The Psychological Reality of the Learned "$p < 0.05$" Barrier

V.N. Vimal Rao, Jeffrey K. Bye, and Sashank Varma
University of Minnesota

Abstract

Traditional statistics instruction emphasizes a 0.05 significance level for inferential tests. Here, we investigate the consequences of this training for researchers' mental representations of probabilities – whether 0.05 becomes a boundary, i.e., a discontinuity of the mental number line. 25 graduate students with statistical training viewed pairs of p-values and judged whether they were "similar" or "different". As predicted, participants were more likely and faster to judge p-values as "different" when they crossed the 0.05 boundary (e.g., 0.049 – 0.051) compared to when they did not (e.g., 0.019 – 0.021). This categorical perception effect suggests that traditional statistical instruction creates a psychologically real chasm between significant and non-significant p-values. This distortion is undesirable given modern approaches to statistical reasoning that de-emphasize dichotomizing p-values.

Keywords: statistics education; statistical significance; mathematics education; rational number processing; categorical perception.

Introduction

The phenomenon of *p*-hacking has spurred debate among researchers (Wasserstein et al., 2019). Beyond these methodological concerns, we consider here the question of whether instruction and practice emphasizing a *statistical boundary* at 0.05 resulted in a *mental boundary* in people's understandings of probabilities. Cognitive science research has investigated how people use categories to divide continuous cognitive representations in the service of learning and communication (Bruner et al., 1956; Gibson, 1969). These categories have consequences. In particular, the *categorical perception effect* (CPE; Harnad, 1987) is the distortion in the way people perceive exemplars on the same vs. different sides of a category boundary (Fleming et al., 2013; Notmon et al., 2005).

Given the predominance of the 0.05 boundary in science and calls for its reform, it is important to understand whether researchers show a CPE on the *p*-value continuum. The presence of distortions around the significance cutoff could affect researchers' statistical interpretations and decisions. The potential existence of such distortions would spur new research on instruction and practice of statistics. In this study, we explore whether a CPE for *p*-values exists in individuals with statistical training.

Background

The use of *p*-values and null hypothesis significance testing (NHST) traces back to the early 20th century (Biau et al., 2010; Gigerenzer, 2004). To identify sufficiently small *p*-values, 0.05 was taken to be a convenient limit to identity a significant deviation from the hypothesized mean (Fisher, 1925). Neyman and Pearson (1933) suggested a method by which in the "long run of experience, we shall not be too often wrong" (p. 291). Together, these practices combined to form the modern practice of NHST, where a hypothesis is rejected if and only if a *p*-value is less than 0.05. In this way, the artificial boundary of 0.05 became a gatekeeper to publication (Tramifow, 2014; Stang et al., 2010).

Methodological critiques of NHST have focused on the logic of the approach. Here, we consider the cognitive consequences of conceptualizing 0.05 as the boundary between "significant" and "null" results. We begin with categorization, which is fundamental to human inference and perception (Bruner et al., 1956). CPEs alter perception such that two values on the same side of a boundary may appear more similar than two values on opposite sides of a boundary, even when the physical or numerical distance between the pairs is the same (Goldstone, 1994; Harnad, 2017). This phantom discontinuity in perception can lead reasoners to make suboptimal decisions in subsequent judgements based on these implicit perceptions (Fleming et al., 2013; Notmon et al., 2005).

To understand how a CPE may warp the mental representation of *p*-values, it is important to first know that the mental representation of numbers in general is continuous. Natural numbers are mapped to points on a logarithmically compressed mental number line (Dehaene et al., 1990). This is also true of rational numbers expressed as decimals (Varma & Karl, 2013). The evidence for a non-linear mental number line comes in part from experiments where people are asked to identify the greater (or lesser) of a pair of numbers. People make faster judgments when the numbers are small versus large (e.g., 1 v 2 is faster than 8 v 9; LeFevre et al., 1996); this is the *size* effect. They make faster judgments when the distance between numbers is far vs. near (e.g., 2 v 8 is faster than 3 v 5; Cohen, 2010); this is the *distance* effect. Finally, they make faster judgments for pairs of multi-digit numbers when the digit in each place of the larger number is greater than its counterpart in the smaller number (e.g., 35 v 46 is faster than 36 v 45; Nuerk et al., 2001); this is the *compatibility* effect. There are also discontinuities of the mental number line caused by the place-value symbol system for naming numbers. For example, determining the midpoint between two numbers is slower and less accurate when the tens digits differ (e.g., bisecting 27 – 35 is harder than 21 – 29; Nuerk et al., 2011); this is the *decade crossing* effect.

