frontiers in Psychology

GENERAL COMMENTARY
published: 01 April 2015
doi: 10.3389/fpsyg.2015.00341

P-values as percentiles. Commentary on: "Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations"

Jose D. Perezgonzalez *

Business School, Massey University, Palmerston North, New Zealand

Keywords: *p*-value, probability, percentile, statistical misinterpretations

A commentary on:

Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations
by Schneider, J. W. (2015). Scientometrics 102, 411–432. doi: 10.1007/s11192-014-1251-5

Edited by:
Lynne D. Roberts, Curtin University, Australia

Reviewed by:
Patrizio E. Tressoldi, Università di Padova, Italy

***Correspondence:**
Jose D. Perezgonzalez, j.d.perezgonzalez@massey.ac.nz

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *02 March 2015*
Accepted: *10 March 2015*
Published: *01 April 2015*

Citation:
Perezgonzalez JD (2015) P-values as percentiles. Commentary on: "Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations". Front. Psychol. 6:341. doi: 10.3389/fpsyg.2015.00341

Schneider's (2015) article is contemporary work addressing the shortcomings of null hypothesis significance testing (NHST). It summarizes previous work on the topic and provides original examples illustrating NHST-induced confusions in scientometrics. Among the confusions cited are those associated with the interpretation of *p*-values, old misinterpretations already investigated by Oakes (1986), Falk and Greenbaum (1995); Haller and Krauss (2000), and Perezgonzalez (2014a), and discussed in, for example, Carver (1978); Nickerson (2000), Hubbard and Bayarri (2003); Kline (2004), and Goodman (2008). That they are still relevant in recent times testifies to the fact that the lessons of the past have not been learnt.

As the title anticipates, there is a twist to this saga, a pedagogical one: *p*-values are typically taught and presented as probabilities, and this may be the cause behind the confusions. A change in the heuristic we use for teaching and interpreting the meaning of *p*-values may be all we need to start working the path toward clarification and understanding.

In this article I will illustrate the differences in interpretation that a percentile heuristic and a probability one make. As guiding example, I will use a one-tailed *p*-value in a normal distribution—$z = -1.75$, $p = 0.04$; **Figure 1**). The default testing approach will be Fisher's tests of significance, but Neyman–Pearson's tests of acceptance approach will be assumed when discussing Type I errors and alternative hypotheses (for more information about those approaches see Perezgonzalez, 2014b, 2015). The scenario is the scoring of a sample of suspected schizophrenics on a validated psychological normality scale. The hypothesis tested (Fisher's H_0, Neyman–Pearson's H_M) is that the mean score of the sample on the normality scale does not differ from that of the normal population (no H_0 = the sample does not score as normal; H_A = the sample scores as schizophrenic, assuming previous knowledge that schizophrenics score low on the scale, by a given effect size). Neither a level of significance nor a rejection region is needed for the discussion.

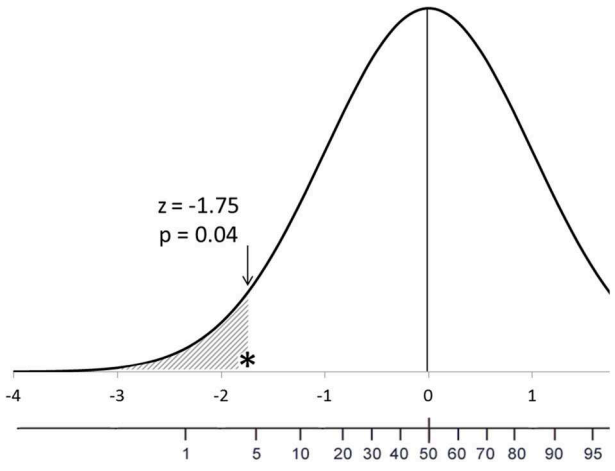

FIGURE 1 | Location of an observed *z*-score and its corresponding *p*-value in the frequency distribution of the hypothesis under test. The accompanying scales are for the theoretical *z*-scores and percentiles, respectively.

P-Values: Probabilities or Percentiles?

Let's start by establishing that *p*-values can be interpreted as probabilities. That is, when hypothetical population distributions are generated from sampling data, those frequency distributions follow the frequentist approach and the associated *p*-values show the appropriate probabilities. This is so because these *p*-values are theoretical—they represent the probability of, for example, a hypothetical human being alive today.

The *p*-value we obtain from our research data, however, is not a theoretical, probabilistic, value, but an observed one: its probability of occurrence is "1," precisely because it has occurred—it represents, for example, the realization that I am alive, not the probability of me being so. Therefore, the observed *p*-value does not represent a probability but a location in the distribution of reference. Among measures of location, percentiles (i.e., percentile ranks) are good heuristics to represent what observed *p*-values really are.

