Methods: Can we be confident in our statistics?

Thom Baguley on the difference between statistically significant and non-significant effects

Last September, Sander Nieuwenhuis and colleagues (2011) published a paper on the failure of neuroscience researchers to conduct proper tests of interaction effects. It caused a bit of stir, and briefly made it into the mainstream media, with a Guardian piece from Ben Goldacre (2011; see also The Psychologist, October 2011, News, p.728). The error in question is one that should be familiar to psychologists. It concerns the case of two differences, one statistically significant and one non-significant. For example, Group 1 may show a significant difference between a drug and placebo condition, while Group 2 do not. A naive interpretation is that the drug works for Group 1 but not Group 2. This is not necessarily true. The proper test of the difference in the effect of the drug between groups is an interaction test (a test, in this context, of differences between differences). The issue is primarily a conceptual one: to claim that one effect (e.g. a difference in means) is bigger than a comparable effect in another sample, you need to support the claim with evidence about the difference in effects.

Psychologists, it has been argued, tend to avoid this error. We have been heavily trained in ANOVA as undergraduates (certainly in the UK and probably also in the US and most of Europe). Even if we fail to learn this, reviewers and editors in our discipline tend to spot the error. Indeed, although Goldacre initially pointed the finger at psychology research, a number of psychologists reacted by suggesting that this error was uncommon in psychology or evidence of something dodgy about neuroscience research, and the piece was later (13 September) amended to refer specifically to neuroscience.

Are psychologists right to feel smug? Perhaps, but only a little. True, we are relatively good performers on this point because we tend to view many statistical analyses through an 'ANOVA lens'. Factorial ANOVA (in which factors are orthogonal) includes the interaction term by default, and 2x2 factorial ANOVA is the workhorse of experimental psychology. Our familiarity with this type of design and analysis makes some errors easy to spot. But our ANOVA lens leads to other errors – notably dichotomising continuous variables (e.g. via median split) in order to squeeze them into an ANOVA design. This always decreases statistical power, and can – albeit infrequently – produce spuriously significant effects (see MacCallum et al., 2002). Such errors are probably less serious than the difference between differences (interaction) error, but are not harmless.

The real test then, is whether psychologists make the same (conceptual) error

in a different context. The obvious context is that of association, rather than difference. If males show a significant correlation between testosterone and aggression (e.g. $r(16) = .44$, $p < .05$) and females don't (e.g. $r(16) = .37$, $p > .05$), the correlation between testosterone and aggression is not significantly bigger for males than females. (If this example seems contrived, I should mention that the figures, but not the context, are taken from an article in a high-impact psychology journal where it was claimed that one association was stronger than the other.) To confirm this you'd need to construct a test or (better still) confidence interval (CI) for the difference in correlations. This is hardly ever done – and, in my experience, psychologists commonly make this kind of claim without backing it up. Also note that the error can work the other way. Two correlations could be non-significantly different from zero but different from each other (e.g. $r(8) = .50$ and $r(8) = -.50$).

Is there any evidence to support my position? Yes. There is anecdotal support (e.g. from editing or reviewing many dozens of papers). But in addition, I conducted a quick survey of articles in psychology. I sampled two recent (2011) issues of Psychological Science – a high-impact journal that publishes a wide range of empirical papers. (The choice of Psychological Science was largely one of convenience. I would expect to find the same kinds of errors in other empirical journals.) These were not chosen at random – indeed one issue was selected because I had previously noticed an article that made an error (failing to report an interaction effect when concluding that a difference was present in one sample but not another). The other issue was selected because it had the same number of articles as the first issue (20). Of the 40 articles I surveyed, 25 claimed or implied differences between differences for one or more factorial ANOVA (or similar) designs. Likewise, 14 claimed or implied differences in correlations (or related effects) one or more times. For studies reporting differences between differences, the proportion appearing to making the error at least once was .12, 95% CI [.03, .30]. For studies reporting differences between correlations, the proportion appearing to make an error was .57, 95% CI [.31, .80]. Ignoring the ANOVA case (where, as expected, the prevalence is lower), the proportion of articles making the error in psychology is comparable to the proportion making the error in Nieuwenhuis et al.'s neuroscience sample (around 50 per cent). Although the journal includes articles by researchers outside psychology, very many of the errors arose in articles written – and presumably reviewed and edited – by researchers based in psychology departments.

