Mind Your p's and Alphas

WILLIAM M. STALLINGS
Georgia State University

ABSTRACT: In the educational research literature, alpha and p *are often conflated. Paradoxically, alpha retains a prominent place in textbook discussions of such topics as statistical hypothesis testing, multivariate analysis, power, and multiple comparisons, whereas it seems to have been supplanted by* p *in the results sections of journal articles. The unique contributions of both alpha and* p *are discussed and a plea is made for using both conventions in summarizing the outcomes of tests of significance.*

When teaching introductory statistics or research methods, I find that my students invariably have difficulty distinguishing alpha from *p*. I sympathize with them, for there is indeed dissonance between what is presented in the textbooks and what is found in the "results" sections of journal articles.

Introductory texts tend to describe alpha as a "ground rule" needed to be set by the researcher *before* gathering data or, in another conceptualization, before calculating the test statistic. Typically, *p* receives little attention, except when it is mechanically compared with alpha for determining whether to reject the null hypothesis. In empirical research the relative importance of alpha and *p* seems to be reversed; more emphasis is placed on such summaries as $t(22) = 2.62$, $p < .01$ than on verbal statements to the effect that the null hypothesis was rejected at the .01 level of significance. Do the two conventions, alpha and *p*, convey the same information and serve identical functions? I will argue that the answer is "No."

William M. Stallings is Professor of Educational Foundations at Georgia State University, University Plaza, Atlanta, GA 30303. His specializations are applied statistics and evaluation.

An earlier version of this paper was presented at the 1985 AERA Annual Meeting, Chicago.

The author wishes to thank William J. Russell, William D. Schafer, William L. Curlette, and two anonymous reviewers for their helpful comments and suggestions.

Let's begin with basics. Alpha and *p* are both probabilities. Alpha, however, is a constant, whereas *p* may be construed as a random variable. Unlike alpha, *p* is sensitive to sample size. Historically, alpha is associated with the Neyman-Pearson theory of hypothesis testing and *p* with the Fisherian theory of significance testing (see Huberty, 1985). Finally, alpha is set during the design phase of a study; *p* is used to interpret the study's outcome.

As presented in our textbooks, applied inferential statistics is a mixture of the once opposing positions of Fisher and Neyman-Pearson (Huberty, 1985; Stallings & Singhal, 1969). It is little wonder that *p*-values sometimes seem to be confused with alphas by professionals as well as by students. Does the most recent edition of the *Publication Manual* of the American Psychological Association (1983) reflect this confusion when it offers as illustrations $p < .05$ and $p < .01$? In the empirical literature, as well as in the *Publication Manual*, *p*'s are often reported in values associated with alpha (e.g., .05 and .01). There are at least three explanations for this: (a) statistical tables (such as those for F) often give the critical values associated with .05 and .01 probabilities; (b) "exact" probabilities are obtainable only if the null hypothesis is true *and* if the mathematical model is correct, thus, the researcher may feel that fallible data do not warrant greater precision in reporting *p*-values; and (c) the conflation between alpha and *p* may be historical. Fisher (1947) seems to have used level of significance as a synonym for *p* and *standard* level of significance for what we call alpha. Of course, standard level of significance and alpha may be, in Thomas Kuhn's conception, "incommensurable."

In current statistical practice, alpha needs to be specified for estimating a priori power, determining sample size, and evaluating multiple comparison techniques. Standard experimental design texts (Keppel, 1982; Kirk, 1982) urge an even greater attention to alpha by recommending the differential allocation of an overall alpha via the Bonferroni inequality. Alpha and beta together are required when operating in a decision theoretic mode and when estimates of the costs associated with the two types of error are available. Clearly, alpha is "still alive and kicking," at least during the design phase.

What do *p*'s do for us? They permit the reader to impose his or her own alpha level on reported findings (Hopkins & Glass, 1978). Parenthetically, a colleague categorized this approach as "Back Door Bayesian." Further, *p*-values are needed in combining probabilities and in conducting some forms of meta-analysis. Some authorities interpret small *p*-values as providing degrees of evidence against the null. For example, Hopkins and Glass (1978) contend that even after a rejection at the .05 level of significance, a *p* value of, say, .01 gives "greater assurance than the .05 level that Ho is false" (p. 223). Similar, but more dramatic, is Lehmann's (1968) characterization that *p* is an "index of surprise." As Lehmann elaborates, "the smaller $\hat{\alpha}$ [our p] is, the more surprising it is to get this extreme a value under H [the null] and, therefore, the stronger the evidence against H" (p. 43).

