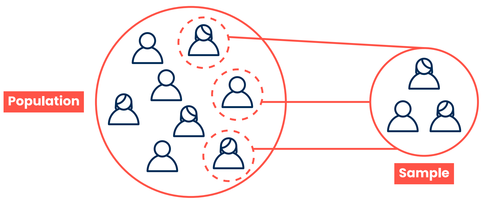
Reference card

Samples and distributions

Samples and populations

In statistics, a population is a **collection of all items** that we are interested in.



A **sample** is any **collection of items from the population**.

The **size of a sample** influences the results:

Large sample (preferred)

- Produces more accurate estimates.
- Increases statistical power.
- Requires more resources.

Smaller sample

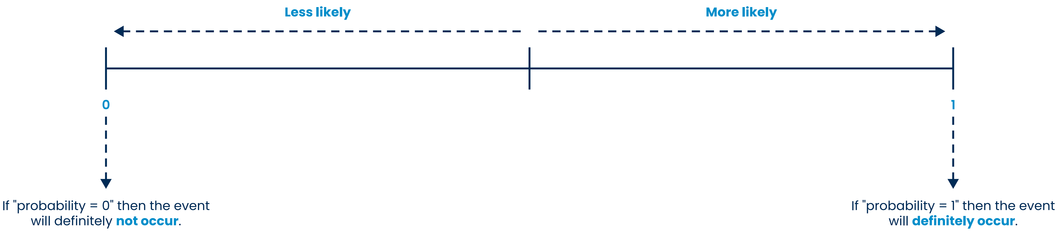
- Cost-effective, time efficient, and manageable.
- Decreases accuracy.

We use **inferential statistics** to help us draw conclusions about a population based on information from a sample. In other words, we select a random sample from the population and use this sample data to make inferences about the entire population.

Probability theory provides the foundation for inferential statistics, as it allows us to quantify the likelihood or chance of various outcomes occurring.

Probability

Probability involves quantifying the **likelihood of an event occurring**, and typically ranges from 0 (representing unlikely events) to 1 (representing very likely events).



If "probability = 0" then the event will definitely **not occur**.

If "probability = 1" then the event will **definitely occur**.

Random variables

A random variable is a variable whose value is **determined by chance or probability**, i.e. it is a numerical value that varies at random and is based on the outcome of a random experiment.

Random variables can either be **continuous** or **discrete**:

Continuous random variable:

Continuous data include complex numbers and varying data values **measured over a particular interval**.

Examples: weight, height

Discrete random variable:

A numerical type of data that includes whole, concrete numbers with specific and fixed data values determined by **counting**.

Examples: number of cars sold, number of voters

How is probability calculated?

There are **three** main ways to calculate probability:

Subjective

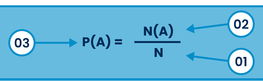
The outcome is derived from **individual perspectives** based on experience and inherent knowledge. This is the easiest approach to implement but not the most reliable.

Example:

There is a **75%** chance the stock market will go up next quarter.

Empirical

A method for calculating probabilities by **counting the number of times** an **event occurs** in a large number of trials. This is the most commonly used approach.



01. Count the number of times the event we are interested in occurs, and call this number **N(A)**.

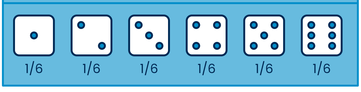
02. Perform an experiment a large number of times, andcall this number **N**.

03. The probability that event A occurs is equal to **N(A)** divided by **N**.

Axiomatic

The outcome is based on the assumption that certain events are **equally** likely to occur.

When rolling a fair-sided dice, the **probability** of getting any particular number is 1/6.



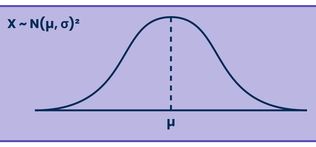
Probability distributions

Probability distributions are mathematical functions that **describe the likelihood of obtaining all possible values** for a random variable.

A probability distribution is usually denoted as f(x), where x is the value of the random variable X.

The normal distribution

In many real-life data situations, the histogram shape is **bell-curved** and **symmetric** (e.g. weight, student scores, height). The normal distribution (Gaussian distribution) is a well-known probability density function that is often used to model the data in such cases.



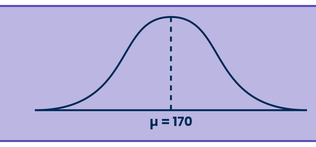
Properties:

Average is μ (the centre of the curve)

Standard deviation is σ (spread around the centre)

Example:

The height measurements of college students follow an approximate normal distribution with a **mean of 170 centimetres** and a **standard deviation of 9 centimetres**.



In probability notation, the statement is written as **X ~ N(170, 81)**

Probability mass function (PMF)

For discrete random variables, the probability distribution is referred to as the **probability mass function**.

If X is a discrete random variable, then **f(x) = P(X = x)**. We can write the PMF in the form of an equation or it can be represented in the form of a table.

The table below provides the probability of each outcome 1,2,3 or 4:

x	1	2	3	4
f(x) = P(X=x)	1/10	2/10	3/10	4/10

a) Find P(X = 2) **Answer**: 2/10

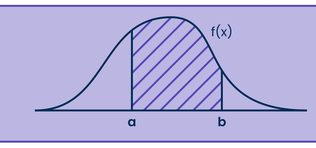
b) Find P(X < 2) **Answer**: 1/10

c) Find P(X > 2) **Answer**: 3/10 + 4/10

Probability density function (PDF)

For continuous random variables, the probability distribution is referred to as the **probability density function**.

The probability density function (PDF) of **X** is defined as f(x) where P(a < X < b) is the area under f(x) over the interval from a to b.



Using the PDF to find probabilities over an interval like P(a < X < b) requires calculus, specifically integration techniques.

Note: If X is a continuous random variable, then P(X = k) = 0