UNDERSTANDING, TEACHING AND USING *p* VALUES

Geoff Cumming
School of Psychological Science, La Trobe University, Australia
g.cumming@latrobe.edu.au

There are many problems with the p *value. Is it an indicator of strength of evidence (Fisher), or only to be compared with* α *(Neyman-Pearson)? Many researchers and even statistics teachers have misconceptions about* p*, although* p *has been little studied, and we know little about how textbooks present it, and how researchers think about it, react to it, and use it in practice. The* p *value varies dramatically because of sampling variability, but textbooks do not mention this and researchers do not appreciate how widely it varies. I discuss the problems of* p *and advantages of confidence intervals, and identify research needed to guide the design of improved statistics education about* p*. I suggest the most promising teaching approach may be to focus throughout on estimation, use confidence intervals wherever possible, give* p *only a minor role, and explain* p *mainly as indicating where the confidence interval falls in relation to the null hypothesised value.*

Many disciplines rely on the *p* value to draw conclusions, yet *p* is often misunderstood and poorly used. It is at the heart of research, so it is surprising and disappointing how little it has been studied. We know little about how researchers think and feel about *p*, and little about how textbooks explain *p* and how that relates to what researchers do. The very large variation in *p* over replication is not widely appreciated, or mentioned in textbooks. I discuss these problems of *p*, and consider what statistical cognition and education research is needed to guide improved teaching about *p*. I conclude that estimation has many advantages over null hypothesis significance testing (NHST), and suggest a promising educational approach may be to focus on effect sizes and confidence intervals (CIs), give *p* only a minor role, emphasize the variability of *p*, and explain *p* mainly in terms of where the CI falls in relation to the null hypothesised value.

APPROACHES TO STATISTICAL INFERENCE

Fisher used .05 and .01 criteria but, more generally, used the *p* value to form an attitude to the null. “It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require…” (Fisher, 1966, p. 13). The *p* value is “a measure of the rational grounds for the disbelief it engenders” (Fisher, 1959, p. 43), and is thus an indicator of strength of evidence against the null. While vehemently rejecting Bayesian statistics, Fisher came close to using *p* to guide formation of, in effect, a subjective posterior probability about a null hypothesis.

Fisher was famously in conflict with Neyman and Pearson (N-P), whose approach was to specify a value of α in advance, and null and alternative hypotheses. Then *p* was compared with α, and the null rejected or not. This was a strict dichotomous choice based on whether $p < \alpha$, and otherwise the size of *p* was irrelevant. Paying any attention to the size of *p*, beyond whether or not it was less than α, invalidated the probabilistic basis of the N-P approach to inference.

Fisher considered the amount of evidence, whereas N-P offered objectivity and clear decisions, but seemed to ignore useful information, for example when *p* was much less than α. Which would science choose? We seem to want to both have our cake, and eat it too! Oakes (1986) spoke of “unsatisfactory conflation of the two approaches” (p. 96). Gigerenzer (1993) described “an incoherent mishmash” (p. 314). Hubbard (2004) claimed that “users of statistical techniques in the social and medical sciences are almost totally unaware of the distinctions” (p. 304) between the two. Given this confusion in what researchers do, what should teachers teach?

PROBLEMS WITH *p* VALUES IN PRACTICE

Statistical cognition is the important research field that studies how people understand statistical concepts and interpret statistical presentations (Beyth-Marom, Fidler, & Cumming, 2008). Examples are the studies of Oakes (1986), who found evidence many psychologists hold severe misconceptions about NHST and *p* values, and Haller and Krauss (2002), who found that even many teachers of statistics hold misconceptions. Kline (2004, Chapter 3) provided an

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

excellent summary of the problems of NHST and how it is used. He described five widespread but fallacious beliefs about *p*, and eight false conclusions often drawn after a null hypothesis test, based mainly on incorrect interpretations of *p*. Given the numerous problems with what researchers believe, and how they use *p* values in practice, what should teachers teach?

