  + **Qualitative (or Categorical) Variable:** Describes a quality or characteristic. It can be placed into categories, such as eye color, gender, or political affiliation.
  + **Quantitative (or Numerical) Variable:** Represents a measurable quantity.
    - **Discrete Variable:** Can only take on specific, countable values (e.g., the number of children in a family).
    - **Continuous Variable:** Can take on any value within a given range (e.g., height, weight, or temperature).
* **Parameter vs. Statistic:**
  + **Parameter:** A numerical value that describes a characteristic of an entire *population*. For example, the average height of *all* men in a country.
  + **Statistic:** A numerical value that describes a characteristic of a *sample*. For example, the average height of a *sample* of 100 men from that country. We use statistics to estimate population parameters.

**Key Terminologies in Descriptive Statistics**

When describing a dataset, we typically look at its central tendency and its spread.

**Measures of Central Tendency (The "Center" of the Data)**

* **Mean:** The arithmetic average of a set of numbers. It is calculated by summing all the values and dividing by the count of the values. It is sensitive to outliers (extreme values).
* **Median:** The middle value in a dataset that has been arranged in ascending order. If there is an even number of values, the median is the average of the two middle numbers. It is not affected by outliers.
* **Mode:** The value that appears most frequently in a dataset. A dataset can have one mode, more than one mode, or no mode.

**Measures of Dispersion (The "Spread" of the Data)**

* **Range:** The simplest measure of spread, calculated as the difference between the highest and lowest values in a dataset.
* **Variance (σ2 for a population, s2 for a sample):** The average of the squared differences from the Mean. A larger variance indicates that the data points are more spread out.
* **Standard Deviation (σ for a population, s for a sample):** The square root of the variance. It is a more intuitive measure of spread because it is in the same units as the original data. A small standard deviation indicates that the data points tend to be very close to the mean, while a large standard deviation indicates that the data are spread out over a wider range of values.

**Key Terminologies in Inferential Statistics**

Inferential statistics allows us to make educated guesses about a population based on sample data.

* **Probability:** The measure of the likelihood that an event will occur. It is the foundation for inferential statistics, with values ranging from 0 (the event will not occur) to 1 (the event is certain to occur).
* **Hypothesis Testing:** A formal procedure for investigating our ideas about the world using statistics. It is used to assess the plausibility of a hypothesis by using sample data.
  + **Null Hypothesis (H0​):** A statement of no effect or no difference, which is assumed to be true until evidence indicates otherwise.
  + **Alternative Hypothesis (Ha​ or H1​):** A statement that contradicts the null hypothesis and is what the researcher is trying to find evidence for.
  + **p-value:** The probability of observing data as extreme as, or more extreme than, what was actually observed, assuming the null hypothesis is true. A small p-value (typically ≤ 0.05) is considered strong evidence against the null hypothesis.
  + **Significance Level (α):** The predetermined threshold for how unlikely a result has to be (if the null hypothesis is true) to be considered statistically significant. It is the probability of making a Type I error (rejecting a true null hypothesis).
* **Confidence Interval:** A range of values, derived from sample statistics, that is likely to contain the value of an unknown population parameter. For example, a "95% confidence interval" for the average height of a population might be [170 cm, 175 cm], meaning we are 95% confident that the true population mean falls within this range.
* **Correlation:** A statistical measure that expresses the extent to which two variables are linearly related, meaning they change together at a constant rate. Correlation does not imply causation.
* **Regression:** A statistical method used to model the relationship between a dependent variable and one or more independent variables. It can be used to predict the value of the dependent variable based on the values of the independent variables.