Significant differences

The last gasp of the academic year, as far as assessment is concerned, is the marking of MSc project reports. At least here in Biomedical Sciences, MSc students carry out two research projects, each about 5 months long and each in a laboratory of their choice (except when too many want the same lab, something that rarely happens). In general, with this amount of time available, the students do very well and produce work that really does advance the frontiers a little. Many students are able to publish their work externally. Whether or not they publish, all students have to write project reports that are marked by two markers not involved in supervision of the project, and these marks are used for classification of the degree (fail/ pass/ distinction).

Marking these reports is usually very enjoyable, partly because the work is interesting and partly because the intolerable noise of Edinburgh's Festival Fringe, which literally surrounds this building, is an excellent excuse to leave the office early and to take marking to the beach. This is much nicer than being in an overheated office! One issue has, however, been annoying me occasionally for some years and, this year, it seems to have come up again and again. It is the perverse use of the word 'significant', usually as part of the phrase 'significant difference'; the problem is not one of literary style but rather the proper separation of the concepts being discussed.

The problem arises from a technical use of the term 'significance' in statistics, a use that seems to have eclipsed the normal meaning of the word in the minds of many students (and professionals, at times). Statistical analysis is common in biology because experiments are “noisy”; no two cells or animals or patients are exactly alike, and our measuring tools and imperfect and add variation of their own. So we run an experiment several times on as many samples as we can manage, and hope that the inherent variation across these 'replicates' is small enough that effects of the actual experiment – the addition of a drug, for example – show through. One method of applying statistical analysis to an experiment, a method that has been very popular for the past few decades, may seem counter-intuitive. We posit a formal hypothesis that there is in fact *no* difference in a measured outcome (eg cell size) between the experimental group (eg cells given the drug) and the control group (eg cells given nothing special). Then we test whether this hypothesis can be *rejected*. There are various statistical tests, each suited to different types of data, that can test the truth of that hypothesis that there is *no* difference. Statistical tests never say 'yes' or 'no'; they give

answers in terms of probability. A test might therefore say something like “there is a probability of less than on in one hundred, that the hypothesis of no difference between the groups is true”. It has become a sort of convention that when a hypothesis of no difference has a less than one-in-twenty chance of being true, we say there is a 'statistically significant difference' between the experiment and control groups. Sometimes people use stricter criteria than one in twenty, but the general pattern and words remain.

There is nothing wrong in principle with doing this, although I have always felt that the 'hurdle' of one in twenty (or whatever) is an unnecessary complication and I prefer simply to report the probability without making any claim of 'significant'. But most people do make the claim (and even some people in my lab insist on doing so, so that we end up compromising by including actual probability and having the claim of statistical significance). But the word 'significant' is really unfortunate. In ordinary English, 'significant' means 'leading to a different result or an important change' (Cambridge English Dictionary), and this is different from statistical significance.

To understand the difference, imagine a tyre manufacturer who claims that Wundarubba tyres last significantly longer than BadYear tyres. You know your usual BadYear tyres last around 20,000 miles of driving. How much longer would Wundarubba tyres have to last in order to be significant to you – significant in the sense that it is worth your bothering to go to a new place to buy them or to pay more. Five thousand miles more? - sure. One thousand? - maybe. One hundred? - hardly worth the bother. Ten miles? - obviously not. If Wundarubba tyres lasted only one mile more on average, but they did a large enough test (feasible with the vast number of tyres sold each year), their manufacturer could almost certainly make a genuine, data-supported claim of a 'statistically significant' increase in tyre life. The point, of course, is that statistical significance is no measure of whether something is actually important. It is a kind of minimum criterion, that's all.

The measure of how much difference something makes (one mile, ten miles, a thousand miles etc) is called the 'effect size'. In most contexts, effect sizes need to be large – at least a few percent - before the common English meaning of 'significant' is satisfied. Nevertheless, I read in project reports, again and again, great claims that a drug has a significant effect on a test system when there strong statistical significance (eg only a less than 1% chance of the hypothesis of no difference being true), but only an utterly pathetic effect size that would make no difference to any practical outcome in the context of cells or patients. This is exactly the kind of thing that fuels so many silly

Tabloid newspaper reports of things that are good and bad for health, and that cumulatively gets medical science such a bad image.

So, I have a New Year's resolution (as in “new academic year”); to try to explain to students that there is a significant difference between a 'significant difference' and a mere 'statistically significant difference'. I wonder if I can find any statistics to prove it….

Jamie Davies
Edinburgh
August 2019