White Paper

Technical brief: Core statistical concepts

As with any expert vocabulary, statistical terminology can be off-putting, and it is easy to tune out until the discussion comes back to something that appears more relevant to our own area of business. However, it is valuable for marketers to have an understanding of the core concepts, so this paper outlines some of the key terms and explains how the underlying principles enable more informed decision-making.

> *Many of us loathed statistics at school – but understanding statistics can help marketers make better decisions. In this occasional series of Technical briefs, we explain some essential concepts and how they are applied.*

Population and Sample

The ***population*** is the total collection of individuals of interest – for example, all UK citizens, or all short-toed sloths. When we make a measurement and calculate statistics such as the average, ideally we would measure every individual – this is known as a ***census***. Typically we cannot measure every individual – instead, we take a ***sample***, calculate the sample average, and use that sample as an estimate of the population average. Note that it is an estimate – we have introduced uncertainty by not measuring every individual. It is therefore important to be clear whether you are talking about population or sample data, because that tells you how certain your conclusion is.

Mean, Median and Mode

Many people recognise these three concepts from school maths:

- ***Mean***: the average – add them all up and divide by the number of measurements (strictly speaking, this is the *arithmetic mean*)
- ***Median***: the middle one, once the measurements are arranged in order
- ***Mode***: the most common value

They are ***summary statistics***, i.e. ways of representing a data set by a single number. They differ in how much of the information they use – the mean uses every measurement, whereas the mode throws away everything except the commonest value. For this reason the mean is usually preferred. However, it is important to remember that any summary measure only gives a partial picture, and that in some cases the mean can give a misleading picture if it is assumed to represent the 'most likely' situation. For example, almost everyone in the UK has more than the average (mean) number of legs (which is a number slightly smaller than 2), but, of greater significance for marketers, most people earn rather less than the 'average' salary as the mean is skewed considerably by a small number of very high earners. The chart below, produced by the UK Department of Work and Pensions, shows the distribution of household post-tax income (adjusted for household size and truncated to exclude the highest earners). It illustrates very clearly the difference between the different measures and the dangers of confusing the mean with the mode or the median.

Each summary measure has its place, and it is important not to be led astray by careless use of the term 'average' to describe a complex situation. A designer of trousers would be well advised to use the ***mode*** when deciding how many legs to attach; a marketer wishing to size a market might use a ***median*** income measure to understand where the bulk of people lie. The mean is undoubtedly a powerful measure, but it needs to be treated with caution. To understand when it is useful and when it is not, we need to examine the concepts of ***variance*** and ***skew.***

Variance

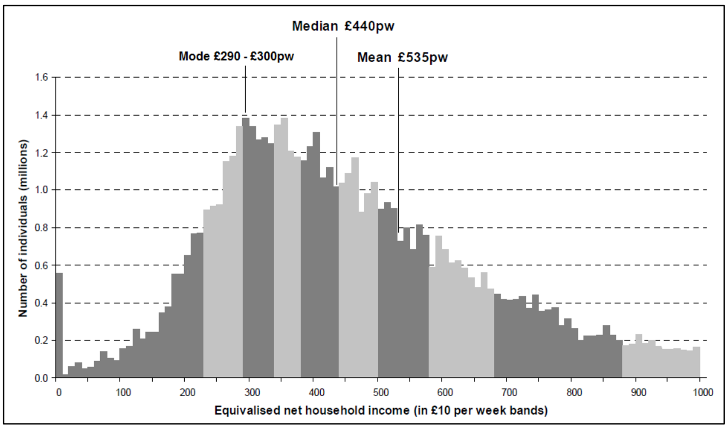
This measures how spread out the data are – whether they are tightly clustered around the mean, or cover a wide range of values. It is typically either measured as the ***variance*** of the data or the slightly more common measure ***standard deviation,*** which is the square root of the former measure. You can think of the standard deviation as the average distance of each measurement from the mean and the higher it is, the less likely a random sample will be representative of the population.

Skew

The concept of ***skew*** is essential to understand whether the mean is a good representation of the 'norm'. In a symmetrical distribution, the median and the mean (and frequently the mode) will be aligned; however the graph of income distribution is heavily skewed to the right by the small number of very high earners (not all shown on the graph). This means that using the term 'average earnings' as if it were interchangeable with 'the earnings of the average worker' is highly misleading.

Null hypothesis

Statisticians start from an inherently boring standpoint – the hypothesis that 'there's nothing interesting going on here' - and then use the sample data to try to disprove that. This absence of an effect is called the ***null hypothesis***, and is usually the opposite of what you are hoping to find in your trial: for example, 'there is no difference in response rate between these marketing concepts'. Statistical analysis then calculates the probability of observing the results seen in the test, if the null hypothesis were true. If it is unlikely that you would see those results, statisticians say we reject the null hypothesis, and so accept the alternative hypothesis – in this example, that there is a difference in response rate. See our Technical brief on ***statistical significance*** for a more detailed discussion on this.


Median £440pw
Mode £290 - £300pw
Mean £535pw
Number of individuals (millions)
Equivalised net household income (in £10 per week bands)