



## Statistical Significance

The expression “statistically significant”, or some variation of it, frequently appears in reports and presentations where there is evidence that it is not well understood. This note seeks to explain the concept and its application.

*NB. To make the principles clear some elements have been over-simplified.*

Any set of consistent data can be described in terms that make it easier to visualize. For example, the ten numbers:

22    34    22    60    17    22    33    57    34    29

Could be described by their:

Count	10	Mode*	22
Total	330	Variance	215.8
Mean*	33	Standard Deviation	14.7
Median*	31	Range (or Spread)	17 - 60

Confusingly, the Mean, Median and Mode are all sometimes referred to as the “average” yet they rarely give the same result.

There are a number of other ways to describe the data.

Quite often figures are based on the “percentage favourable response” i.e. the percentage of respondents who agreed or tended to agree with a positive statement (or who disagreed or tended to disagree with a negative statement). The score for a Category or an Index is then based on the mean percentage favourable response for the all the questions in that category or index.

Assume that the following represent Index or Category scores for a company and for a particular department:

Item	Company (20,000 respondents)	Department A (400 respondents)
1	36	41
2	67	64
3	47	52
4	50	47
Category Score (Mean of 1 to 4)	50	51

Clearly, Department A has a slightly better Category score and some ups and downs at the individual question level. The question is “Are these differences meaningful?”

All the figures are true. However, differences can be caused by natural random variation. This concept is best understood by thinking of a machine that is designed to produce items of a specific weight. Tiny variations in the machine or the material will produce differences in the finished product - usually accepted as being within or outside set “tolerances”.

Any representative number is “correct” plus or minus a bit of random variation and we use the laws of probability to judge whether any representative number is within those tolerances. We say “We are confident that the “correct” number lies between an upper and lower limit”. We can go further and express exactly how confident we are but as we express more confidence the spread between upper and lower limits gets bigger.



For example, I might be 5% confident I can hit the bull's eye with my dart, 50% confident I can hit the board and 99% confident I can hit the wall!

This spread between the upper and lower limit is the "Confidence Interval".

So in our example are we confident that the difference between the Department A Category score of 51 and the Group score of 50 is real?

The Confidence Interval is calculated separately for each figure and depends on three things:

1. The degree of confidence we want

The degree of confidence is the biggest influence and we often set this at 95% (i.e. we are 95% certain that the difference is not due to random variation).

2. The number of respondents

The number of respondents is the second biggest influence - the Confidence Interval for Department A would be around 7 times larger than that for the Group.

3. If the score is particularly high or low

The Confidence Interval is highest for scores of 50% and gets lower as scores increase or decrease.

In our example, the 95% Confidence Interval for the Category score for Department A is around + or - 5 percentage points (i.e. 46% to 56%); the difference of 1 is within this interval so is **not** statistically significant at the 95% confidence level.

In fact, for the 1 point difference to be significant one would need:

- 10,000 respondents or,
- a score of 1% (or 99%) or,
- a very low level of confidence (around 30%) or,
- some combination of these three.

*Suppose that for **Department A** we substituted **Team B** with the same scores but only 30 respondents. This time the 95% Confidence Interval would be a massive + or - 17 percentage points making all but the most extreme comparisons unreliable!*

In the department example, at the individual item level the differences for items 1 and 3 are statistically significant at the 95% confidence level; items 2 and 4 are **not**.

What conclusions can we draw from this?

1. Any results that come from small groups are very difficult to compare.
2. Even for large groups, small differences must always be treated with healthy scepticism.
3. Research often involves a choice between measurement and practicality - there is often no point in trying for both!