Technical brief: Calculating statistical significance

The control mailing had a 1.0% response rate, and the test had 1.2%. That looks good, but then someone asks if it is statistically significant. What do they mean, and how is it calculated?

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.

When we make a measurement and calculate statistics such as the mean, typically we can't measure every individual – instead, we take a *sample*. So, if we want to know the average height of 10-year-olds in England, we would measure the height of a random sample of 10-year-olds, find the mean, and use that as our *estimate* of the true mean in the total population.

Sampling error introduces uncertainty

However, if we take a different random group of 10-year-olds, we will get a different estimate – just because we have got a different mix of individuals in the sample. This is known as *sampling error* – which does not mean that the researcher has done anything wrong, it is just the inevitable consequence of taking a sample. Because we haven't measured everyone, we have introduced *uncertainty*.

Similarly, a test mailing is a sample – so the 1.2% above suffers from sampling error, and there is uncertainty in the response rate you expect to get when you scale up.

Now, assume we take this measuring experiment a step further, and continue taking samples, measuring, and calculating the mean for each one until we have a long list of means. If we count how often each value of the mean crops up and plot the results on a bar chart, we get a graph that looks something like Figure 1. Importantly, the pattern of the frequencies of these means is predictable, with extreme values of the mean being less likely than the ones in the middle – as shown by the arrows on the chart. That makes intuitive sense: if you are picking a random sample of children to measure, you are less likely to get a bunch of really tall ones, and more likely to pick a mixture, with a mean somewhere in the middle of the range – but just by chance, it is possible you could pick a sample of 6-footers.

Because that pattern is predictable, and can be described mathematically, we can calculate how likely it is that we will get a mean above a particular value – and that is the basis of statistical significance. It is a concept that is relevant whenever a sample is used to draw a conclusion about a larger population – because of the uncertainty due to sampling error.

Quantifying the uncertainty

So, going back to the test mailing: was the response rate really higher, or was it just a high-responding sample? We can treat a proportion (i.e. percentage) like a mean, so applying the logic above we ask that question mathematically:

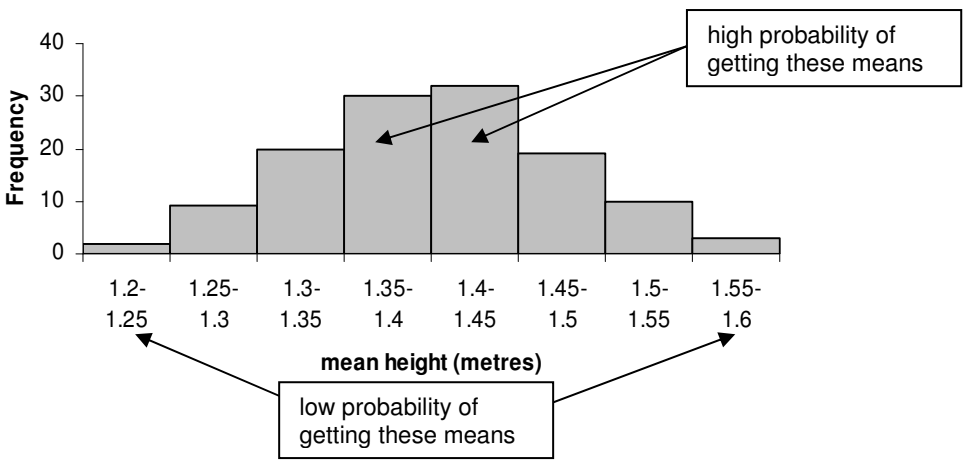
> *"How likely is it that we would observe a response rate as extreme as this, if the true response rate to the test was the same as the control?"*

If, for example, we calculate that there is a 9 out of 10 chance of observing 1.2% when the true rate is 1.0%, then it is entirely within the bounds of possibility that the two mailings are equivalent, and the difference is just sampling error - you have got more responders in your sample by chance. However, if there's only a 1 in 1000 chance, then it is more likely the test does in fact generate a higher response.

Notice that none of these are certainties – all we are saying is whether it is likely that there is a difference. So what is a reasonable probability to use, to decide one way or the other? Conventionally, it is 5% – i.e. if there is less than a 5% chance of getting such an extreme result in the test, we will call that a difference. We then say that the difference is statistically significant at the 5% level.

That 5%, though, is entirely arbitrary; it is conventionally accepted as a reasonable level of risk – because remember, even if there is a less than 5% chance of observing that response rate, there is still a chance that the two mailings

Figure 1: mean heights from repeat sampling



actually generate equivalent response rates. You could choose where to set the level at which you declare significance, depending on the level and type of risk of the decision.

So what is the decision on this test mailing – are we sure it is better than the standard one? That depends on the sample size – how many people you sent the test to.

- If you sent it to 10,000 people, then NO – there is an 18% probability of getting those results when there is no difference in reality
- You would need a response rate of 1.3%, at this scale, to say it is statistically significant
- If the test was 21,000 people, then PROBABLY – there is only a 4.9% chance that the response rates are the same.....but is that enough for your CEO to go with the new version?

Nuts and bolts – calculating significance for response rates

Work with the *difference* between the rates, and determine whether it is possible that the true difference is zero: if it is, we say it is not significant. The most straightforward way – avoiding using statistical tables – is calculating a *confidence interval*.

- Calculate the *standard error* of the difference. This is a measure of the uncertainty, and is calculated as:

$$\text{SE(diff)} = \sqrt{\frac{p_t(1-p_t)}{n_t} + \frac{p_c(1-p_c)}{n_c}}$$

  where p_t, n_t are the proportion and sample size for the test, and similarly for the control.

- Calculate the *95% confidence interval*. This is the range of values of the difference, within which there is a 95% chance that the truth lies, and the edges of the range are:

  difference - 1.96 x SE
  difference + 1.96 x SE

- If this range includes zero, then your result is not statistically significant at the 5% level