

Simple Statistics

Contributed by Joanne Birchall

Simple Statistics Written by Joanne Birchall from Rainbow Research. This article covers why we need statistics and also examines the following in detail:

- Average (Mean)
- Median and Mode
- Measures of Dispersion
- The range
- The mean deviation
- The standard deviation

Why we need Statistics

The number of people we speak to during quantitative Market Research projects can often run into thousands. Inevitably, this can result in very large sets of data. The human brain is limited in its capacity to deal with rapid incoming information, and when faced with large groups of numbers, most people cannot normally hold them all in mind at once. It is difficult to make any conclusions by simply looking at the data in its raw state; therefore it is useful to glean some kind of overall picture or summary of what is going on. The main purpose of statistics is to accurately summarise the data into easily interpretable fewer numbers. Some of the simplest statistics are described below.

Average (or mean)

Once Market Research data has been collected, a good starting point is to tabulate the data to find numbers of respondents's answers to each question, for example the number of cans of varying types of dog food purchased from a particular supermarket each week. However with large data sets, this could result in numerous sheets of data and could be difficult to take in and mentally summarise. The next step would be to get an indication of the 'normal' or 'usual' number of cans of dog food purchased from the supermarket each week. These figures are called averages and the research might say: on average about 600 cans of Pedigree Chum are bought each week from the supermarket. Notice the word about is used, as you would not expect exactly 600 cans of Pedigree Chum to be bought every week, but that there would be some variation around the figure of 600. For example if the research spanned a four week period, there may have been 535 cans sold in the first week, 692 in the second week, 550 in the third week and 623 in the fourth week.

The average figure is calculated as:

The sum of all of the numbers (535+692+550+623)

The number of numbers (or weeks) (4)

The average is one kind of descriptive statistic, which indicates a 'typical' or 'central' figure for a group of numbers. It is officially called a 'measure of central tendency'.

Median and Mode

On the whole, when numbers in a particular group cluster closely around the central value, the mean is a good way of indicating the 'typical' score, i.e. it is truly representative of the numbers. If however, the numbers are very widely spread, are very unevenly distributed, or contain extreme values, e.g. 9, 10, 13, 17, 23, 30, 45; or a hundred values of 10 and one value of 50 then the mean can be misleading, and other measures of central tendency such as the median or the mode should be used instead.

Median: If you have a set of values, and wish to obtain a figure which represents the central point, then a sensible way of doing this may be to arrange the numbers in order of size and pick the number which falls in the middle as being of typical value. For example if we had seven apples weighing 120g, 100g, 200g, 80g, 130g, 160g and 140g, if we arrange them in order of size, we get 80g, 100g, 120g, 130g, 140g, 160g and 200g. The value in the middle i.e. the fourth from the end weighs 130g, our median value.

If however there had been eight apples, we would take the weight of apples four and five as our two central numbers, and find the halfway point between them. For example, if the two central numbers were 140g and 180g, the median would be 160g i.e. you could find this mid point by adding the two numbers in question and then dividing by two, that is, by finding their average (or mean)!

Advantages of the median are:

- If one of the extreme values changes (and often in experiments it is the extreme values which are least reliable), then the median remains unaltered. Whereas the mean would be affected hugely.
- If a set of numbers has a lop-sided pattern – if for example, most of the scores are small, several medium sized, but only one or two high – then the median may again be more appropriate than the mean, as its value will be close to the majority of numbers

The disadvantages of the median are however:

- If there you have a large set of numbers, it would be time consuming to place each in order of size
- If one of the numbers near the middle of the distribution moves even slightly, then the median would alter, unlike the mean, which is relatively unaffected by change in one of the central numbers.

Mode: an alternative average is called the 'mode'. Mode means 'fashionable', which describes very well just what the statistical mode is. It is simply the value in any set of scores that occurs most often – or is the most 'popular'. Take the following set of numbers: 5, 6, 7, 8, 8, 8, 9, 10, 10, 12. As the number 8 occurs most often (three times), 8 is the mode of these set of numbers. If one of the 8s vanished, we would be left with two 8s and two 10s. In this case, there would be two modes of values 8 and 10 and this is known as bimodal. An example of a bimodal distribution is the height of six children in a nursery all aged 2 years. They are 40, 40, 40, 40, 80, 80, 80. Although the mean is 60, this figure is not a good indication of the height of any single child in the nursery aged 2 years. We would be far better off knowing that there are two modes of values 40 and 80.

If on the other hand the numbers were 5, 6, 7, 8, 9, 10, in which there is no single number that occurs more than once, there is no mode as all the numbers appear with the same frequency.

Advantages of the mode:

The mode can be a very useful statistic. One of its main assets is that it can be used to indicate a 'normal' or 'usual' figure. It is exactly opposite to the mean in this respect, as the modal value must be a commonly occurring figure. Often the value of a mean is a number with a decimal point, and sometimes may not remotely resemble any of the values in the data set – as in the nursery children. Often people use the mode as an average as in the 'average person', the figure quoted being the usual or typical value, and quite often will not be the mean. The mode is also a useful descriptive statistic when the numbers in a distribution are not evenly spread around a central value (as is the median). Such a lopsided distribution is called a 'skewed' distribution.

Disadvantages of the mode are however:

Firstly and sadly, the mode is hardly ever used, due to its instability as it can swing wildly through the whole set of numbers at the drop of a hat. Take the numbers 1, 1, 6, 7, 8, 10. The mode here is 1, which is not a very representative figure of the group as a whole. However if we change the score of 1 to a 10, the mode shifts right to the other end of the scale. Thus a single number change can alter the mode dramatically. This is in great contrast to the mean (average) and the median, where number changes can take place and leave them virtually unaffected.

