Marketing brief: What is a 'statistically significant' test result?

We all know that it is good practice to run a test before deploying a new marketing concept in bulk, but how should we decide whether the test has been successful or not?

Most people know that it isn't enough for the result of the test simply to be better than the control: that the result should be 'statistically significant', but what exactly does that mean? What does it mean if the result is better but is not statistically significant? Does it mean that the tested concept is no good? In this briefing note, we explain the key concepts behind statistical significance in terms that are relevant to direct marketers. For the more mathematically minded, see our technical brief on statistical significance, to see more details and how it can be calculated.

The key point to remember is that any marketing test can only examine a small fraction of the market. We all know that when looking at a small subset of any particular group, you are not guaranteed to get a representative sample: if you count the number of left-handed people in a street it will approximate to, but probably not be equal to, the level in the population. However, if you count the number of left-handers in a town, you will get a more reliable estimate of the number in the country at large.

It's only an estimate…

The number of people in the test is important to deciding how much faith you can put in the result. The important thing to remember though, is that whatever result you get from your marketing test, it is only an estimate of the response rate you would receive if you rolled out exactly the same test in bulk.

The second thing to remember is that exactly the same considerations apply to the control group, if one has been created. It is good practice to create a control group that replicates as far as possible the makeup of the test cell in terms of age, gender etc. because you want to minimise differences between the groups so that you can attribute any difference in response rate to the different marketing material that they have received as far as

possible. The key point to remember is that the observed response rate from the control group is also subject to the vagaries of random sampling. So, if the response rate from the control group was 1.0%, it is quite possible that this was a rather poor-performing sample, and that if you scaled that up, you might see a response rate of, say, 1.1%. Similarly, if the response rate from the test group was 1.2%, does this represent the expected response rate for that marketing material in the population at large, or is this particular sample better than the average? With a response rate of 1.2% from the test group, as a cautious marketer you should accept that if scaled up, you might not get that same rate again. For example, you might only get 1.1% - no better than the response rate from the control material.

An arbitrary choice

The question we need to answer when looking at a positive marketing test is: how *likely* is it that this result is genuine, and is not just a result of the random nature of the test and control samples? Although people talk about whether something 'is' or 'is not' statistically significant, it isn't really as clear-cut as that. Statistical analysis can tell us exactly how likely the observed result is to have come about by pure chance, and what we are really asking is whether this possibility is below a certain acceptable level of risk.

Exactly how certain you want to be depends on a number of factors, typically a weighing up of the possible downside if you are mistaken, with the potential upside if you are correct. So, it is not strictly accurate to say that a result 'is' or 'is not' statistically significant; what you actually do is to decide how confident you want to be in the result, and then see whether or not the test passes that threshold.

The confidence level conventionally applied for many purposes is 95%: i.e. if there is only a 5% chance that the result you are seeing has come about by chance then we say that it is statistically significant (technically, we should add the confidence level to be more precise). Two things will be immediately apparent from this:

- The choice of 95% is a somewhat arbitrary decision and not fixed in stone
- Whether something 'passes' or 'fails' is not as black and white as it might appear.

When a decision is a matter of life or death, such as the safety of a new medicine, then it makes sense to require a very high level of confidence before proceeding. With a marketing test, it is a purely commercial decision and will depend, amongst other things, on the relative cost of the different options. If the test material is a lot more expensive than the control, then your commercial director might want to be 95% certain that the response rate will definitely be higher. However, if it is roughly the same cost, a lower confidence interval might be perfectly acceptable. If it is significantly cheaper, then all you really need to know is that it won't actually harm response rates.

And if it fails the test?

The important thing to remember is that if your resident statistician tells you that your positive result is not 'statistically significant', that does not mean that the marketing test has failed. It just means that you can't be 95% confident that it has succeeded. In fact, your statistician should be able to tell you just how certain you can be about the results. If you still feel personally convinced by your marketing material then you can run another test and hope that this time the results are more conclusive. Or you can consider increasing the scale of the test or control. In general the sampling error decreases with the square root of the number of people - so if you have four times as many people in your test cell, the likely error is halved and a result that was not statistically significant before may become so.