Statistical literacy guide
Measures of average and spread

Last updated: February 2007
Author: Richard Cracknell

Am I typical?

A common way of summarising figures is to present an average. Suppose, for example, we wanted to look at incomes in the UK the most obvious summary measurement to use would be average income. Another indicator which might be of use is one which showed the spread or variation in individual incomes. Two countries might have similar average incomes, but the distribution around their average might be very different and it could be useful to have a measure which quantifies this difference.

There are three often-used measures of average:

- **Mean** – what in everyday language would think of as the average of a set of figures.
- **Median** – the 'middle' value of a dataset.
- **Mode** – the most common value

Mean

This is calculated by adding up all the figures and dividing by the number of pieces of data. So if the hourly rate of pay for 5 employees was as follows:

£5.50, £6.00, £6.45, £7.00, £8.65

The average hourly rate of pay per employee is

$$\frac{5.5+6.0+6.45+7.0+8.65}{5} = \frac{33.6}{5} = £6.72$$

It is important to note that this measure can be affected by unusually high or low values in the dataset and the mean may result in a figure that is not necessarily typical. For example, in the above data, if the individual earning £8.65 per hour had instead earned £30 the mean earnings would have been £10.99 per hour – which would not have been typical of those of the group. The usefulness to the mean is often as a base for further calculation – estimated the cost or effect of a change, for example. If we wanted to calculate how much it would cost to give all employees a 10% hourly pay increase, then this could be calculated from mean earnings (multiplied back up by the number of employees).

Median

If we are concerned with describing a set of date by giving an average or *typical* value then it is sometimes preferable to use the median rather than the mean. The median is the value such that exactly half the data items exceed it and half are below it.

The conventional way of calculating the median is to arrange the figures in order and take the middle value. If there is no middle value because there is an even number of figures, then, conventionally, the median is take to be mid-way between the two

middle points. In the earnings example the middle value is £6.45 and this is the median for that data:

£5.50, £6.00, £6.45, £7.00, 8.65

The median is less affected by values at the extremes than the mean. It can therefore be a better guide to *typical* values.

Mode

The mode is the value that occurs most frequently. It is often thought of as not particularly useful in statistical textbooks! But in real life we often use the mode, without realising we are using a measure of average. The 'top 10', 'most popular', '2nd favourite' are simply looking at the most common, or 2nd most common values, ie. modal measures..

Grouped data

Sometimes we do not have exact values, instead the data have already been grouped into bands – 1 to 10, 11 to 20, 21 to 30 …etc. While it is not possible to exactly calculate the mean from grouped data, an estimate can be made by assigning the mid-point of each band to the observations in that group. This rests on the assumption that the actual values are spread evenly across within each band. Sometimes these classes include open-ended groups – over 50, less than 5 etc. In these cases you have to make some intelligent guess at an appropriate value. Where you have done this, you can assess how sensitive your estimate is to the assumed value for open classes by re-calculating the average using an alternative assumption (using a spreadsheet to do the calculations also makes it easy to investigate this).

It also possible to estimate the median for grouped data, by looking for the class above and below which 50% fall. Sometimes it is necessary to estimate where the 50% boundary is within a class.

Other averages

There are a number of other measures of average, some of which are briefly described below:

- **Geometric mean** - the nth root of the product of n data values
- **Harmonic mean** - the reciprocal of the arithmetic mean of the reciprocals of the data values
- **Quadratic mean** or **root mean square** (RMS) - the square root of the arithmetic mean of the squares of the data values
- **Generalized mean** - generalizing the above, the nth root of the arithmetic mean of the nth powers of the data values
- **Weighted mean** - an arithmetic mean that incorporates weighting to certain data elements
- **Truncated mean** - the arithmetic mean of data values after a certain number or proportion of the highest and lowest data values have been discarded
- **Interquartile mean** - a special case of the truncated mean
- **Midrange** - the arithmetic mean of the highest and lowest values of the data or distribution.
- **Winsorized mean** - similar to the truncated mean, but, rather than deleting the extreme values, they are set equal to the largest and smallest values that remain

Weighted average/mean

An average calculated as the arithmetic mean assumes equal importance of the items for which the average is being calculated. Sometimes this is not appropriate and you have to allow for differences in size or importance. A simple example would be if you were looking at incomes of pensioners. It the average income of female pensioners were £150 per week and the average for male pensioners £200 – it would be wrong to say that the average for all pensioners was £175 [(150+200)/2]. There are around twice as many women in this age group than men and this needs to be taken into account in calculating the overall average. If we give twice as much weight to the value for women than for men, the overall average comes to £167. The calculation of this is set out below:

	£pw	Weight	Weight x value
Women	150	2	300
Men	200	1	200
	Total	3	500

(Total, weight x value) / (Total weights) = 500 / 3 = **£167**

Measures of variation / spread

Range and quantiles

The simplest measure of spread is the *range.* This is the difference between the largest and smallest values.

