
Lecture 1: Descriptive Statistics

INF1004 Mathematics II

Name: Timothy Chia

Date: 05/01/2025

1. What is Statistics?

Statistics is the art of collecting, analysing, and interpreting data. It allows us to discover hidden patterns and make informed decisions, even when we face uncertainty in the real world.

Broadly, statistics can be divided into three closely related areas:

Descriptive Statistics	Probability Theory	Inferential Statistics
This branch involves collecting, organising, summarising, and presenting the data you actually have.	This branch links descriptive statistics to inferential statistics by focusing on modelling randomness and uncertainty using mathematical frameworks.	This branch involves making predictions, testing hypotheses, and drawing conclusions about a population based on a smaller sample.

Data Sources

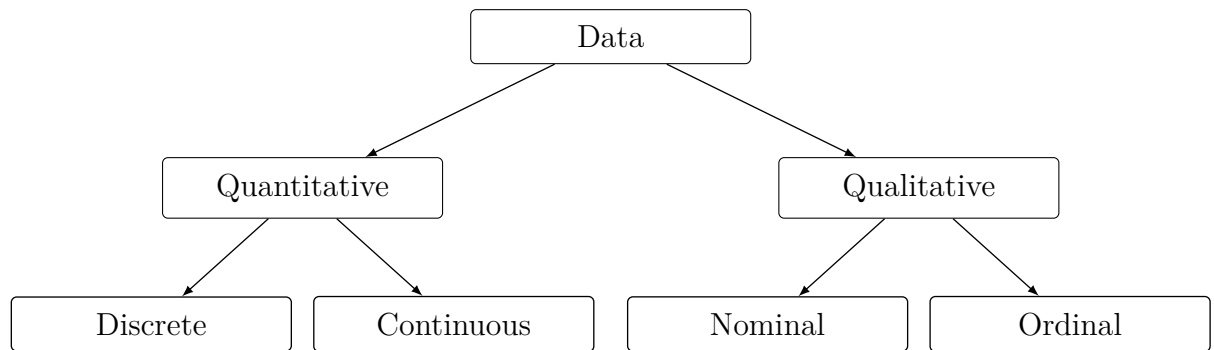
- **Population:** The entire collection of all individuals or items of interest. Measuring every single item in a population is called a *census*, which is often expensive and time-consuming.
- **Sample:** A smaller subset selected from the population. We study the sample to estimate properties of the population. The process of selecting this subset is called *sampling*.

To describe populations and samples more precisely, we can use simple mathematical notation.

Definitions

Let P be the sample space of a particular population being investigated. Let S be a sample drawn from the population such that $S \subset P$.

2. Types of Data



- **Quantitative (Numeric):** Expresses numeric data and represents amounts or quantities.
 - **Discrete:** Numeric data that are specific, countable values.
 - **Continuous:** Numeric data that can take any value within a range such that it can be measured with increasing precision.
- **Qualitative (Categorical):** Expresses non-numeric data and generally represents some quality, characteristic, or category.
 - **Nominal:** Categories that are just names or labels with no inherent order or ranking.
 - **Ordinal:** Categories that have a meaningful order or ranking, but the difference between the ranks is not necessarily equal.

3. Graphical Presentation of Data

Categorical Data:

- **Frequency Table:** A simple summary that lists each category alongside its frequency (f_i , the count) or relative frequency ($p_i = f_i/N$, the proportion).
- **Graphical:**
 - **Bar Charts:** Use separate bars (with gaps between them) to compare different categories.
 - **Pie Charts:** Show how a whole is divided into different categories (best for showing proportions).

Quantitative Data:

- **Frequency Distribution:** Because numeric data can be vast, we group it into "classes" or "intervals" to see the pattern.
- **Histogram:** The most common way to visualize numeric distributions. Unlike bar charts, the bars in a histogram *touch each other* to indicate that the data is continuous number-line data.
- **Unequal Class Widths:** Sometimes, grouping intervals have different sizes. If we simply plotted frequency, wider bars would look visually "heavier" or more important than they should. To correct this, we use **Density** for the height.

$$\text{Density} = \frac{\text{Relative Frequency}}{\text{Class Width}}$$

Note: In a density histogram, the *area* of the bar represents the frequency, not just the height.

Bivariate Data (Two Variables):

- **Crosstabulation (Contingency Table):** A table that displays the frequency distribution of two variables simultaneously (e.g., "Gender" vs. "Subject Choice").
- **Charts: Stacked or Clustered Bar Charts** allow you to compare subgroups within the data side-by-side.

4. Measures of Central Tendency

Data may be presented numerically to succinctly describe it. There are three basic measures – central tendency (or position), spread and relative standing.