This conclusion rests on concluding that a proportion in one sample is lower than that in another without evaluating the difference in the proportions directly. If the two proportions are independent, the lack of overlap of their respective 95% CIs suggests that their difference is greater than zero ($p < .05$) – though overlap would not necessarily imply non-significance. As the

proportions are an awkward mix of independent and dependent observations it may be better to focus on the 11 articles where both types of error could arise. Of these, one shows both errors (a tie) and seven show only the correlation error ($p < .05$ by a McNemar exact test).

Note also that presence or absence of overlap between CIs around two effects is not sufficient to determine statistical significance or non-significance of their difference. If the effects are independent, the joint width of the two separate CIs is usually around 40% larger than the CI for their difference, and thus the procedure is too conservative. If the effects are dependent, then the width of the CI for their difference depends on the degree to which the effects are correlated. For a more detailed discussion see Baguley (in press). Lessons

There are, I think, some interesting lessons to be learned here. Interactions are bit more complicated than psychologists (particularly those very familiar with ANOVA) often think. This concept (that the difference between significant and non-significant is not necessarily also statistically significant – see Gelman & Stern, 2006) is probably quite tricky. It is worth exploring the factors that lead people to make the error.

At the heart of it is the shift from thinking about the estimate of an effect (e.g. a difference or correlation) to the outcome of a hypothesis test. The outcome of the test is a dichotomous decision that conceals the uncertainty inherent in the original data. This is a more subtle point than that hypothesis tests (and significance tests in particular) encourage dichotomous thinking or that there is a 'cliff effect' at the threshold of the decision (Rosenthal & Gaito, 1963) – though these are likely to be facets of the problem.

It is also the case that researchers may misunderstand even quite familiar procedures. With respect to ANOVA, it is common to treat simple main effects as if they are ways to partition an interaction effect into its constituent parts. This is not correct. Simple main effects are decompositions of variance attributable to one of the main effects plus variance attributable to the interaction. Thus simple main effects are not 'pure' components of an interaction (except when a main effect explains no variance, and even then they tend to lack power relative to a focused interaction test).

The ANOVA version of the error may therefore arise through a misunderstanding of simple main effects (e.g. the mistaken belief that a combination of significant and non-significant simple main effects implies a significant interaction effect).

Unfortunately, methods for testing or constructing CIs for differences between correlations are not widely known. The standard procedures are, in fact, a bit fiddly (e.g. depending on overlap or lack of overlap in the

measurements), and rarely taught at undergraduate or postgraduate level. The methods that are taught are also often rather inefficient (e.g. see Zou, 2007, for some better alternatives). It is not surprising that people fail to report results from procedures that are rarely found in popular research methods texts.

The error may also sometimes arise in the presentation of results, rather than in the interpretation. This might happen if the pattern of effects is clear-cut and obtained from a large sample. Authors (possibly at the behest of reviewers or editors) may have omitted the crucial information required to demonstrate the effect because it seems obvious, or to reduce clutter in the results. This is perhaps defensible for peripheral results – particularly if sufficient information is reported to allow readers to check for themselves. Unfortunately, many of the errors I observed (and many of those documented by Nieuwenhuis et al.) are neither accompanied by appropriate descriptive statistics nor relate to peripheral findings.

On a positive note, it appears that the error can be avoided with appropriate training (although it appears that both the ability to detect the problem and the tools to avoid it need to be taught). The problem is also likely to be reduced if researchers report interval estimates rather than point estimates of effects.

A CI draws attention to the uncertainty in the estimate and therefore should make it less likely that a non-significant effect is automatically assumed to differ from a significant one.

Thom Baguley is Professor of Experimental Psychology at Nottingham Trent Universitythomas.baguley@ntu.ac.uk **References**

Baguley, T. (in press). Calculating and graphing within-subject confidence intervals for ANOVA. Behavior Research Methods. doi: 10.3758/s13428-011-0123-7

Goldacre, B. (2011, 9 September). The statistical error that just keeps on coming. The Guardian. Retrieved from tinyurl.com/3rzdua6

Gelman, A. & Stern, H. (2006). The difference between 'significant' and 'not significant' is not itself statistically significant. American Statistician, 60, 328–331.

MacCallum, R.C., Zhang, S., Preacher, K.J. & Rucker, D.D. (2002). On the practice of dichotomization of quantitative variables. Psychological Methods, 7, 19–40.

Nieuwenhuis, S., Forstmann, B.U. & Wagenmakers, E-J. (2011). Erroneous analyses of interactions in neuroscience: A problem of significance. Nature Neuroscience, 14, 1105–1107.

Rosenthal, R. & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. Journal of Psychology, 55, 33–38.

Zou, G.Y. (2007). Toward using confidence intervals to compare correlations. Psychological Methods, 12, 399–413.