Goodbye P value—is it time to let go of one of science's most fundamental measures?

March 10 2015, by Lewis Halsey

Credit: Karolina Grabowska from Pexels

How should scientists interpret their data? Emerging from their labs after days, weeks, months, even years spent measuring and recording, how do researchers draw conclusions about the results of their experiments? Statistical methods are widely used but our recent research in *Nature Methods* reveals that one of the classic science statistics, the P value, may not be as reliable as we like to think.

Scientists like numbers, because they can be compared with other numbers. And often these comparisons are made with statistical analyses, to formalise the process. The broad idea behind all statistical analyses is that they allow the researcher to make somewhat objective assessments of the results of their experiments.

Which drug is more effective?

Scientists often conduct experiments to investigate whether there is a difference between two conditions: do people get better more quickly after taking the blue pill (condition one) or the red pill (condition two)? The most common method for assessing if the pills differ in their effectiveness is to undertake statistical analysis of where some patients were given the blue pill and some the red, and from this determine whether there is strong evidence that one colour is more effective than the other.

To assess experimental results, scientists very often use a "P value" (P is for probability). This is used to show how convincing these results are: if the P value is small, they think that the findings are real and not just a fluke. In our pill example, if P is small this is considered good evidence that there is a difference in effectiveness of the two colours of pill.

Although P is never proof that there is a difference – scientific studies never prove things, they only provide a degree of evidence for them – studies with low P values are thought to be convincing, and so are not often repeated to be sure the results are correct. This might seem reasonable because there is limited money and time in science – results from a study that seem very clear perhaps do not warrant doublechecking when there are new discoveries out there to be made.

P values are fickle friends

However, we have used simple models to show that the P value often varies dramatically if a study is replicated. Our models depict a simple scenario. Samples have been measured from two conditions. A statistical test called a t-test is conducted to investigate whether there is good evidence that the conditions are different, and the test result is interpreted by the generation of a P value.

The two conditions in our scenario are indeed somewhat different and so we might expect a reasonable sample size to uncover this difference. That is, a reasonable sample size will return a low P value associated with the t-test. However, when we repeat the model experiment many times over, we find that the P value varies dramatically each time.

If your friend has invited you round for dinner next week but in the preceding days keeps contacting you and giving dramatically differing arrival times, you will soon conclude you have very little idea of what time dinner will actually be. Similarly, if P varies considerably each time an experiment is conducted, this makes the P value unreliable, and a poor measure of how strong the evidence is from a single run of that experiment.

The implication is huge for data analysis –- a low P value returned from a study is likely to have as much to do with luck as it has to do with the presence of an important pattern in the data, and in turn a re-run of the experiment might well result in a very different P value. Therefore, a low P value for a single experiment cannot be taken as good evidence that there is a difference between the conditions.

This weakness could well explain why famous scientific findings from the past, central to the foundations of many disciplines, are not being confirmed now that the original studies are finally being re-examined.

These include a lack of reproducibility in cancer research, as well as the apparent loss of the phenomenon called "verbal over-shadowing" whereby people shown a face and asked to describe it are less likely to recognise the face later on than if they had simply looked at it.

So why is the P value so variable, so fickle? Unfortunately it seems that some degree of variability between the samples for each occurrence of an experiment creates an unstable P value.

Moving on

So if not the P value, what should we use to analyse and interpret our data? We argue for a fundamental shift in thinking away from asking the question "is there a difference?" and towards asking "how big is the difference?". After all, scientists rarely want to know simply whether there is a difference between conditions.

There is always a difference, even if extremely small. It is more pertinent to ask whether the difference is big enough to be of interest, to be of importance. If the effectiveness of the red pill is just 0.01% greater than that of the blue pill, there is a difference between them but it isn't noteworthy – in practice one pill colour is as good as the other.

The P value can be ditched and scientists can focus instead on how big the difference is between the conditions according to their experiment. They can also provide simple-to-calculate values on how precise that difference is likely to be when generalised beyond the laboratory.

Thus once data collection has finished, scientists should focus on estimating how big the difference is in the effectiveness of the blue and red pills, and how precise this estimate is likely to be. Researchers already know about these simple concepts – effect sizes and confidence intervals – they just need to start emphasising them, and let the P value become a thing of the past.

Unfortunately, while a smattering of journals have now started to outlaw the P value in recognition of some of its failings, recently at least one journal has also banned the use of the confidence interval, apparently because its precise statistical definition risks it being over-interpreted and misunderstood.

A reasonable counter to this point of view is that confidence intervals are a valuable tool for estimating the margin of error around our findings – they are a crucial measure when translating our sample of data collected in the laboratory into an understanding of real world scenarios, where results really matter.

This story is published courtesy of The Conversation *(under Creative Commons-Attribution/No derivatives).*

Source: The Conversation

Citation: Goodbye P value—is it time to let go of one of science's most fundamental measures? (2015, March 10) retrieved 1 May 2024 from https://phys.org/news/2015-03-goodbye-p-valueis-science-fundamental.html