Mean Definition

Let x_1, x_2, \dots, x_m be a collection of m numerical observations. The mean of these observations is defined as

$$\frac{1}{m} \sum_{i=1}^m x_i$$

If the observations constitute the entire population of size N , the mean is called the **population mean** and is denoted by μ , where

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i$$

If the observations form a sample of size n drawn from the population, the mean is called the **sample mean** and is denoted by \bar{x} , where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Here, N denotes the population size, n denotes the sample size with $n < N$, and the distinction between μ and \bar{x} reflects whether the mean is taken over the full population or over a sample.

Median Definition

Let x_1, x_2, \dots, x_m be a collection of m numerical observations arranged in non-decreasing order. The **median** is defined as the central value of the ordered data.

- If m is odd, the median is the value in position $\frac{m+1}{2}$.
- If m is even, the median is defined as the arithmetic mean of the values in positions $\frac{m}{2}$ and $\frac{m}{2}+1$.

Mode Definition

The **mode** is defined as the value(s) that occur with the greatest frequency in the data set. A data set may have no mode if all values occur with equal frequency, one mode if a single value occurs most frequently, or multiple modes if several values share the highest frequency.

Histogram Skewness

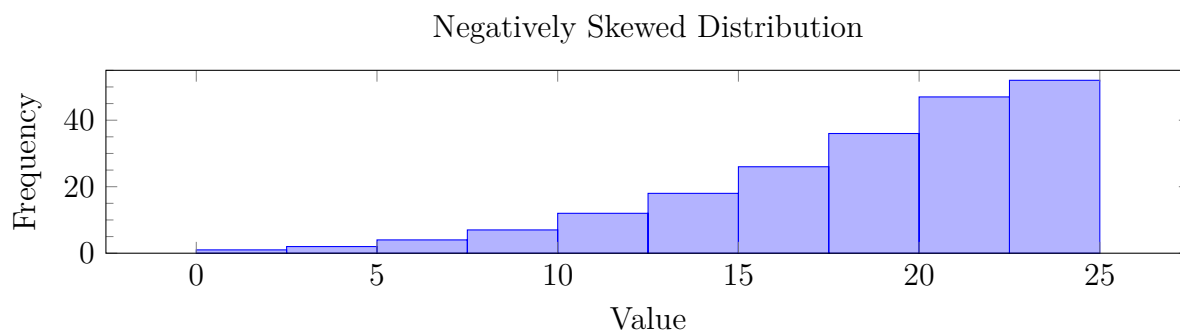


Figure 1: A negatively (left) skewed distribution with a long tail towards smaller values;
 $mean < median < mode$.

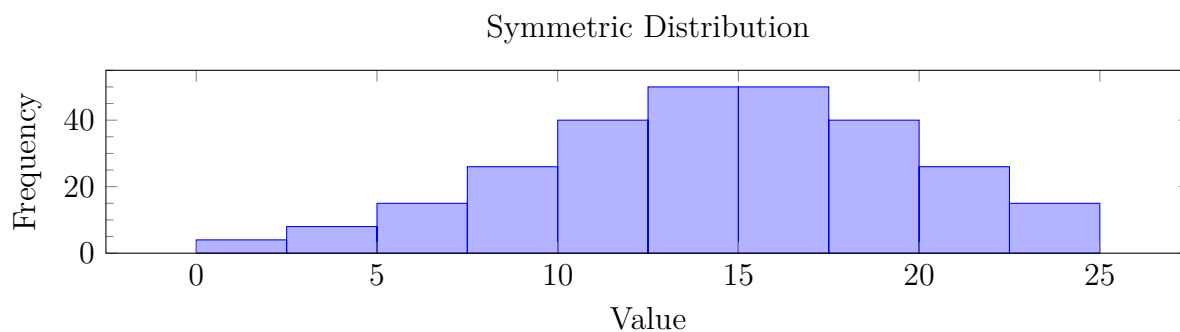


Figure 2: An approximately symmetric distribution with no pronounced tail;
 $mean = median = mode$.

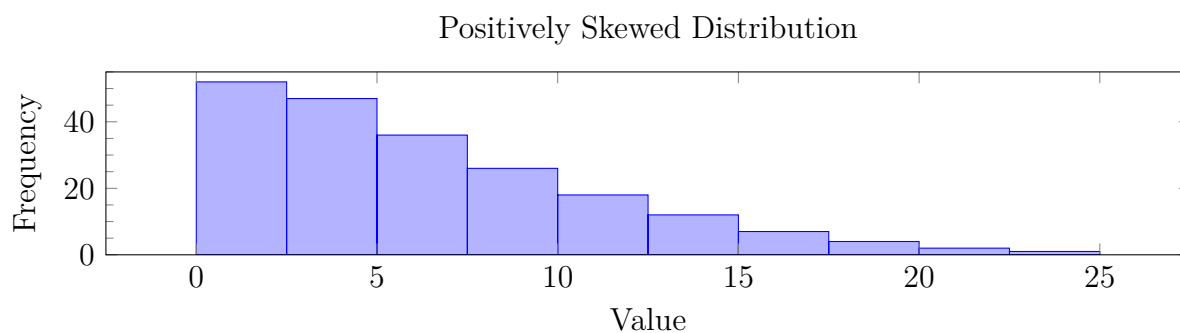


Figure 3: A positively (right) skewed distribution with a long tail towards larger values;
 $mean > median > mode$.