P-Values' Correct and Incorrect Misinterpretations

As **Figure 1** shows, a percentile describes a fact: the sample scored in the 4th percentile. As a probability, however, the *p*-value is often misinterpreted as, the observed result has a 4% likelihood of having occurred by chance—the *odds-against-chance fantasy* (Carver, 1978)—which also elicits a further misinterpretation as, the observed result has a 96% likelihood of being a real effect (Kline, 2004).

The percentile heuristic also conveys the correct interpretation of the *p*-value as a cumulative percentage in the tail of the distribution: 4% of normal people will score this low or lower. As a probability, the *p*-value is often misinterpreted as, the sample has only a 4% likelihood of being normal—the *inverse probability error* (Cohen, 1994).

Consequently, because the percentile only provides information about location in the distribution of the normal scores hypothesis, it is impossible to know the probability of making a mistake if this hypothesis is rejected. As a probability, the *p*-value is often misinterpreted as, there is only a 4% likelihood of making a mistake when rejecting the tested hypothesis. This is further confused as, the probability of making a Type I error in the long run (alpha, α) is 4%; which then leads to the belief that α can be adjusted a posteriori—roving α (Goodman, 1993)—as a lower than anticipated Type I error (Kline, 2004; Perezgonzalez, 2015).

Furthermore, the percentile is circumscribed to its hypothesis of reference—normal scores on the normality test—and makes no concession for non-tested hypotheses. As a probability, the *p*-value is often misinterpreted as, there is a 96% likelihood that the sample scored as not normal—Fisher's negation of H_0, the *valid research hypothesis fantasy* (Carver, 1978)—or scored as schizophrenic—Neyman–Pearson's H_A, the *validity fallacy* (Mulaik et al., 1997).

Finally, the percentile heuristic helps ameliorate misinterpretations regarding future replicability, if only because we normally have enough experience with percentiles in other spheres of life as to realize that the big fish in this pond is neither necessarily big all the time nor equally big in all ponds. As a probability, the *p*-value is often misinterpreted as, there is a 96% likelihood that similar samples will score this low in future studies—the *replicability or reliability fallacy* (Carver, 1978).

Conclusions

The percentile heuristic is a more accurate model both for interpreting observed *p*-values and for preventing probabilistic misunderstandings. The percentile heuristic may also prove to be a better starting point for demystifying related statistical issues—such as the relationship among *p*-value, effect size and sample size—and epistemological issues—such as statistical significance, and the proving and disproving of hypotheses. All in all, the percentile heuristic matters for better statistical literacy and better research competence, allows for clearer understanding without imposing unnecessary cognitive workload, and has a positive effect in fostering the teaching and practice of psychological science.

References

Carver, R. P. (1978). The case against statistical significance testing. *Harv. Educ. Rev.* 48, 378–399.

Cohen, J. (1994). The earth is round (p <.05). *Am. Psychol.* 49, 997–1003. doi: 10.1037/0003-066X.49.12.997

Falk, R., and Greenbaum, C. W. (1995). Significance tests die hard: the amazing persistence of a probabilistic misconception. *Theor. Psychol.* 5, 75–98. doi: 10.1177/0959354395051004

Goodman, S. N. (1993). *P* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. *Am. J. Epidemiol.* 137, 485–496.

Goodman, S. N. (2008). A dirty dozen: twelve P-value misconceptions. *Semin. Hematol.* 130, 995–1004. doi: 10.1053/j.seminhematol.2008.04.003

Haller, H., and Krauss, S. (2000). Misinterpretations of significance. A problem students share with their teachers. *MPR Online* 7, 1–20.

Hubbard, R., and Bayarri, M. J. (2003). Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing. *Am. Stat.* 57, 171–178. doi: 10.1198/0003130031856

Kline, R. B. (2004). *Beyond Significance Testing. Reforming Data Analysis Methods in Behavioral Research.* Washington, DC: APA. doi: 10.1037/10693-000

Mulaik, S. A., Raju, N. S., and Harshman, R. (1997). "There is a time and place for significance testing," in *What If There were No Significance Tests?* eds L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Mahwah, NJ: Erlbaum), 65–116.

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychol. Methods* 5, 241–301. doi: 10.1037/1082-989X.5.2.241

Oakes, M. (1986). *Statistical Inference: A Commentary for the Social and Behavioral Sciences*. Chichester: John Wiley & Sons.

Perezgonzalez, J. D. (2014a). Misinterpretation of the *p*-value and of the level of significance (2000). *Knowledge* 2014, 24–25. doi: 10.6084/m9.figshare.1320686

Perezgonzalez, J. D. (2014b). A reconceptualization of significance testing. *Theor. Psychol.* 24, 852–859. doi: 10.1177/0959354314546157

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Front. Psychol.* 6:223. doi: 10.3389/fpsyg.2015.00223

Schneider, J. W. (2015). Null hypothesis significance tests. A mix–up of two different theories: the basis for widespread confusion and numerous misinterpretations. *Scientometrics* 102, 411–432. doi: 10.1007/s11192-014-1251-5

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.