If data are arranged in order we can give more information about the spread by finding values that lie at various intermediate points. These points are known generically as quantiles. The values that divide the observations into four equal sized groups, for example, are called the *quartiles.* Similarly, it is possible to look at values for 10 equal-sized groups, *deciles,* or 5 groups, *quintiles*, or 100 groups, percentiles, for example. (In practice it is unlikely that you would want all 100, but sometimes the boundary for the top or bottom 5% or other value is of particular interest)

One commonly used measure is the *inter-quartile range*. This is the difference between the boundary of the top and bottom quartile. As such it is the range that encompasses 50% of the values in a dataset.

Mean deviation

For each value in a dataset it is possible to calculate the difference between it and the average (usually the mean). These will be positive and negative and they can be averaged (again usually using the arithmetic mean). For some sets of data, for example, forecasting errors, we might want our errors over time to cancel each other out and the mean deviation should be around zero for this to be the case.

Variance and standard deviation
The variance or standard deviation (which is equal to the variance squared) is the most commonly used measure of spread or volatility.

The standard deviation is the root mean square (RMS) deviation of the values from their arithmetic mean, ie. the square root of the sum of the square of the difference between each value and the mean. This is the most common measure of how widely spread the values in a data set are. If the data points are all close to the mean, then the standard deviation is close to zero. If many data points are far from the mean, then the standard deviation is far from zero. If all the data values are equal, then the standard deviation is zero.

There are various formulas and ways of calculating the standard deviation – these can be found in most statistics textbooks or online[1]. Basically the standard deviation is a measure of the distance from of each of the observations from the mean irrespective of whether then differences is positive or negative (hence the squaring and taking the square root).

The standard deviation measures the spread of the data about the mean value. It is useful in comparing sets of data which may have the same mean but a different range. For example, the mean of the following two is the same: 15, 15, 15, 14, 16 and 2, 7, 14, 22, 30. However, the second is clearly more spread out and would have a higher standard deviation. If a set has a low standard deviation, the values are not spread out too much. Where two sets of data have different means, it is possible to compare their spread by looking at the standard deviation as a percentage of the mean.

Where the data is "normally distributed", the standard deviation takes on added importance and this underpins a lot of statistical work where samples of a population are used to estimate values for the population as a whole (for further details see *Statistical significance/confidence intervals* in this series).

Excel functions to calculate averages and spread
While it is possible to calculate these from first principles, there are a number of statistical functions in Excel which are useful shortcut ways of calculating averages and spread. Excel includes a "wizard" which can be used to insert these functions into a cell of spreadsheet. Useful functions include:

AVERAGE Returns the average of its arguments – example =Average(A1..A4)

COUNT Counts how many numbers are in the list of arguments

LARGE Returns the k-th largest value in a data set, where you determine k.

MAX Returns the maximum value in a list of arguments

MEDIAN Returns the median of the given numbers

MIN Returns the minimum value in a list of arguments

MODE Returns the most common value in a data set

[1] For example http://www.beyondtechnology.com/tips016.shtml

PERCENTILE Returns the k-th percentile of values in a range

PERCENTRANK Returns the percentage rank of a value in a data set

QUARTILE Returns the quartile of a data set

RANK Returns the rank of a number in a list of numbers

SMALL Returns the k-th smallest value in a data set

STDEV Estimates standard deviation based on a sample

STDEVP Calculates standard deviation based on the entire population

Other statistical literacy guides in this series:

- What is a billion? and other units
- How to understand and calculate percentages
- Index numbers
- Rounding and significant places
- Measures of average and spread
- How to read charts
- How to spot spin and inappropriate use of statistics
- A basic outline of samples and sampling
- Confidence intervals and statistical significance
- A basic outline of regression analysis
- Uncertainty and risk
- How to adjust for inflation