WHAT THOUGHTS AND FEELINGS DO *p* VALUES ELICIT?

Remarkably, this question has rarely been asked! There has been sadly little investigation of how researchers think about, and react to *p* values; statistical cognition research is required. Above I mentioned evidence of misconceptions about *p*. Beretvas and Robinson (2004) used focus groups to explore how professors of education thought about *p* values, and found basic confusions. Otherwise, close to the only cognitive study of *p* started with Rosenthal and Gaito (1963), who found a cliff effect, meaning researchers' belief that an effect is real made a jump at $p = .05$, consistent with an N-P approach. Poitevineau and Lecoutre (2001) found evidence of a cliff for some participants, but in others a Fisherian gradual increase in belief with decrease in *p*. Lai, Kalinowski, Fidler, and Cumming (2010) found evidence of various shapes of function between judgments an effect is real, and *p* values. Both N-P jumps and Fisherian slopes were common.

I suspect there is much more to learn about thoughts and feelings that *p* values elicit. For example, *p* values may often lead to strong emotions, perhaps of disappointment, frustration, relief, pleasure and exhilaration—as *p* ranges from large to small. Given such incomplete understanding of the cognition and affect of *p*, how can we design what to teach about *p* and how best to teach it?

REPLICATION AND *p* VALUES

Replication is at the heart of science, and the best strategy for overcoming sampling variability. Cumming, Williams, and Fidler (2004) explained that a 95% CI is, on average, an 83% prediction interval for the mean of a replication experiment. CIs thus give useful information about replication, and thinking about what is likely to happen on replication is one useful way to interpret any CI. But what about *p* values? If you repeat an experiment, what *p* value is likely? Cumming (2008) showed theoretically and by simulation that *p* varies enormously with replication.

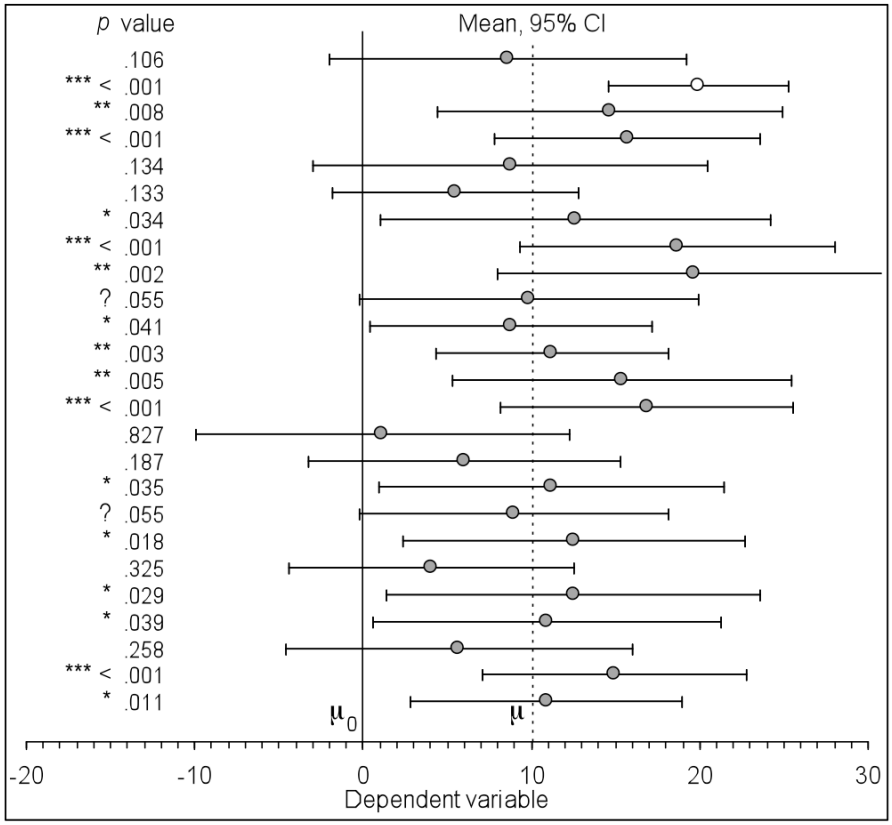
Figure 1 shows sample means, and 95% CIs, for 25 replications of a fictitious single-sample experiment. Sample size is $N = 18$. The normally distributed population has mean $\mu = 10$ and standard deviation $\sigma = 20$. The CIs bounce around the population mean. In the long run, 95% of intervals would include $\mu = 10$, and 5% would miss. (In Figure 1, just one interval misses; its mean is shown with an open circle.) The interval widths also vary, because intervals are calculated using sample SDs. The bouncing around over replication is familiar, and is illustrated in any textbook that gives a good explanation of what the 95% level of confidence means.

