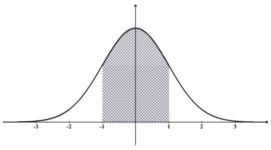
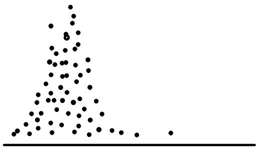
Distribution of the data

This generally refers to the shape of the data when it is represented graphically.

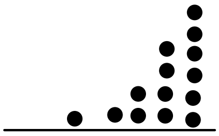
The most common type of distribution is a normal distribution. This is usually represented by a bell curve. It is based on numerical data that is continuous. The mean and median are at the centre.



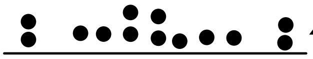


A distribution skewed to the right

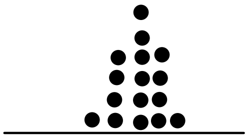
(The skew is the tail end of the data)



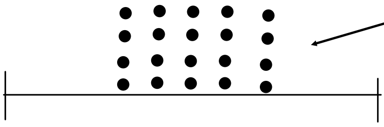
A distribution skewed to the left



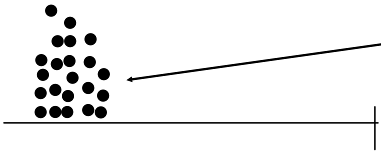
A widely spread distribution



A symmetric distribution



A uniform distribution



A tightly clustered distribution

Measures of Location

When a set of data is perfectly normal, the mean, median and mode are identical.

When the data is skewed (one tail is much longer than the other), it loses its ability to provide the best central location. Skewness is common is data sets that are essentially positive e.g. income.

The **mean** takes into account every entry in a data set and can therefore be susceptible to the influence of outliers.

The **median** is the middle value. It can be more variable than the mean but it is more robust against the effects of extreme values.

Think of some examples where using the mean instead of the median is important and vice versa.

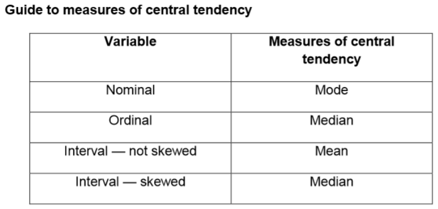
Measures of Location

Mode: The mode is a useful measure of the average in come cases. It is the most freqent in a data set.

Can you think of times when it wouldn't be useful?

There are usually many numbers in real data that aren't the same so the mode tends to be useful when we are dealing with a small number of distinct values. It is useful for categorical data when specifying the most common category.

Measures of Location



Guide to measures of central tendency

Variable	Measures of central tendency
Nominal	Mode
Ordinal	Median
Interval — not skewed	Mean
Interval — skewed	Median

Measures of Dispersion

The **range** is the difference between the max. and min. values of a data set.

Standard Deviation is the average amount of variation around the mean. It is a measure of spread of the data.

The **inter-quartile range** gives a more accurate measure of the range. It excludes outliers to the data.

(we use this when there are extreme values in the data as it is more robust)