I have argued that (a) alpha and p are conceptually distinct, (b) the two conventions can serve different functions, and (c) alpha and p sometimes seem conflated. The separate nature and particular virtues of alpha and p can be maintained and the confusion eliminated by reporting more exact values for p, together with explicit alpha values.

Consider three arguments for more exact p-values coupled with stated alphas. First, the researcher can adjudicate in such cases as $\hat{\alpha} = .05$ and $p = .06$. Both Tatsuoka (1982) and Kempthorne (1972) seem to suggest that such a finding might be evaluated as a statistically significant one. Second, the reader would *not* have to assume, as did one anonymous referee of this paper, "by convention, that unless otherwise told, the author has specified $\hat{\alpha} = .05$." After all, as Huberty (1985) observes, a large alpha might be used for an exploratory study and a very small alpha for a more thoroughly investigated topic. Finally, if an explicit alpha were given, the researcher would be free to use an "exact" p-value. Otherwise, uses like $p < .05$ seem to show p-values playing an additional role of the alpha level.

Consider three calls for reporting more exact p-values. Huberty (1985) recommends that "when combining results of independent studies using Fisher's method of adding logarithms of P-values or Edgington's method of adding P-values. . . , three decimal places seem to be the minimum needed" (p. 9). Gibbons and Pratt (1975) argue that "reporting a P-value, *whether exact or within an interval* [italics added], in effect permits each individual to choose his own level of significance as the maximum tolerable probability of a Type I error" (p. 21). Again, a similar position is articulated by McNeil, Kelly, and McNeil (1975):

> Now if the author reports the actual probability of .007, the conservative reader knows that the research hypothesis is tenable for him as well as for the author. On the other hand, if the actual probability is reported as .030, then the research hypothesis is *not* tenable for the conservative reader, even though he realizes that the research hypothesis is tenable for the author. (pp. 191–192)

This brings us to the issue of how more exact p-values should be reported. I recommend "exact" p's (e.g., $p = .031$) or probability intervals (e.g., $.04 > p > .03$). Each of these seems consistent with Gibbons and Pratt's (1975) definition of a p-value as "the smallest level at which the observations are significant in a particular direction" (p. 20).

In summary, the "results" section of an empirical article will be more informative if it reports that $p = .022$ *and* that alpha was set at .05. Perhaps we should adopt a new convention. A first approximation might be one illustrated by $p = .022 < .05$ where the researcher's alpha level would follow the inequality sign. If the test were two-sided and the null distribution symmetrical, we would divide the alpha value by two (e.g., $p = .022 < .05/2$) or we could double the obtained value of p and add the words "two-tailed" after the alpha value (e.g., $p = .044 < .05$ two-tailed).

Alpha and p are different; *together* they can provide non-redundant information to the researcher and to the reader of research. This non-redundancy is most evident if "exact" p's or probability intervals are given.

References

American Psychological Association. (1983). *Publication manual of the American Psychological Association* (3rd ed.). Washington, DC: Author.

Fisher, R. A. (1947). *The design of experiments* (4th ed.). New York: Hafner.

Gibbons, J. D., & Pratt, J. W. (1975). *P*-values: Interpretation and methodology. *American Statistician, 29,* 20–25.

Hopkins, K. D., & Glass, G. V (1978). *Basic statistics for the behavioral sciences.* Englewood Cliffs, NJ: Prentice-Hall.

Huberty, C. J. (1985, April). *On statistical testing.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Kempthorne, O. (1972). Theories of inference and data analysis. In T. A. Bancroft (Ed.), *Statistical papers in honor of George W. Snedecor.* Ames: Iowa State University.

Keppel, G. (1982). *Design and analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kirk, R. E. (1982). *Experimental design: Procedures for the behavioral sciences* (2nd ed.). Belmont, CA: Brooks/Cole.

Lehmann, E. L. (1968). Hypothesis testing. In D. L. Sills (Ed.), *International encyclopedia of the social sciences: Vol. 7.* New York: Macmillan and Free Press.

McNeil, K. A., Kelly, F. J., & McNeil, J. T. (1975). *Testing research hypothesis using multiple linear regression.* Carbondale: Southern Illinois University.

Stallings, W. M., & Singhal, S. (1969). Confidence level and significance level: Semantic confusion or logical fallacy? *Journal of Experimental Education, 37,* 57–59.

Tatsuoka, M. M. (1982). Statistical methods. In H. E. Mitzel (Ed.), *Encyclopedia of educational research: Vol. 4.* New York: Free Press.