Figure 1 also reports two-tailed *p* values, for a zero null hypothesis. It may be surprising to many readers that *p* varies so dramatically, from < .001 to .827. Virtually any value seems possible! However, Cumming (2008) found this extent of variation in *p* is typical. Figure 1 thus shows variation in CIs, which is considerable but familiar, and variation in *p*, which is surprisingly large. The two forms of variation are closely tied, because of course each *p* is determined by where the CI falls in relation to $\mu_0 = 0$. If the interval includes zero, $p > .05$, and if not, $p < .05$. The further the interval falls from zero, the smaller is *p*. The closer the sample mean to zero, the larger is *p*. Given either CI or *p*, (and *N*, sample mean, and μ_0) we can calculate the other.

In Figure 1 the population effect size is $10/20 = 0.50$, a medium-sized effect. For $N = 18$, the power to find this effect, with two-tailed $\alpha = .05$ and σ unknown, is .52. Maxwell (2004, p. 148) cited evidence that median power of much social and behavioural sciences research to detect a medium-sized effect is around .5. The example of Figure 1 is thus typical of much published research, and illustrates the extent of *p* variation that researchers should expect with replication.

Any single CI gives some idea of the infinite sequence of possible results—25 of which are shown in Figure 1—because its width gives some idea of the extent of bouncing around. In stark contrast, any single *p* gives hardly any idea of the infinite sequence of *p* values. Cumming et al. (2004) found researchers have a reasonable understanding of how a CI can be used to forecast where means of replication experiments are likely to fall, even if they underestimate somewhat the variability of replication means. By contrast, Lai, Fidler, and Cumming (2009) found researchers

on average severely underestimated the variability of *p* with replication. Given these weaknesses of *p* in relation to replication, what should teachers teach?



The normally distributed population has mean $\mu = 10$ and standard deviation $\sigma = 20$. The null hypothesized value μ_0 is zero, so the population effect size is 10/20 = 0.50, a medium-sized effect. Means and 95% CIs, and two-tailed *p* values are shown, and marked with three, two, or one asterisks, using the conventional boundaries of .001, .01, and .05 respectively; '?' indicates $.05 < p < .10$. The only sample mean whose CI does not capture μ is shown with an open circle.

Figure 1. Twenty five replications of a single-sample experiment with sample size $N = 18$

WHAT DO STATISTICS TEXTBOOKS SAY ABOUT *p* VALUES?

I have been able to find very little systematic study of how textbooks present *p* values. Brewer (1985) found numerous errors in the ways five statistics textbooks widely used in the behavioural sciences described NHST. Huberty (1993) examined more than 50 statistics textbooks published between 1910 and 1992, with particular attention to the relative influence of Fisherian and N-P ideas. He identified a predominance of N-P, but many deficiencies in the presentations. Gliner, Leech, and Morgan (2002) examined 12 graduate-level textbooks widely used in education, and found most mentioned problems involving *p*, at least minimally, but few gave details or recommendations. Gliner et al. concluded that "Most disheartening was the failure of almost all of these recent texts to acknowledge that there is controversy surrounding NHST" (p. 90).

