Statistics

Describing the **distribution**
- Normal Distribution, symmetrical, bell curve
- Skewed left or negatively skewed (your left foot)

Percentage Error $= \frac{\text{Error}}{\text{Actual Value}} \times \frac{100}{1}$

Central Tendencies – are different ways of working out the average
- **Mean** $(\bar{x}) = \frac{\text{sum of all the values}}{\text{number of values}}$

Use when data is numerical and there is NO extreme values.
- **Mode** = the most common value

Use when data is categorical. An example would be hair colour.
- **Median** = the middle value when they are arranged in order (ranking them)

Use when data is numerical and there are extreme values.

Measures of Spread – measures the spread of the data
- The **Range** of a set of data is the difference between the highest and lowest amounts. **Range** = Maximum value – Minimum Value.
- The **Interquartile** range $= Q_3 - Q_1$. Quartiles divide the data into 4 parts with the interquartile range the difference between values of the 1st and 3rd quarters.
- The **Standard Deviation** (σ) measures the average spread from the mean of all the values (see calculator notes for standard deviation).
- **Outliers** are extreme values that are not typical of other values in a data set.

Percentiles
Being in the 90th percentile means 90% of the people scored lower than you.

To find the kth percentile, first rank the data.

Find c, which is k% of the total numbers in the set.
$c = \frac{n \times k}{100}$
n= sample size, k = required percentile

If c is whole number find the average of c and $c + 1$
If c is not a whole number round up. Find this value in the data set.

z-score
is the number of standard deviations a given value is away from the mean

$$z = \frac{x - \mu}{\sigma}$$

x is a population value
μ is the population mean
σ is the population standard deviation

Unusual values for z-scores are $z < -2$ and $z > 2$

The **Empirical** rule states that in any Normal distribution:
- 68% of the population lie within one standard deviation of the mean $[\bar{x} - \sigma, \bar{x} + \sigma]$
- 95% of the population lie within one standard deviation of the mean $[\bar{x} - 2\sigma, \bar{x} + 2\sigma]$
- 99.7% of the population lie within one standard deviation of the mean $[\bar{x} - 3\sigma, \bar{x} + 3\sigma]$

Margin of Error and Hypothesis Testing
Is the maximum likely difference (to a 95% confidence) between the sample proportion, $\hat{p}$ and the population proportion, p

$$E = \frac{1}{\sqrt{n}} \qquad \hat{p} - \frac{1}{\sqrt{n}} < p < \hat{p} + \frac{1}{\sqrt{n}}$$

Where n is the sample size.

Hypothesis test is a procedure to test a claim about a population.
State **Null Hypothesis** H_0 (statement which defines population)
Calculate $\hat{p}$, the sample proportion.
Set up a **confidence interval** for p, the population proportion
If proportion is within confidence interval, accept H_0
If it is outside confidence interval reject it.

Mid-Interval Values – In some tables you will have to use mid-interval values to perform the statistics. Example would be finding the **mean** of the table below:

Amount (€)	0-20	20-40	40-60	60-100
Frequency	20	5	25	60

$$\frac{(10 \times 20) + (30 \times 5) + (50 \times 25) + (80 \times 60)}{20 + 5 + 25 + 60} = €58.18$$

Misuse of Statistics

Statistics can sometimes be **misleading**

- errors or omissions (always check)
- sample size (make sure the sample is large enough to be representative)
- misleading comparisons (must compare like with like)
- sources (always check sources of information)
- misleading graphs (some charts can exaggerate differences)
- response bias (occurs when people choose to take part in a survey. An example would be people ringing in to vote for favourite singer)

Commentating on Graphs

- For a summary of the various ways to **graph** statistics see the attached handout. Remember that when graphing tables, the top of the table goes on the bottom of the graph.
- Quote the **range** of the data.
- Can we estimate the **mean**? Is the **mode** obvious?
- Comment on **standard deviation** of the data. Large standard deviation means the data is well spread. Low standard deviation gives more of a cluster. Are there any outliers.
- We use a **scatter** graph when we have data that can be paired together (**bivariate** data). An example would be heights and weights or age and salary. We measure how well they are related through **Correlation**. A **line of best fit** is one that comes as close as possible to the points. We can find the equation of this line by selecting two points on the graph and using co-ordinate geometry.
- A **stem and leaf** diagram represents data by separating each value into two parts. The stem and the leaf (the final digit). View on their side to describe the distribution.

Definitions

Population – is the entire group being studied
Census – is a survey of the whole population
Sample – is a group that is selected from the population
Parameter – is a numerical measurement describing some characteristic of a population.
Statistic – is a numerical measurement describing some characteristic of a sample.

Simple random sample – selected a sample of size n, in such a way that every sample of size n has an equal chance of being selected.
Stratified random sample – first divide population into subgroups so that individuals within each subgroup share characteristics. Then a sample random sample is drawn from each group. Eg. We might first divide population by gender.
Systematic random sample – We select the sample based on random starting point and select a fixed periodic interval. Eg Select every 5^{th} entry.
Cluster sample – population is divided by sections or clusters. Then some of those clusters are randomly selected and all members from those clusters are chose. Eg. We want a sample of students. We get a list of schools and then select a school and use those students.
Quota sample – Non probability sampling. We select to fill a quota of a certain type of subgroup. Eg Selecting men between age 30 and 40.