There are **2 types of statistics**: Descriptive & Inferential

**Statistics**

* **Descriptive** - Consists of organizing & summarizing the data

You don’t make predictions, you just show what the data says.

**Examples**:

* + - Average age of people in a class
    - Highest and lowest test scores
    - Graphs, charts, tables
    - Measures like mean, median, mode, standard deviation

👉 Think of it as “telling the story of the data you see.”

* + **Measure of central tendency**
* **Mean** = (Sum of all values) / (Number of values)

A black background with white text

AI-generated content may be incorrect.

* + - * + ∑xi​ = Sum of all values
        + n = Number of values
* The **median** is the middle value when the data is arranged in order.

**A black background with white text

AI-generated content may be incorrect.**

* The **mode** is the value(s) that appear **most frequently** in the data.

There's no fixed formula; just identify the number(s) with the highest frequency.

**Note**: If the data have outliers, we can choose median or mode else, mean.

* + **Measure of dispersion**
    - **Variance** tells us **how spread out** the data is from the mean.

**A black background with white text

AI-generated content may be incorrect.**

**Note:** For sample data, we divide by n-1 instead of n because using n tends to underestimate the true population variance. n-1 is known as Bessel's correction, and it compensates for this underestimation. Dividing by n-1 corrects this bias, giving a better (unbiased) estimate of the population variance.

* + - **Standard deviation** is the **square root of variance**. It gives dispersion in the same units as the data. Meaning, how far a datapoint is away from the mean.

**A black screen with white text

AI-generated content may be incorrect.**

* **Inferential**

After collecting sample data, it’s about **making predictions or conclusions** about a **larger group** based on a **small sample** of data.

You use the data to make **guesses, test hypotheses, or estimate future outcomes**.

using some experiments (z test, t test etc), These conclusions are for other data called **population** data.

**Examples**:

* Predicting election results based on a survey
* Testing if a new drug works better than the old one
* Estimating the average income of a country using a sample

👉 Think of it as “making educated guesses based on data.”

**Variable**

Variable is a property that can take up any value. 2 different types of variable **Quantitative** & **Qualitative** / **Categorical** Variables

* **Quantitative Variables**
  + **Discrete Quantitative Variable**
    - A variable that takes **specific whole numbers** (counts).
      * A have 5 houses (can’t say 5.5 houses)
      * B have 2 children (not 2.5 children)
  + **Continuous Quantitative Variable**
    - A variable that can take **any value within a range,** including decimals.
      * Height = 172.5 cm
      * Weight = 65.8 kg
      * Temperature = 36.6°C
* **Qualitative / Categorical Variables**

Variables that describe categories or qualities. These are **not numeric** and can be divided into two type **Nominal** & **Ordinal.**

* + **Nominal:**

Categories with no specific order.

* + - Gender (Male, Female, Non-binary)
    - Eye colour (Blue, Brown, Green)
  + **Ordinal:**

Categories with a **meaningful order** but without a specific numeric difference between them.

* Education level (High school, Bachelor's, Master's, PhD)
* Rating (Poor, Fair, Good, Excellent)

**Random Variables**

A **random variable** is a function whose values are derived from the outcomes of different processes or experiments or trials.

* **Discrete Random Variable**

Takes **specific, distinct values** (typically whole numbers).

* + - The number of heads in 10-coin tosses (can be 0, 1, 2, ..., 10).
    - The number of cars that pass a checkpoint in an hour (e.g., 5, 10, 20).
* **Continuous Random Variable**

Takes **any value** within a range, including decimals.

* + - The height of a person (could be any value, such as 170.5 cm, 172.3 cm).
    - The time it takes to run a race (e.g., 9.12 seconds, 9.13 seconds).

**Percentiles & Quartiles**

A **percentile** is a value below which a certain percentage of observations lie.