We use Aron, Aron, and Coups (2008) in our introductory statistics course for psychology and social work students. It introduces NHST by describing its "opposite-to-what-you-predict, roundabout reasoning… something like a double negative" (p. 147). It explains *p* correctly, then says a researcher selects in advance a significance criterion for rejecting the null hypothesis, with 5% and 1% identified as conventional levels. The obtained *p* is then compared against the chosen criterion. The alternative hypothesis, Type I and Type II errors and statistical power are covered. That all fits with N-P, even if the criterion is called a significance level, not α. There is also a description of how research is reported in journal articles, which gives a broader view. It notes that

$p < .001$ may be reported even if a 5% criterion had been chosen, that $p < .10$ may be described as "approaching significance", and that exact p values, such as $p = .03$ or $p = .27$, are often given—although no comment is made about how these might be interpreted.

Moore (2004) and Moore and McCabe (2006) are widely-used and highly-regarded textbooks that first define p correctly then describe it as a measure of strength of evidence against the null hypothesis. The final step of inference they describe is to compare p against a preselected α, referred to as the significance level. They thus combine elements of Fisher and N-P.

There is a strange anomaly about how sampling variability is presented to students. Every introductory textbook and course carefully explains sampling variability of the mean, sampling distributions, and standard error. Students see figures and perhaps simulations that illustrate the extent of sampling variability. Similarly, if CIs are covered there is probably an illustration of the sampling variability of intervals, as in Figure 1. In stark contrast, however, I know of no textbook that even mentions that p varies from sample to sample, let alone how greatly it varies. Instead, the focus is on p as an exact value to be compared against a sharp cutoff, in order to make a dichotomous decision to reject, or not reject, the null.

The American Psychological Association (APA) *Publication Manual* is very widely used across many disciplines. The sixth edition (APA, 2010) recommends "when reporting p values, report exact p values (e.g., $p = .031$)…. The tradition of reporting p values in the form $p < .10$, $p < .05$, $p < .01$… was appropriate in a time when only limited tables of critical values were available" (p. 114). This implies a Fisherian view, although there is no comment about how such exact p values should be interpreted. Cumming and Finch (2005), in the context of discussing CIs and their advantages, noted with approval that: "Reporting exact p values encourages a move from NHST as dichotomous decision making to the consideration of p values as useful input to interpretation" (p. 172)—again a Fisherian position.

However p values "can be highly misleading measures of the evidence… against the null hypothesis" (Berger & Sellke, 1987, p. 112). Hubbard and Lindsay (2008), amongst others, also made a detailed and strong case that p values are not a useful measure of evidence against the null. Even so, if p is to be used at all, 'strength of evidence' may be the least bad interpretation.

The signs summarised above support my observation that textbooks tend to take an N-P approach, with a requirement to specify α (or a significance criterion) in advance, with or without a nod in the direction of Fisher. However it seems researchers in journals usually take a largely Fisherian approach by reporting exact p, or giving as many asterisks as p permits—in either case p is serving as a measure of strength of evidence. There are exceptions and middle positions, but to some extent we seem to teach N-P but practise Fisher. This is a strange anomaly, and we need further evidence on the extent it is true. In the light of blindness to the variability of p, and seeming disagreement between textbooks and research practice, what should teachers teach?

WHAT SHOULD WE TEACH ABOUT *p* VALUES?

My first conclusion is the obvious one that we must start by choosing a goal. Should we teach N-P, Fisher, or some mixture—or something else entirely? A Bayesian would respond gleefully to the problems I have described by pronouncing NHST dysfunctional and overdue for replacement. Critics of the p value, for example Wagenmakers (2007), greatly outnumber its defenders, for example Dixon (1998), but anyone planning a textbook or course based on NHST—nearly everyone in the business—should explain their goal and rationale, and how they will counter the problems I have identified. Otherwise mere inertia and fashion are shaping statistics education.

