Psychological Methods
2000, Vol. 5, No. 4, 411–414

1082-989X/00/$5.00 DOI: 10.1037//1082-989X.5.4.411

A Sensible Formulation of the Significance Test

Lyle V. Jones
University of North Carolina at Chapel Hill

John W. Tukey
Princeton, New Jersey

The conventional procedure for null hypothesis significance testing has long been the target of appropriate criticism. A more reasonable alternative is proposed, one that not only avoids the unrealistic postulation of a null hypothesis but also, for a given parametric difference and a given error probability, is more likely to report the detection of that difference.

Procedures for testing the significance of the difference between two sample means have been widely used, and also widely criticized, throughout most of the 20th century. To better codify the procedures, the Neyman–Pearson approach to choosing between two point hypotheses emerged. Although attractive mathematically, that approach is troublesome in practice because investigators are often unwilling to specify a value for the alternative to the null hypothesis and because, of equal importance, both hypotheses, like all point hypotheses about parameters in real-world populations that are subjected to different treatments, are always untrue when calculations are carried to enough decimal places.

The introduction of hypothesis testing did lead to an attempt to reformulate the significance test as a test of a point null hypothesis against all possible alternatives. However, that reformulation failed to silence the critics, for example, Bakan (1966), Bechhofer (1954), Berkson (1938, 1942), Cohen (1994), Lykken (1968), Meehl (1967), Rozeboom (1960), Schmidt (1996), and Schmidt & Hunter (1997), among many others. Jones (1955) suggested that

> an investigator would be misled less frequently and would be more likely to obtain the information he seeks were he to formulate his experimental problems in terms of the estimation of population parameters, with the establishment of confidence intervals about the estimated values, rather than in terms of a null hypothesis against all possible alternatives. (p. 407)

Many other critics have echoed that advice, to which we also subscribe, especially when the outcome measure is well defined.

However, when faced with several mean differences, that is, with multiple comparisons, straightforward procedures for establishing confidence intervals may not be available. Also, in some situations, the scale of the outcome measure may be so untrustworthy that the primary interest then may appropriately reside in just the direction of a treatment effect rather than in its size. We propose an alternative to the conventional formulation for whenever a test of significance is to be performed for these or other reasons. (Additional arguments in support of significance testing have been set forth by Abelson, 1997; Hagen, 1997; Mulaik, Raju, & Harshman, 1997; and Wainer, 1999, among others.)

A common formulation for the conventional test-of-hypothesis version of the test of significance is the traditional test for the equality of two population means, using Student's t distribution. A null hypothesis is set forth, H_0: $\mu_A - \mu_B = 0$ versus an omnibus alternative H_1: $\mu_A - \mu_B \neq 0$. From samples of sizes n_A and n_B, an estimate of $\mu_A - \mu_B$, $y_A - y_B$ is obtained. Usually, y_A and y_B are sample means, but they might be sample medians, sample midmeans, or other estimates of μ_A and μ_B. An estimated standard error

Lyle V. Jones, L. L. Thurstone Psychometric Laboratory, Department of Psychology, University of North Carolina at Chapel Hill; John W. Tukey, Princeton, New Jersey.

John Tukey was recovering from a mild stroke when unexpected complications led to his death on July 26, 2000. Final revisions to the manuscript were completed after that date, and thus are the sole responsibility of the first author.

This project was supported in part by the National Institute of Statistical Sciences with National Science Foundation Grant RED-9350005. We thank Thomas S. Wallsten and Forrest W. Young for constructive criticisms of a draft of this article.

Correspondence concerning this article should be addressed to Lyle V. Jones, CB 3270, Department of Psychology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-3270. Electronic mail may be sent to lvjones@email.unc.edu.

of $y_A - y_B$, s_d, is calculated, and a statistic, $(y_A - y_B)s_d$, is formed.

A two-tailed rejection region of the sampling distribution of this statistic, typically Student's t with $df = n_A + n_B - 2$, is established by setting a value for α, the probability of rejecting H_0 when it is true, often set as .05. When the t statistic falls in either tail of the rejection region, each with area $\alpha/2$, H_0 is rejected in favor of H_1. The only allowable conclusion is (a) reject H_0 or (b) fail to reject that null hypothesis, thereby withholding judgment.

