Emily Dickinson and monkeys on the stair Or: What is the significance of the 5% significance level?

'Once is happenstance, twice is coincidence, three times is enemy action.¹' But when does statistical significance kick in? Statisticians set the 5% level as Biblical; but as **Mark Kelly** points out, it ain't necessarily so.

Significance levels are what statisticians live by. They say that anything that stands more than one chance in twenty of having happened accidentally is not statistically significant. If it stands less than one chance in twenty of having happened by chance, then it is statistically significant. One in twenty is 5%. On that cut-off point statisticians divide sheep from goats, success from failure, those ideas that shall live from those that shall vanish into darkness and die.

They call these chances p-values, and write than as decimals, .05 instead of 5%; but that is only their jargon. The point is that this is the level by which they judge the events and the laws of the universe. God, or Nature, or the Human (or statistician's) Mind says that 5% is what divides the random from the planned. Are they right?

There is only one thing more tedious than complaining about the abuse of p-values and that is complaining about the amount of complaints about the abuse of p-values. That's why a paper I read recently² on the origins of the 5% significance level is so interesting. Anyone who has explained significance testing will have heard themselves uttering something along the lines of 'there's nothing magical about the 5% level, it's an arbitrary threshold'. The paper that intrigued me, by Michael Cowles and Caroline Davis, covers the source of why we routinely use 5% as the standard level of significance. So is there something special about 5%? No. There isn't. What is interesting though is that other levels are special, or at least are more objective.

For a start Sir Ronald Fisher, the person most credited with promulgating the 5% level, was quite clear that 'Deviations exceeding twice the standard deviation are thus formally regarded as significant.' Exceeding two standard deviations equates to testing at the 4.55% significance level. The difference between 4.55% and 5% may seem trivial but, in the world of significance testing where we are expected to close the valves of our attention like stone (see below: Emily Dickinson was an intuitive hypothesis tester) to p-values above our threshold, these things matter. The distribution of p-values in published research has been shown to have a preponderance of scores just below the 5% significance level³. Lowering to an arguably more objective 4.55% level might just be shifting goalposts of course, but on a practical level it would have an impact. Fisher went on to say 'Personally, the writer prefers to set a low standard of significance at the 5 per cent point, and ignore entirely all results which fail to reach this level.' So an entire field has been guided by one (albeit remarkable) individual's personal preference. He continues: 'A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance.'

So, what do other luminaries have to say? Here's Karl Pearson (of chi-square fame) describing a p-value of 0.1: 'Not very improbable', he says, of this value that could have happened ten per cent of the time by chance. And a p-value of 0.01 – one time in a hundred by chance – was 'This very improbable result'. Cowles and Davis in their paper suggest choosing the point halfway between these two – which is inconveniently 5.5%. How about William Sealy Gosset, statistician and brewer, responsible for the t-test as well



Daguerreotype of Emily Dickinson, c. 1848. (Yale University Manuscripts & Archives Digital Images Database)

The soul selects her own society, Then shuts the door; On her divine majority Obtrude no more. Unmoved, she notes the chariot's pausing At her low gate;

Unmoved, an emperor is kneeling Upon her mat.

I've known her from an ample nation Choose one; Then close the valves of her attention Like stone.

Emily Dickinson

Dickinson has been called the poet of exclusion. The soul chooses just one to love from all the millions and ignores all others, no matter how tempting. Do statisticians do the same? as some fine Guinness? He ventured 'Three times the probable error in the normal curve, for most purposes would be considered significant'. Three times the probable error, being equivalent to 2.02368 standard deviations, implies testing at the 4.3% level. The paper goes on to quote Wood and Stratton⁴, who in 1910 opined '30 to 1 as the lowest odds which can be accepted as giving practical certainty that a difference is significant'. Obviously this implies testing at the 3.33% level.

No one therefore has come up with an objective statistically based reasoning behind choosing the now ubiquitous 5% level, although there are objective reasons for levels above and below it. And no one is forcing us to choose 5% either.

A one in ten chance is 'not very improbable'; a one in a hundred chance is 'very improbable', said Pearson

The ICH E9 guidelines that harmonise statistical practice in clinical trials⁵ are explicit: 'Conventionally the probability of type I error is set at 5% or less or as dictated by any adjustments made necessary for multiplicity considerations ... Alternative values to the conventional levels of type I and type II error may be acceptable or even preferable in some cases'. Having checked the quidelines for authors of many of the medical

The 'Monkey on the Stairs' experiment placed a group of monkeys in a cage with a set of stairs leading toward a hanging banana. When any monkey stepped on the stairs all of the monkeys were sprayed with ice-cold water. After a while any monkey who attempted to step on the stairs was attacked by the group. Slowly the original monkeys were replaced with new monkeys, until eventually none of the monkeys in the cage had been sprayed with water. They continued to attack any monkey that attempted the stairs – even though none of them now knew why.

Interestingly, this story itself is apocryphal and has turned into a kind of monkeys and the stairs story itself, with people repeating it having never seen the original paper. The original experiment⁶ was similar but involved blasts of air as punishment and the monkeys merely warning each other instead of attacking.



© iStockphoto.com/foryouinf

journals, none are prescriptive about the level of significance testing.

Are we collectively then like the monkeys in the stairs experiment? It is the classic example of mob mentality, maintained long after any reason for it has been lost (see above). Do we chose 5% as significant just because it has always been like that?

No one is stopping us from choosing other significance levels. We choose 5% because changing to any other level would require justification – which is illogical as there is no justification for 5% in the first place. Are any of us comfortable with a one in twenty chance of doing everything right and still getting it wrong? Should we be demanding more of the information we get from our experiments?

If we insist on sticking to 5% we might remember that the operators of the Large Hadron Collider declined to announce their finding of a Higgs Boson-like particle until they had reached five standard deviations worth of probability – which is a one in 3.5 million chance.

References

1. Fleming, I. (1959) Goldfinger.

2. Cowles M. and Davis, C. (1982) On the Origins of the .05 Level of Statistical Significance. *Am Psychol* **37**, 553–558.

3. Gerber, A. S. and Malhotra, N. (2008) Publication incentives and empirical research: do reporting standards distort the published results? *Socio Meth Res* **37**, 3–30.

4. Wood T. B. and Stratton, F. J. M. (1910) The interpretation of experimental results. *Journal of Agricultural Science*, **3**, 417–440.

5. www.ich.org/fileadmin/ Public_Web_Site/ICH_Products/ Guidelines/Efficacy/E9/Step4/E9_ Guideline.pdf

6. Stephenson, G. R. (1967) Cultural acquisition of a specific learned response among rhesus monkeys. In: Starek, D., Schneider, R. and Kuhn, H. J. (eds.), *Progress in Primatology*, Stuttgart: Fischer, pp. 279–288.

Dr Mark Kelly is a statistician working in the South East Wales Trials Unit in Cardiff University. His research interests include weight loss maintenance, early life influences on body mass index, and mental health, as well as statistical methodology including multilevel modelling, systematic reviews and meta-analysis.



Illustration: Andrew Tapsell (www.andrewtapsell.com)