My second conclusion is that, although introductory statistics is taken by millions of students every year, there are giant gaps in the evidence we need to design excellent statistics education. We have little idea how researchers respond to p values, and little idea what textbooks currently present and how that relates to what researchers actually do. There is evidence of NHST misconceptions held by researchers, students and even their teachers, but the thorough review by Sotos, Vanhoof, Van den Noortgate, and Onghena (2007) could find no study of, for example, how students interpret numerical values of p. It is very strange that p is the basis for inference and research decision making across many disciplines, but the cognition of p has hardly been studied. Interesting and important questions still await statistical cognition and statistical education research.

My third conclusion is to agree with Kline (2004, Chapter 3) that the problems are so severe we need to shift as much as possible from NHST. Bayesian and model-fitting techniques are attractive, but the first shift should be to estimation: Report and interpret effect sizes and CIs (Cumming & Fidler, in press), as the APA *Publication Manual* now recommends (APA, 2010). Fidler and Loftus (2009), and Coulson, Healey, Fidler, and Cumming (2009) reported evidence that reporting results as CIs can give better interpretation than reporting the same results using NHST.

I have for years taught an introductory course for first year psychology undergraduates that includes basic ideas of experimental design, descriptive statistics, sampling and CIs, and meta-analysis (Cumming, 2006) without any mention of NHST or *p* values. Data interpretation is based on figures with CIs, using the rules of eye described by Cumming and Finch (2005) but without reference to *p* values. Students use the ESCI software (www.latrobe.edu.au/psy/esci) to explore concepts and analyse data. I cannot report a controlled evaluation, but student ratings are high and informal feedback is good. Students encounter NHST and *p* values in their second semester.

Schmidt and Hunter (1997) made the bold claim that: "Any teacher of statistics knows that it is much easier for students to understand point estimates and CIs than significance testing with its strangely inverted logic" (p. 56). I know of no proper evaluation of this claim, but it raises the intriguing possibility that even if NHST is chosen as the primary approach to inference, it may still be most effective to teach estimation first, then *p* values. Teaching NHST after CIs is not novel—it seems common in some disciplines, although not in psychology. Moore (2004) and Moore and McCabe (2006) take this approach. They mention the link between CIs and *p*, then consider both estimation and NHST in subsequent chapters, with varying emphasis on one or the other.

I propose, however, a much stronger focus on estimation, which should be presented throughout as the approach to be used wherever possible. At least for some time students will need to know about NHST, if only to understand existing literature. I suggest *p* should be given only a marginal role, its problems explained, and it should be interpreted primarily as an indicator of where the 95% CI falls in relation to a null hypothesised value. Cumming (2007) illustrated that relationship, and suggested benchmarks that allow easy estimation of *p* from observing a CI and the null value. That seems to me a promising approach that may minimise the detrimental impact of the problems I described earlier. It requires educational development, then empirical evaluation.

ACKNOWLEDGEMENTS

This research was supported by the Australian Research Council. I thank Pav Kalinowski for valuable comments.

REFERENCES

American Psychological Association. (2010). *Publication manual of the APA* (6th ed.). Washington, DC: Author.

Aron, A., Aron, E. N., & Coups, E. J. (2008). *Statistics for the behavioural and social sciences: A brief course* (4th ed.). Upper Saddle River, NJ: Pearson Education.

Beretvas, S. N., & Robinson, D. H. (2004). How are effect sizes and *p*-values interpreted by professors of education? *Research in the Schools, 11,* 41-50.

Berger, J. O., & Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of *p* values and evidence. *Journal of the American Statistical Association, 82,* 112-122.

Beyth-Marom, R., Fidler, F., & Cumming, G. (2008). Statistical cognition: Towards evidence-based practice in statistics and statistics education. *Statistics Education Research Journal, 7,* 20-39.

Brewer, J. K. (1985). Behavioral statistics textbooks: Source of myths and misconceptions? *Journal of Educational Statistics, 10,* 252-268.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2009). *Confidence intervals permit, but don't guarantee, better inference than statistical significance testing*. (Submitted for publication.)

Cumming, G. (2006). Meta-analysis: Pictures that explain how experimental findings can be integrated. In A. Rossman & B. Chance (Eds.), *ICOTS-7 Proceedings*. Online: www.stat.auckland.ac.nz/~iase/publications/17/C105.pdf.

