Statistics – collection of methods for planning experiments, obtaining data, and then organizing, summarizing, presenting, analysing, interpreting and drawing conclusions.

The population includes all objects of interest whereas the sample is only a portion of the population. Parameters are associated with populations and statistics with samples. Parameters are usually denoted using Greek letters (sigma σ) while statistics are usually denoted using Roman letters (x, s).
There are several reasons why we don't work with populations. They are usually large, and it is often impossible to get data for every object we're studying. Sampling does not usually occur without cost, and the more items surveyed, the larger the cost.
We compute statistics, and use them to estimate parameters.

Population – all subjects possessing a common characteristic that is being studied.
Sample – a subgroup or subset of the population.
Parameter – characteristic or measure obtained from a population.
Variable – characteristic or attribute that can assume different values.
Random variable – a variable whose values are determined by chance.
Qualitative variable – variables which assume non-numerical values, e.g. eyes colour, hair colour, etc.
Quantitative variable – variables which assume numerical values, e.g. age, height, etc.
Discrete variables – variables which assume a finite or countable number of possible values, usually obtained by counting. I.e. there are a finite or countable number of choices available with discrete data, e.g. data on shoes sizes. You can't have 2.63 people in the room.
Continuous variables – variables which assume an infinite number of possible values, usually obtained by measurement, e.g. length, weight, and time. Since continuous variables are real numbers, we usually round them. This implies a boundary depending on the number of decimal places. For example, 64 is really anything $63.5 \leq x \leq 64.5$. Boundaries always have one more decimal place than the data and end in a 5.

SAMPLE
Random sampling – sampling in which the data is collected using chance methods or random numbers. It's analogous to putting everyone's name into a hat and drawing out several names. Each element in the population has an equal chance of occurring.

Systematic sampling – sampling in which data is obtained by selecting every k^{th} object. I.e. the list of elements is 'counted off'. This is similar to lining everyone up and numbering off "1, 2, 3, 4; 1, 2, 3, 4; etc." When done numbering, all people numbered 4 would be used.
Stratified sampling – sampling in which the population is divided into groups (called strata) according to some characteristic. Each of these strata is then sampled using one of the other sampling techniques.

DATA may be described as discrete or continuous (see 'variables'). We will need to find the **average** and the **range**.
Average could mean one of four things. The arithmetic mean, the median, mid-range or mode.
Mean is what people usually intend when they say 'average'.

- ⇨ population mean: $\mu = \frac{\sum x}{N}$
- ⇨ weighted mean: $\bar{x} = \frac{\sum xf}{\sum f}$, when frequency of data is taken into account

For data in categories like colour, you can only find one *average*:

- The **mode**, i.e. the category with the highest frequency, or in other words, the most common value.

For numerical data, you can find several averages and the range:

- The **mode** x^ is the most frequent data value. There may be no mode, if no one value appears more than any other. There may also be two modes (bimodal), three modes (trimodal), or more than three modes (multi-modal).
- The **range** is the difference between the highest and lowest values, i.e. Max – Min
- The **mid-range** is the mean (mid-point) of the highest and lowest values, i.e. $\frac{Max - Min}{2}$
- The **median** $\tilde{x}$ is the middle value when the data is in ascending order (for an even number of values, take the median as halfway between the middle pair of values). There are as many numbers below the median as above the median.
- The **mean** is the sum (total) of all the values divided by the number of values
- The **weighted mean** is the mean when each value is multiplied by its weight and summed. This sum is then divided by the total of the weights.

DATA

Raw data are data collected in original form.

Frequency – the number o times a certain value or class of values occurs.

Frequency distribution – the organisation of raw data in table form with classes and frequencies (a set of data presented in a frequency table).

The **standard deviation** is the most widely used measure of dispersion. It is the square root of the **variance** – the average of the squares of the distances from the population mean. It is the sum of the squares of the deviations from the mean divided by the population size. The units on the variance are the units of the populations squared.

$$\sigma^2 = \frac{\sum (x - \bar{x})2}{N}$$

To calculate the standard deviation:

- ⇨ Calculate the mean of the values
- ⇨ Calculate the squared deviation for each value, i.e. the square of the difference between the value and the mean
- ⇨ Calculate the square root of the mean squared deviation

I.e. the standard deviation σ can be written as a formula:

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})2}{n}}$$, where σ is the Greek letter 'sigma'

$\sum (x - \bar{x})^2$ is the total of the squared deviations and

n is the number of values

Standard deviation of a frequency distribution

- ⇨ Calculate the mean of all the values
- ⇨ Calculate the squared deviation for each different value, i.e. the square of the difference between the value and the mean
- ⇨ Multiply each squared deviation by its frequency
- ⇨ Calculate the square root of the mean squared deviation

I.e. the standard deviation σ can be written as a formula:

$$\sigma = \sqrt{\frac{\sum f(x - \bar{x})2}{\sum f}}$$, where $\sum f(x - \bar{x})^2$ is the total of the squared deviations and

$\sum f$ is the total frequency

Coefficient of variation is the standard deviation divided by the mean, expressed as a percentage.