The question we consider here is whether traditional statistics instruction results in a discontinuity in the mental number line at 0.05. If this is ***not*** the case, participants should perceive $p = 0.049$ and $p = 0.051$ to be *more similar* than $p = 0.019$ and $p = 0.021$ because of the size effect; than $p = 0.028$ and $p = 0.033$ because of the size and distance effect; and than $p = 0.012$ and $p = 0.023$ because of the size, distance, and compatibility effects. However, if a CPE exists, individuals may perceive $p = 0.049$ and $p = 0.051$ to be *more different* than any of the other pairs because only in this case do the two *p*-values cross the putative 0.05 boundary. Thus, the goal of this study is to identify whether a CPE effect exists in the perception of *p*-values.

Methods

We recruited graduate students associated with an Educational Psychology department at a public university in the midwestern US. Eligible participants where those who reported having experience in their own research or teaching with hypothesis tests and *p*-values. Of 40 initial respondents, 25 were deemed eligible and completed the study in full. Participants were recruited on a voluntary basis and not compensated for completing the study.

Each stimulus was a pair of *p*-values. The within-pair distance between the two values ranged from 0.002 to 0.009 in thousandths increments. Between 10 to 12 stimuli pairs were created for each within-pair distance. The critical variable was boundary-crossing. Of these 10-12 pairs, three or four crossed the 0.05 boundary, three or four were below the boundary, and three or four were above the boundary. All pairs were of the form “0.0XY” where X and Y were non-zero digits to control for a potential effect of a different number of leading zeros on reaction time (Schulze et al., 1991). Each pair of *p*-values had different digits in the hundredths place. We included additional filler stimuli so that participants would not notice and begin attending to the uneven distribution of digits in the stimulus set.

Pairs of *p*-values were presented sequentially to participants, with the first *p*-value shown for one second and the second shown until either a response was made or five seconds elapsed. Participants were asked to judge whether the two *p*-values were “similar” to or “different” from each other, and to indicate their choice as quickly as possible by pressing either F (for similar) or J (for different) on their keyboard. This format is known as an AX categorical discrimination task and is widely used in CPE studies (e.g., Hary & Massaro, 1982). Of the 10-12 trials for each distance, in one half the first *p*-value presented was smaller than the second one, and in the other half it was larger. To induce a statistical mindset, *p*-values were presented in the form “$p = 0.036$”. Stimuli were blocked into six sets of 18 pairs, with a 20-second break

between each block. An additional six stimuli were presented in an initial training phase to familiarize participants with the task procedures.

We looked for categorical perception in two ways. First, we investigated whether similar vs. different judgments (for which there is no objectively correct response) changed as a function of whether the *p*-values crossed the 0.05 boundary with a log-binomial mixed effects model (e.g., Cudeck, 1996; Marschner & Gillet, 2012). To isolate the effect of 0.05 boundary-crossing, the model adjusted for size of the *p*-values, the distance between them, the pair's digit compatibility, and whether the smaller *p*-value was presented first.

Second, we looked for differences in participants' reaction times as a function of crossing the 0.05 boundary using a linear mixed effects model, which analyzed the relationship between reaction time and whether the stimuli crossed the boundary, and again adjusted for the size, distance, compatibility, and order.

Both mixed effects models converged without singularities, so ANOVA Type III Sums of Squares tests were used to generate *p*-values for factors (Matuschek et al., 2017). All analyses were completed using R (R Core Team, 2019) and the lme4 package (Bates et al., 2015).

Results

Judgments of similarity vs. difference showed the predicted categorical perception effect. For a given distance, participants were more likely to label the pair as different when the *p*-values crossed the boundary ($p < 0.0001$), as shown in Figure 1. The point estimate was 2.42 (*95% CI: 1.49 – 3.95*), implying that for a given distance between stimuli, participants were approximately two and a half times more likely to indicate that the pair were different from each other when the *p*-values crossed the boundary (compared to both being below or both above). Among the other variables, participants were more likely to indicate that a stimulus pair was

different as the distance between stimuli in the pair increased ($p < 0.0001$), analogous to the distance effect. No other covariates were significant.