**Percentile of value x** = (no of values below x / n) \* 100

If a percentile is given, then we can fine the value:

**Value** = (Percentile / 100) \* (n+1)

* n = total number of values
* If the position is a fractional value (e.g., 3.5 or 3.75), we identify the values at the positions surrounding it (e.g., the 3rd and 4th values). For 3.5, we take the **average** of the 3rd and 4th values, since it lies exactly halfway between them. For other fractions like 3.75, we apply **linear interpolation** between the 3rd and 4th values, weighted by the decimal portion (in this case, 0.75), to estimate the value at that position.

**Linear Interpolation Formula Example:**

Consider the sorted dataset with 6 values:

**[2,2,4,6,8,10]**

We need to fine the 25th percentile value

**Value** = (25/100) \* (6+1)

= 0.25 \* 7 = 1.75TH position

The position 1.75 means the 25th percentile is between the 1st and 2nd data points.

We are looking for the **1.75th value**.

* 1st value = 2
* 2nd value = 2

We are 75% of the way between 1st and 2nd, so:

Interpolated Value= Interpolated Value=2+0.75 × (2−2) = 2+0 = 2

**✅ Final Answer:**

The value at the **1.75th position is 2**, which says, **the 25th percentile** of the values of this distribution **is below 2**

**Quartiles** divide a sorted data set into **four equal parts**, with each part containing **25% of the data**. Quartiles help describe the **spread** and **distribution** of your data.

Let's say we have this sorted list:  
[2, 4, 6, 8, 10, 12, 14, 16]

* **Q1** = value at the 25th percentile = 5
* **Q2 (Median)** = value at the 50th percentile = 9
* **Q3** = value at the 75th percentile = 13

(Exact values depend on the method used for interpolation, but you get the idea.)

**5 Number Summary**

**E.g. [1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9,27]**

We can see that **27** seems much higher than the rest, it may be an **outlier**. We can use **quartiles** to detect and remove outliers.

**The formula is:**

*(Note: The data is already sorted)*

Q1 (25th percentile) = value at the 5th position = **3**

Q3 (75th percentile) = value at the 15th position = **7**

IQR = **Q3 – Q1** = 7-3 = 4

**Lower fence** = Q1 – 1.5(IQR) = 3 – 1.5(4) = -3

**Upper fence** = Q3 + 1.5(IQR) = 7 + 1.5(4) = 13

* **Q1**: value at the 25th percentile
* **Q3**: value at the 75th percentile
* **IQR**: Interquartile range, difference between Q3 & Q1

Any value **below –3** or **above 13** is considered an outlier. In this case, **27** is above 13, so it's an outlier.

After removing Outlier, the data become:

**[1,2,2,2,3,3,4,5,5,5,6,6,6,6,7,8,8,9]**

So, the **Five number summary** is:

**Minimum: 1**

**1st Quartile: 3**

**Median: 5**

**3rd Quartile: 6**

**Maximum: 9**

**Covariance & Correlation**

**Covariance & Correlation** are 2 statistical measures used to determine the **relationship** **between 2 variables**. Both are used to understand how changes in one variable are associated with changes in another variable.

**Covariance**

Covariance is a measure of how much 2 random variables change together. If the variables tend to increase and decrease together, the covariance is **positive**. If one increases the other decreases, the covariance is **negative**. If there's **no clear pattern**, covariance is **close to zero**

A black background with white numbers and symbols

AI-generated content may be incorrect.

* xi, yi​ are values of the two variables
* xˉ, yˉ​ are the means of X and Y
* n is the number of data points

The **covariance** of a variable with **itself** is just its **variance**.

A black background with white text

AI-generated content may be incorrect.

**Advantages**

* Quantify the relationship between x & y

**Disadvantages**

* Covariance does not have a specific limit value. (-inf to inf)

**Correlation**

There are2 types of correlations

* Pearson Correlation coefficient
* Spearman Rank Correlation

**Pearson Correlation coefficient**

The **Pearson Correlation Coefficient** (often written as r) measures the strength and direction of the linear relationship between two variables.

