A version of this article appeared in The Psychologist (2008) in volume 21, part 3 (March) pages 224-7.

Thom Baguley *(Nottingham Trent University)*

Statistical practice is peppered with examples of ‘magic’ numbers such as the infamous alpha = .05 of null hypothesis significance testing. Similarly, sample size calculations typically assume a desired statistical power of .80 and nearly always use ‘canned’ effect sizes such as d = 0.5 (Lenth, 2001; Baguley, 2004). Some of the most interesting cases are for commonly used cut-off criteria, such as keeping factors where the eigenvalue > 1 in factor analysis; requiring reliability > .70 in scale construction; or a goodness-of-fit-index or GFI > .90 in structural equation modelling (Lance et al., 2006). What’s interesting is that these ‘magic’ numbers all have several things in common.

First, they all seem to have arisen by a process of academic Chinese Whispers (see Vicente & Brewer, 1993, for empirical evidence of this kind of effect in use of citations). Take the requirement for reliability > .70. Lance et al. (2006) show that where a source is given for this cut-off it is almost invariably Nunnally’s Psychometric Theory. However, the closest Nunnally comes to saying this is:

In the early stages of research . . . one saves time and energy by working with instruments that have only modest reliability, for which purpose reliabilities of .70 or higher will suffice…
(Nunnally, 1978, p.245)

Nunnally even notes that reliabilities in excess of 0.90 may be inadequate in many situations. The problem is that while early citations of Nunnally’s book were largely accurate, later authors cited him without consulting the original text. Fairly soon the nuance and context of the original source is lost and the citation becomes the ritual acknowledgement of a common authority. Eventually reliability > .70 becomes commonly accepted wisdom– more accurately myth or urban legend (Vandenberg, 2006) – and a citation is no longer required.

Second, they are – as far as I can tell, without exception – wrong. To be fair, they do differ in degree of wrongness: ranging from mildly misleading to terminally myopic. In addition, they are all wrong for more or less the same reason. In every case they involve taking a continuous quantity and reducing it to a single number. Take, for instance, the a = .05 threshold. As Rosnow & Rosenthal put it:

Surely, God loves the .06 nearly as much as the .05. Can there be any doubt that God views the strength of evidence for or against the null as a fairly continuous function of the magnitude of p?
[Rosnow & Rosenthal, 1989, p.177]

This isn’t just the prerogative of null hypothesis significance testing (though it can claim some of the most widespread and most egregious cases). For example, in a recent paper on alternative (Bayesian) approaches to statistical inference Wagenmakers (2007) wryly comments that ‘people apparently find it difficult to deal with continuous levels of evidence’ before discussing the lumping of Bayes factors into discrete categories, such as that suggested by Raftery (1995). Wagenmakers points out that ‘statistically, this desire [to reduce a Bayes factor to a dichotomous decision] is entirely unfounded’. I would argue that this conclusion could readily be generalised to other statistics. It isn’t always a good idea to present findings as the output of such a dichotomous decision. Even when it is necessary (or convenient) to reach a decision on this way, it makes no sense to use the same threshold or cut-off in all situations. My own favourite example is collinearity in multiple regression. Collinearity does not simply arise when two predictors are correlated > .90 (though very high values can make computing the coefficients tricky depending on the software you use). As long as there is any shared variance between predictors it will be somewhat difficult to tease the effects of the predictors apart, and the precision with which their individual effects are measured will suffer (and thus statistical power to detect effects of individual predictors is compromised).

It isn’t all doom and gloom, though. I recognise that people sometimes find simple rules of thumb useful, and many of the examples I’ve used started off in just this way. Cohen’s effect size guidelines and recommendations for power calculations are a case in point (e.g. Cohen, 1969). In particular, estimating statistical power for supposedly ‘small’ (d = .2), ‘medium’ (d = .5) or ‘large’ (d = .8) effects was at first intended only as a method of last resort. The problem is that when these rules of thumb become articles of faith in the rituals of statistics, the benefits they provide are lost. The solution is simple. Just remember that although these ‘magic’ numbers can occasionally be useful, they are always wrong. Don’t end up doing statistics by numbers.

References

Baguley, T. (2004). Understanding statistical power in the context of applied research. Applied Ergonomics, 35, 73–80.

Cohen, J. (1969). Statistical power analysis for the behavioral sciences. New York: Academic Press.

Lance, C.E., Butts, M.M., & Michels, L.C. (2006). The sources of four commonly reported cutoff criteria: What did they really say? Organizational Research Method, 9, 202–220.

Lenth, R.V. (2001). Some practical guidelines for effective sample size determination. The American Statistician, 55, 187–193.

Nunnally, J.C. (1978). Psychometric Theory (2nd Ed.). New York: McGraw-Hill.

Raftery, A.E. (1995). Bayesian model selection in social research. Sociological Methodology, 25, 111–196.

Rosnow, R.L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. American Psychologist, 44, 1276–1284.

Vandenberg, R.J. (2006). Introduction: Statistical and methodological myths and urban legends: Where, pray tell, did they get this idea. Organizational Research Methods, 9, 194–201.

Vicente, K.J. & Brewer, W.F. (1993). Reconstructive remembering of the scientific literature. Cognition, 46, 101–128.

Wagenmakers, E-J. (2007). A practical solution to the pervasive problems of p-values. Psychonomic Bulletin & Review, 14, 779–804.