AS Statistics – Chapter 3 – Representations of Data

Outliers

An outlier is an extreme value that lies outside the overall pattern of the data.

There are various definitions used for outliers, depending on the nature of the data available and the calculations that you are asked to carry out. There are generally based on measures of location (averages, quartiles) and spread (IQR, standard deviation).

If you know the median and quartiles, a common definition for an outlier is any value that is a particular multiple of the interquartile range either below the lower quartile or above the upper quartile:

- greater than $Q_1 - k(Q_3 - Q_1)$
- less than $Q_3 + k(Q_3 - Q_1)$ where $Q_3 - Q_1$ is the interquartile range, IQR

Most commonly, the value of the multiplier k is 1.5, but you may be given a different value to work with.

If you know the mean μ and standard deviation σ, a common definition for an outlier is any value that is more than two standard deviations from the mean:

- greater than $\mu + 2\sigma$
- less than $\mu - 2\sigma$

Anomalies and Cleaning the Data

Sometimes outliers are legitimate values which could still be correct. For example, there may be a 70-year-old member of a tennis club whose other members are generally in their twenties and thirties.

However, there are occasions when an outlier should be removed from the data since it is clearly an error and it would be misleading to keep it. For example, if the list of ages of the members of the aforementioned tennis club includes a value of 138, this is clearly not correct. These data values are known as **anomalies**.

The process of removing anomalies from a data set is known as cleaning the data.

Be careful not to remove data values just because they are outliers. You must justify why a value is being removed.

Comparing Data

When comparing data, you should always comment on a measure of location and a measure of spread.

- Mean and standard deviation
- Median and interquartile range

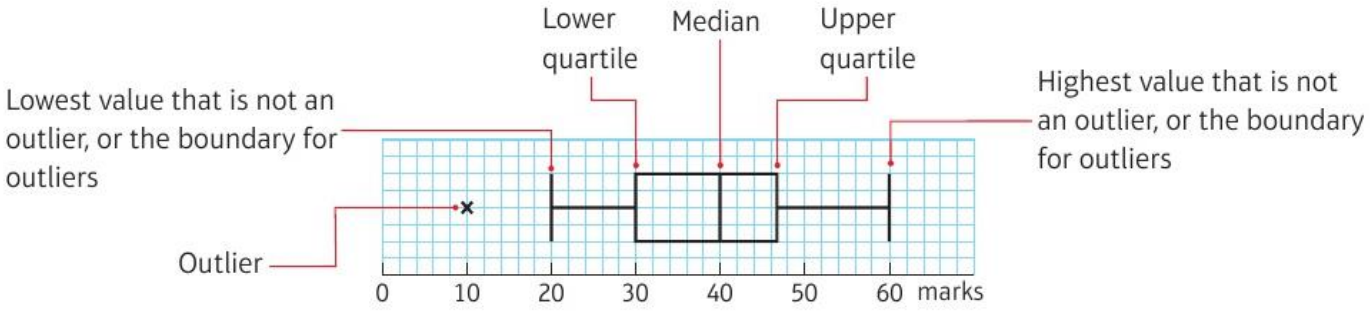
Generally the question and the data you are given will determine which pair of statistics to compare.

If the data set contains extreme values this can affect the mean and standard deviation, so the median and interquartile range are more appropriate statistics to use.

Representations of Data

You need to be familiar with box plots, cumulative frequency diagrams and histograms, all of which were covered at GCSE. The key features of each are summarised here.

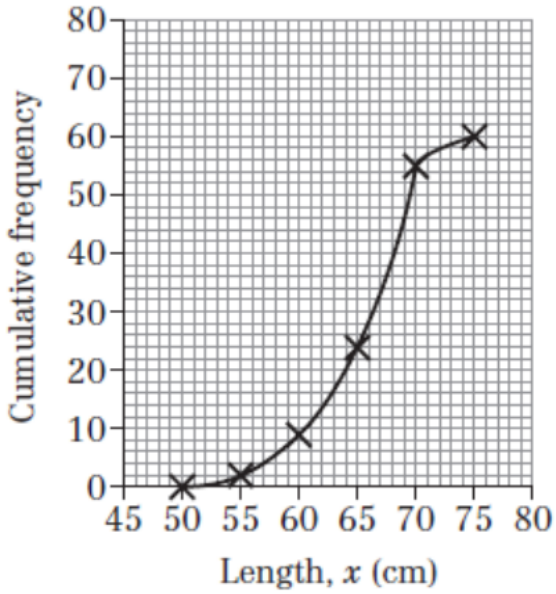
Box Plots



Cumulative Frequency Diagrams

Length, x (cm)	Frequency
$50 \leqslant x < 55$	2
$55 \leqslant x < 60$	7
$60 \leqslant x < 65$	15
$65 \leqslant x < 70$	31
$70 \leqslant x < 75$	5

Points are plotted at the **upper bound** of each group and are joined with a smooth curve.



Histograms

The **area** of each bar is equal or proportional to the frequency it represents.

- **The vertical scale on a histogram shows the frequency density:**
  frequency density = $\frac{\text{frequency}}{\text{class width}}$
- **Joining the middle of the top of each bar in a histogram with equal class widths forms a frequency polygon.**