Analysis of participants' reaction times indicated large individual differences that appeared to obscure group differences. Nevertheless, after controlling for covariates, participants may have been slightly faster at identifying pairs of stimuli as different when they crossed the 0.05 boundary ($p = 0.0825$). On average, participants were approximately 0.282 seconds faster (*95% CI: -.048s – .612s faster*) in such cases, an estimated 24% reduction in reaction time against the model-adjusted mean. As expected, reaction times were faster as the distance between the pairs of stimuli increased ($p = 0.0043$). No other covariates were significant.

Discussion

This study investigated whether there is a CPE on the mental representation of the *p*-value continuum, with traditional statistical training resulting in a boundary at 0.05. In fact, this appears to be the case. Participants were more likely to judge a pair of *p*-values as 'different' (vs. 'similar') when they crossed the 0.05 boundary. They may also be slightly faster to make this judgment. These effects were present even after controlling for the size, distance, compatibility, and order effects in the mathematical cognition literature.

The finding of a CPE shows that the *p*-value continuum contains a discontinuity at 0.05. This is a discontinuity in participants' perception of *p*-values beyond a simple application of a label that fundamentally affects their statistical reasoning and decision making about *p*-values. We hypothesize that this is a consequence of traditional statistics instruction, and with reading a scientific literature still dominated by NHST. We are in the process of collecting additional data to more strongly test this learning hypothesis. Specifically, if this instructional and experiential hypothesis is true, then the CPE for *p*-values should be relatively small for first-year graduate

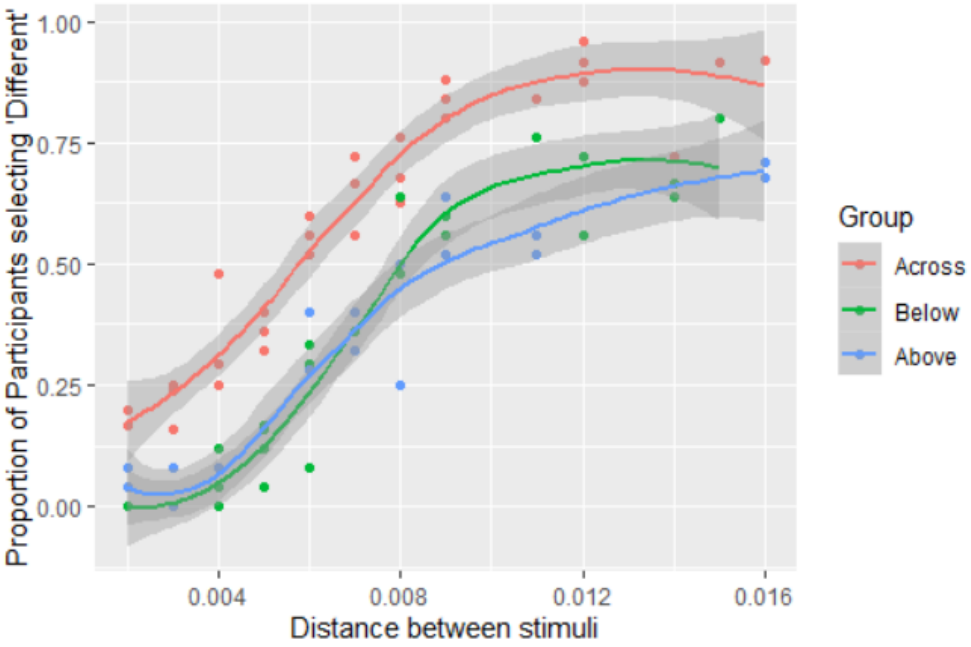
students, should increase over graduate training and statistical coursework, and should be relatively large in working scientists.

This study had several limitations. First, it included only 25 participants who made only 108 judgments. Thus, the data were noisy, and the reaction time data exhibited unexpected patterns most likely explained by measurement variability occluding a possible CPE. Second, the CPE found here might be paradigm-specific, and further studies are needed showing the 0.05 remains a boundary across a range of discrimination tasks (Gerrits & Schouten, 2004). Nevertheless, the current study provides a first insight into the cognitive models underlying interpretation of *p*-values, and sets the stage for future research.