Cumming, G. (2007). Inference by eye: Pictures of confidence intervals and thinking about levels of confidence. *Teaching Statistics, 29,* 89-93.

Cumming, G. (2008). Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*, 286-300.

Cumming, G., & Fidler, F. (in press). From hypothesis testing to parameter estimation: An example of evidence-based practice in statistics. In A. T. Panter & S. Sterba (Eds.), *Handbook of Ethics in Quantitative Methodology*. London: Taylor and Francis.

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals, and how to read pictures of data. *American Psychologist, 60*, 170-180. Online: www.apastyle.org/manual/related/cumming-and-finch.pdf.

Cumming, G., Williams, J., & Fidler, F. (2004). Replication, and researchers' understanding of confidence intervals and standard error bars. *Understanding Statistics, 3*, 299-311.

Dixon, P. (1998). Why scientists value *p* values. *Psychonomic Bulletin & Review, 5*, 390-396.

Fidler, F., & Loftus, G. (2009). Why figures with error bars should replace *p* values: Some conceptual arguments and empirical demonstrations. *Zeitschrift für Psychologie, 217*, 27-37.

Fisher, R. A. (1959). *Statistical methods and scientific inference* (2nd ed.). Edinburgh, UK: Oliver and Boyd.

Fisher, R. A. (1966). *The design of experiments* (8th ed.). Edinburgh: Oliver & Boyd.

Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Keren & C. Lewis, (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Hillsdale, NJ: Erlbaum.

Gliner, J. A., Leech, N. L., & Morgan, G. A. (2002). Problems with null hypothesis significance testing (NHST):What do the textbooks say? *Journal of Experimental Education, 71*, 83-92.

Haller, H., & Krauss, S. (2002). Misinterpretations of significance: A problem students share with their teachers? *Methods of Psychological Research, 7*, 1-20.

Hubbard, R. (2004). Alphabet soup. Blurring the distinction between *p*'s and α's in psychological research. *Theory & Psychology, 14*, 295-327.

Hubbard, R., & Lindsay, R. M. (2008). Why *p* values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology, 18*, 69-88.

Huberty, C. J. (1993). Historical origins of statistical testing practices: The treatment of Fisher versus Neyman-Pearson views in textbooks. *Journal of Experimental Education, 61*, 317-333.

Kline, R. B. (2004). *Beyond significance testing. Reforming data analysis methods in behavioral research*. Washington, DC: APA Books. Chapter 3. Online: www.apastyle.org/manual/related/kline-2004.pdf.

Lai, J., Fidler, F., & Cumming, G. (2009). *Subjective* p *intervals: Researchers underestimate the variability of* p *values over replication*. Manuscript submitted for publication.

Lai, J., Kalinowski, P., Fidler, F., & Cumming, G. (2010). *Dichotomous thinking: A problem beyond NHST*. Paper presented at ICOTS-8, Slovenia, July.

Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods, 9*, 147-163.

Moore, D. S. (2004). *The basic practice of statistics* (3rd ed.). New York: W. H. Freeman.

Moore, D. S., & McCabe, G. P. (2006). *Introduction to the practice of statistics* (5th ed.). New York: W. H. Freeman.

Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences.* Chichester: Wiley.

Poitevineau, J., & Lecoutre, B. (2001). Interpretation of significance levels by psychological researchers: The .05 cliff effect may be overstated. *Psychonomic Bulletin & Review, 8*, 847-850.

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

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in analysis of research data. In L. Harlow, S. Mulaik, & J. Steiger (Eds.). *What if there were no significance tests?* (pp. 37-63). Mahwah, NJ: Erlbaum.

Sotos, A. E. C., Vanhoof, S., Van den Noortgate, W., & Onghena, P. (2007). Students' misconceptions of statistical inference: A review of the empirical evidence from research on statistics education. *Educational Research Review, 2*, 98-113.

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