If a distribution of numbers has more than two modes, and with large sets of numbers, it might be possible to have many modes – then the modal values themselves could need summarising, and so the usefulness of the mode as a descriptive statistic begins to dwindle.

Measures of Dispersion

Another type of descriptive statistic is used to qualify the word about as in the sentence 'on average about 600 cans of Pedigree Chum are bought each week from the supermarket' – in the section on averages (means). As we have already established over a four-week research period, there were between 535 and 692 cans of Pedigree Chum sold. If there were 440 cans of Pal (another type of dog food) sold in the first week, 589 sold in the second week, 670 sold in the third week and 701 sold in the fourth week, the average number of cans of Pal sold over the period would also be 600.

However, the word about signifies that there may be, and are, large departures of actual cans sold from the averages for each type of dog food. Used by itself, the word about is far too vague, and we need some means of giving more details about the variation. The solution is to use one of the descriptive statistic known as the measure of spread; these simply indicate just how much the word 'about' means for a particular set of figures, and indicates how widely scattered the numbers are. If one of these measures is used together with one of the averages, then the two summary numbers together will give an extremely concise and useful description of the particular distribution. There are three commonly used measures of dispersion (see below).

The Range

The range tells you over how many numbers altogether a distribution is spread. It is easily obtained by subtracting the smallest score from the largest. For example, if I found that various kinds of potatoes for sale in several greengrocers' shops were priced at 10p, 25p, 12p, 8p, 14p, 14p, 14p 24p and 15p per lb, the range for these prices would be $25p - 8p = 17p$.

The problem with the range is that extreme value have a very big effect on the descriptive statistic and outliers (atypical extreme values) may cause distributions which overall look very different to have similar ranges. Clearly then, the range can only be used sensibly as a descriptive statistic when all the scores are fairly well bunched together.

The Mean Deviation

The mean deviation is a number that indicates how much, on average, the scores in a distribution differ from a central point, the mean. Suppose you take the numbers 8, 9, 10, 11, 12. The mean is 10 and the range is 4. The number 8 is 2 points away from the mean, and so is the number 12. Numbers 9 and 11 are both 1 point away from the mean, and 10, the remaining number in the set is the mean, so does not differ at all. Listing these differences, you get $2 + 1 + 0 + 1 + 2 = 6$. There are 5 numbers in the group and so you can say that the average (mean) amount they all vary from the mean is 6 divided by 5 i.e. 1.2 points. The differences, 2, 1, 0, etc., which were obtained are called deviations. The larger the mean deviation is, the more spread out the scores in the distribution are. Given the calculation of the mean deviation is based on all the numbers in a distribution, it is a much more stable statistic than the range, which is only based on two of them.

The Standard Deviation

When we calculated the mean deviation (above), we did not take into account the fact that some numbers are higher than the mean and some are lower, i.e. our true deviations should have been $-2, -1, 0, +1, +2$. However, if we add these together, our total would have been zero, so we ignored the signs for this reason. Ignoring signs is not especially pleasing to mathematicians and partly for this reason, but mainly because the mean deviation is a very simple figure without any powerful mathematical properties, it is rarely used as a measure of dispersion. The preferred measure, known as the standard deviation is used far more often, usually in conjunction with the mean and the range.

In principle, the standard deviation (often shortened to 'sd') is very similar to the mean deviation. It summarises an average distance of all the scores from the mean of a particular set. But it is calculated in a slightly different manner.

As just discussed, if we take into account the signs (+ or -) of the deviations from the mean, the mean deviation will always be zero. There is however a solution to the problem. If you multiply two negative numbers together, you get a positive result. The same applies to squaring a negative number (multiplying the number by itself).

Now suppose you have the set of deviations $-2, -1, 0, +1, +2$; these equal zero when added together, or 6 when the signs are ignored. If instead of adding the deviations, you square each one you get $+4, +1, 0, +1, +4$ for the squared deviations, and all are positive numbers. Magic! Now you need to add these numbers and find their mean, just as you did with the mean deviation calculations. You should get 10, divided by 5, to obtain the value 2. The figure 10 is called the sum of squares, and 2, the mean of the sum of squares is called the variance. Notice that 2 is not the standard deviation. As we squared all the differences, the figure 2 must be 'unsquared' (find its square root), to bring it back in perspective. In this case, the square root of 2 is 1.4142. This then is the standard deviation. N.B. this is slightly higher than the figure 1.2 which was calculated as the mean deviation for the same set of numbers.

A slight (but not serious) complication with the standard deviation is the use of N (the number of scores in the particular set). The majority of quantitative Market Research methodologies involve taking a sample from a population (as described above), however because this is only a sample, some error will inevitably occur in the results. The sensible way to compensate for this is to make an allowance for it in the calculation of the standard deviation. Some statistical books tell us to divide the sum of squares by N-1 instead of N i.e. dividing by a smaller number will give a larger variance, and hence a larger standard deviation, which compensates for the margin of error in taking a sample. Thus a formula for the standard deviation involving N-1 rather than N is preferred whenever we are working with samples rather than a set of scores which is absolutely complete (a population).

If you would like any further advice about statistics for Market Research. Please do not hesitate to get in touch with Rainbow Research on +44 (0) 1772 743235.