In recent years, statisticians have argued that comparisons of statistical significance and non-statistical significance should not be made (Wasserstein et al., 2019). The advice is meant to obviate the phenomenon of *p*-hacking, wherein researchers make self-serving decisions to produce attractive *p*-values that may increase their odds of publishing (Simonsohn et al., 2014). However, such dichotomous categorizations can be helpful to novice learners (Gibson, 1969), and some statisticians have pushed back on the abandonment of statistical significance in the classroom for this reason (e.g., Krueger & Heck, 2019). Our results show that CPEs for *p*-values may exist. However, CPEs that result from early categorizations are not permanent – categorizations and perceptions can change (Goldstone, 1994). Thus, a promising direction for future research would be to develop classroom activities and run an instructional study showing the NHST can be taught to students without distorting their perceptions of the probability of theories given evidence. This would be an important step towards a world beyond "$p < 0.05$".

Figures

Figure 1. Unadjusted mean propensity to identify a pair as different



References

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., ... & Bolker, M. B. (2015). Package 'lme4'. *Convergence, 12*(1), 2.

Biau, D. J., Jolles, B. M., & Porcher, R. (2010). P value and the theory of hypothesis testing: an explanation for new researchers. *Clinical Orthopaedics and Related Research, 468*(3), 885-892.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*. New York: Wiley.

Cohen, D. J. (2010). Evidence for direct retrieval of relative quantity information in a quantity judgment task: Decimals, integers, and the role of physical similarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(6), 1389.

Cudeck, R. (1996). Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate behavioral research, 31*(3), 371-403.

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of experimental Psychology: Human Perception and performance, 16*(3), 626.

Fisher, R.A. (1925). *Statistical methods for research workers.* London: Oliver and Boyd.

Fleming, S. M., Maloney, L. T., & Daw, N. D. (2013). The irrationality of categorical perception. *Journal of Neuroscience, 33*(49), 19060-19070.

Gerrits, E., & Schouten, M. E. H. (2004). Categorical perception depends on the discrimination task. *Perception & psychophysics, 66*(3), 363-376.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587-606.

Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General, 123*(2), 178.

Harnad, S. (1987) Psychophysical and cognitive aspects of categorical perception: A critical overview. In Harnad, S. (ed.) (1987) *Categorical Perception: The Groundwork of Cognition*. New York: Cambridge University Press.

Harnad, S. (2017). To cognize is to categorize: Cognition is categorization. In *Handbook of categorization in cognitive science* (pp. 21-54). Elsevier.

Hary, J. M., & Massaro, D. W. (1982). Categorical results do not imply categorical perception. *Perception & Psychophysics, 32*(5), 409-418.

Krueger, J. I., & Heck, P. R. (2019). Putting the p-value in its place. *The American Statistician, 73*(S1), 122-128.

LeFevre, J. A., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(1), 216.

Marschner, I. C., & Gillett, A. C. (2012). Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics, 13*(1), 179-192.

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94,* 305-315.

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character, 231*(694-706), 289-337.

Notman, L. A., Sowden, P. T., & Özgen, E. (2005). The nature of learned categorical perception effects: a psychophysical approach. *Cognition, 95*(2), B1-B14.

Nuerk, H. C., Moeller, K., Klein, E., Willmes, K., & Fischer, M. H. (2011). Extending the mental number line: A review of multi-digit number processing. *Journal of Psychology, 219*(1), 3.

Nuerk, H. C., Weger, U., & Willmes, K. (2001). Decade breaks in the mental number line? Putting the tens and units back in different bins. *Cognition, 82*(1), B25-B33.

R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schulze, K. G., Schmidt-Nielsen, A., & Achille, L. B. (1991). Comparing three numbers: The effect of number of digits, range, and leading zeros. *Bulletin of the Psychonomic Society, 29*(4), 361-364.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: a key to the file-drawer. *Journal of experimental psychology: General, 143*(2), 534.

Stang, A., Poole, C., & Kuss, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. *European journal of epidemiology, 25*(4), 225-230.

Trafimow, D. (2014). Editorial. *Basic and Applied Social Psychology, 36*(1), 1-2.

Varma, S., & Karl, S. R. (2013). Understanding decimal proportions: Discrete representations, parallel access, and privileged processing of zero. *Cognitive Psychology, 66*(3), 283-301.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a World Beyond “p< 0.05”. *The American Statistician, 73* (S1), 1-19.