Mean, Median and Mode

A set of numerical data is not particularly useful on its own. What patterns and mysteries does it reveal? The simplest answer to this problem is to compute an *average*, i.e. reduce the data to a single number. There are three general types of average: *mean, median and mode.*

Mean

This average is the easiest to compute. It is the **sum of all the numbers divided by the size of the dataset**.

e.g. consider a data set $\mathbf{x} = \{1, -3, 5, 0, 1, 2, 2, 4, -1\}$ There are 9 numbers in the data set, so the mean is $\bar{x} = \frac{1-3+5+0+1+2+2+4-1}{9} = \frac{11}{9} = 1\frac{2}{9}$

We can use the sum notation to generalize what we 'mean by a mean' (!) $\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i$

Is the mean meaningful? The mean will only be a *sensible average* if the data tends to be *distributed* (i.e. bunched or clustered) about the mean. An average ought to be the 'average number' after all. In situations when the dataset has *outliers*, this can distort the mean.

e.g. consider the data set $\mathbf{y} = \{1, 2, 1, 3, 2, 1, 2, 2, 100\}$ Clearly 100 is atypical and the average should be between 1 and 2.

However the mean is significantly distorted by the 100 outlier: $\bar{y} = \frac{1+2+1+3+2+1+2+2+100}{9} = \frac{114}{9} = 12\frac{2}{3}$ i.e. the mean is nowhere near any of the numbers. As an average it is *meaningless* (!)

Mode

This flavour of average describes the **most frequent number in a dataset.** Clearly this is *only useful* if it is likely that numbers *repeat*. i.e. a mode can be useful when a dataset consists of integers, but less useful when numbers are decimals (e.g. measurements of pressure, temperature etc). Of course a dataset could be *rounded* to the nearest integer, which might then result in a representative mode average.

For larger data sets the mode is slightly harder to compute than a mean, since one needs to construct a frequency table first.

$\mathbf{z} = \{1, 1, 5, 0, 1, 2, 2, 4, 2, 1\}$

Number	Frequency
0	1
1	4
2	3
4	1
5	1

In this case the **mode is 1** since it has the highest number of occurrences.

If more than one number shares this maximum frequency then the data is *bimodal*, *trimodal* etc.

Median

This is the hardest average to compute, but possibly the most representative as it is not distorted by outliers like the mean, or potentially irrelevant like the mode when no (or very few) data values are exactly the same.

The median is **the 'middle' number when the data is sorted* in ascending order.** If there is an *even* number of data values, the median is the *mean average of the middle two numbers in the sorted list.*

$\mathbf{a} = \{1.2, 4.1, 3.7, 1.1, 2.5, 6.1, 0.9, 2.3\}$

$\mathbf{a}_{sort} = \{0.9, 1.1, 1.2, 2.3, 2.5, 3.7, 4.1, 6.1\}$

There are 8 elements in data set **a**, so the median is the mean average of the 4th and 5th numbers in the sorted list

i.e. median is $\frac{2.3+2.5}{2} = 2.4$ The mean is 2.7, slightly higher due to the distorting effect of the outlier 6.1

*For large data sets, the sorting operation makes it harder to compute than a mean or mode