* + It ranges from **–1 to +1**
  + +1 = perfect positive linear relationship
  + –1 = perfect negative linear relationship
  + 0 = no linear relationship

**A black background with white text

AI-generated content may be incorrect.**

* Cov (X, Y) is the covariance between X and Y
* σX ​ is the standard deviation of X
* σY ​ is the standard deviation of Y

**Spearman Rank Correlation**

Pearson correlation may not accurately capture the relationship in **nonlinear** data. For example, if the data points **increase at a slower rate** (i.e., follow a nonlinear trend), Pearson may still return a high coefficient (like **0.88**), but it won't truly represent the nature of the relationship. In such cases, **Spearman rank correlation** is better because it can detect **monotonic relationships**, whether linear or not.

A black background with white text

AI-generated content may be incorrect.

* di = difference between ranks of xi and yi
* n = number of observations

**OR**

A black background with white text

AI-generated content may be incorrect.

This looks just like the **Pearson formula**, but instead of using the original X and Y values, you use their **ranks**.

**Why Correlation is Important?**

* **Finds Relationships Between Variables**  
  → Helps us see if two things move together (like study time and exam marks).
* **Helps in Decision Making**  
  → Businesses can check if more ads lead to more sales, or if price drops attract more customers.
* **Useful in Building Models**  
  → In machine learning, we remove features that are too similar to avoid confusion in predictions.
* **Detects Patterns in Data**  
  → Even if it's not obvious, correlation can show trends (like if higher temperature increases electricity usage).
* **Saves Time and Effort**  
  → Helps us focus only on useful data by identifying which variables are really connected.
* **Used Across Fields**  
  → From finance and health to marketing and sports — it's used everywhere to understand data better.

**Probability**

**Probability** is the chance of something happening.

* **Mutual Exclusive Event**

Two events are **mutually exclusive** if they **cannot happen at the same time**.

**Example:**

* Tossing a coin: Getting **heads** and **tails** at the same time is not possible.
* Rolling a die: Getting a **3** and a **5** in one roll can't happen.

So, if one happens, the other won’t.

* **Additive Rule of Probability**

This rule helps us find the chance of **either** of two events happening.

There are two versions:

* 1. **For Mutually Exclusive Events:**

If **A** and **B** can't happen together, then:

**P (A or B) = P(A) + P(B)**

**Example:**

* Rolling a die: What's the probability of getting a **2** or a **5**?

P (2) = 1/6, P (5) = 1/6

Since they can’t happen at the same time:

P (2 or 5) = 1/6 + 1/6 = 2/6 =1/3

* 1. **For Non-Mutually Exclusive Events:**

If **A** and **B** *can happen together*, then:

P (A or B) = P(A) + P(B) – P (A and B)

Note: We subtract the overlap because it gets counted twice.

**Example:**

* Drawing a card: What's the chance of drawing a **heart** or a **king**?
  + Total cards = 52
  + Hearts = 13
  + Kings = 4
  + One king is also a heart (King of Hearts)

P(Heart) = 13/52, P(King) = 4/52, P (Heart and King) =1/52

​P (Heart or King) = 13/52 + 4/52 − 1/52 = 17/52 - 1/52 = 16/52 = 4/13

* **Complementary Events:**If A is an event, then not A (its complement) is the **opposite**.

P (Not A) =1 − P(A)

* **Independent Events:**Two events are independent if **one happening doesn’t affect the other**.  
  Example: Tossing a coin and rolling a die.
* **Dependent Events:**

When one event **does affect the outcome of another**.  
Example: Drawing cards without replacement.

* **Multiplication Rule**

It helps you find the probability of two events happening together (i.e., both A and B happening).

1. **If the events are independent:**

P (A and B) = P(A) × P(B)

**Example**:

* + - Tossing a coin and rolling a die:

P (Heads and 4) = P(Heads) × P (4) = 1/2 × 1/6 = 1/12

1. **If the events are dependent:**

P (A and B) = P(A) × P(B/A)

**P(B|A)** means the probability of **B happening after A has already happened**.

**Example:**

* Drawing two cards from a deck **without replacement**:
  + Probability of getting an ace first = 4/52
  + After drawing one ace, 3 aces are left out of 51 cards:

P (Second ace ∣ First ace) = 3/ 51

P (Both aces) = 4/52 × 3/51 = 12/2652 = 1/221