5. Measures of Relative Standing

Measures of relative standing describe the position of an observation relative to the other values in a dataset, rather than its absolute size. Common measures include quartiles and percentiles.

Quartiles: Quartiles divide an ordered dataset into four equal parts (quarters). Each part contains 25% of the data.

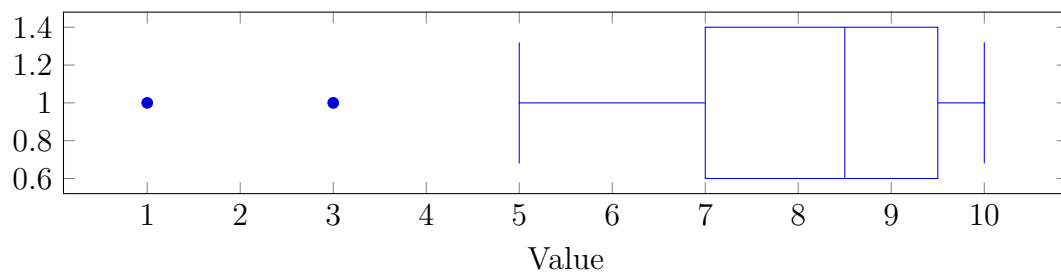
- Q_1 (Lower Quartile): the value below which 25% of the data lie
- Q_2 (Median): the value below which 50% of the data lie
- Q_3 (Upper Quartile): the value below which 75% of the data lie

Percentiles: Percentiles generalise the idea of quartiles. The p th percentile is the value below which $p\%$ of the data lie. For example, the 90th percentile is the value below which 90% of observations fall.

Box-and-Whisker Plot: Quartiles and the overall distribution of the data are commonly visualised using a box-and-whisker plot.

- The box spans from Q_1 to Q_3 and represents the interquartile range (IQR)
- The line inside the box marks the median (Q_2)
- The whiskers extend to the smallest and largest non-outlier values
- Points beyond the whiskers (if shown) represent outliers

Example Box-and-Whisker Plot



From the boxplot, we can quickly identify the median, the spread of the middle 50% of the data, and whether the distribution appears symmetric or skewed.

6. Measures of Spread

Measures of spread (also called measures of dispersion) describe how variable or spread out a dataset is around its centre.

Range: The simplest measure of spread. It only considers the extreme values in the dataset and is therefore highly susceptible to outliers. It also ignores all values between the minimum and maximum.

$$\text{Range} = \text{Highest Value} - \text{Lowest Value}$$

Interquartile Range (IQR): Measures the spread of the middle 50% of the data after the data have been ordered. Because it ignores the extremes, it is a robust measure of spread.

$$\text{IQR} = Q_3 - Q_1$$

- **Outliers:** The IQR is commonly used to identify unusual values. A data point is considered an outlier if it lies below $Q_1 - 1.5(\text{IQR})$ or above $Q_3 + 1.5(\text{IQR})$. This rule is a convention rather than a strict mathematical law.

Variance (s^2): Ideally, we want to measure the average distance of data points from the mean. Variance does this by calculating the average *squared* distance from the mean, where squaring ensures all values are positive. A larger variance indicates greater dispersion around the mean.

- **Sample Variance Formula:** We divide by $n - 1$ instead of n to obtain a better estimate of the population variance. This adjustment is known as Bessel's correction.

$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$$

- **Alternative Formula (Computational):** A mathematically equivalent formula often used in manual calculations to reduce rounding errors.

$$s^2 = \frac{1}{n - 1} \left[\sum x_i^2 - \frac{(\sum x_i)^2}{n} \right]$$

Standard Deviation (s): Because variance is measured in squared units (for example, dollars squared), it can be difficult to interpret. Taking the square root returns the measure to the original units of the data.

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

Coefficient of Variation (CV): Standard deviation is an absolute measure of spread. The coefficient of variation is a *relative* measure, expressed as a percentage. It is particularly useful for comparing variability or risk between datasets with very different means. However, it should not be used when the mean is close to zero.

$$CV = \frac{s}{\bar{x}} \times 100\%$$