Cohen (1994) advised, "Don't look for a magic alternative to" this formulation of the null hypothesis significance test; "it doesn't exist" (p. 1001). In the same article, Cohen cited Tukey (1991) but seems to have overlooked the alternative suggested in Tukey's article (see also Tukey, 1993). As applied to multiple comparisons, that alternative has been more explicitly developed and illustrated by Williams, Jones, & Tukey (1999). As shown below, the formulation is entirely suitable for use with a single mean difference as well as with multiple comparisons.

When A and B are different treatments, μ_A and μ_B are certain to differ in some decimal place so that $\mu_A - \mu_B = 0$ is known in advance to be false and $\mu_A - \mu_B \neq 0$ is known to be true (Cohen, 1990; Tukey, 1991). An extensive rebuttal to this claim has been provided by Hagen (1997), who stated that "I agree that A and B will always produce differential effects on some variable or variables that theoretically could be measured. But I do not agree that A and B will always produce an effect on the dependent variable." (p. 20). We simply do not accept that view. For large, finite, treatment populations, a total census is at least conceivable, and we cannot imagine an outcome for which $\mu_A - \mu_B = 0$ when the dependent variable (or any other variable) is measured to an indefinitely large number of decimal places. (We come to a similar conclusion with respect to the mean of a single population when H_0: $\mu = k$, whether $k = 0$ or any other definite value.) For hypothetical treatment populations, $\mu_A - \mu_B$ may approach zero as a limit, but as for the approach of population sizes to infinity, the limit is never reached. The population mean difference may be trivially small but will always be positive or negative. As a consequence, we should not set forth a null hypothesis because to do so is unrealistic and misleading. Instead, we should assess the sample data and entertain one of three conclusions: (a) act as if $\mu_A - \mu_B > 0$; (b) act as if $\mu_A - \mu_B < 0$; or (c) act as if the sign of $\mu_A - \mu_B$ is indefinite, that is, is not (yet) determined. This specification is similar to "the three alternative decisions" proposed by Tukey (1960, p. 425).

With this formulation, a conclusion is in error only when it is "a reversal," when it asserts one direction while the (unknown) truth is the other direction. Asserting that the direction is not yet established may constitute a wasted opportunity, but it is not an error. We want to control the rate of error, the reversal rate, while minimizing wasted opportunity, that is, while minimizing indefinite results.

Consider the probability of erroneously concluding that the population mean difference is in one direction when in truth it is in the other. If $\mu_A - \mu_B$ is positive, an error will occur only when the value of t falls in the lower tail of the distribution, with area $\alpha/2$, which is also the limiting probability of error for that case. Likewise, if $\mu_A - \mu_B$ is negative, an error occurs only when the t statistic falls in the upper tail of the distribution, also of area $\alpha/2$. The reversal rate, also the overall probability of error, is then

$$\begin{aligned} P(\text{error}) &\leq (\alpha/2)\, P[(\mu_A - \mu_B) > 0] + (\alpha/2) P[(\mu_A - \mu_B) < 0] \\ &\leq (\alpha/2)\, P[(\mu_A - \mu_B) \neq 0] \leq \alpha/2. \end{aligned}$$

This P(error) is just one half the probability of Type I error for the traditional null hypothesis test and is the same as that for a one-tailed test (see Jones, 1952). However, the procedure is symmetric, equally sensitive to a mean difference in either direction.

A task force on statistical inference is currently preparing guidelines for standards to be adopted in psychology journals. In an earlier article, Wilkinson and The Task Force on Statistical Inference (1999) recommended that, when hypothesis tests have been performed, effect sizes and p values always be presented. The adoption of this proposed three-alternative conclusion procedure lends itself to reporting the effect size as the estimated value of $\mu_A - \mu_B$ (either standardized or not). Regardless of the size of that estimate, and regardless of whether or not the calculated value of t falls in the rejection region, it seems appropriate to report the p value as the area of the t distribution more positive or more negative (but not both) than the value of t obtained from $(\mu_A - \mu_B)/s_d$. (The limiting values of p are then 0, as the absolute value of t becomes indefinitely large, and 1/2, as the value of t approaches zero.)

For any specified positive or negative population mean difference, there may be found in the usual way the probability of a Type II error, of withholding judgment when the parametric difference is as specified. For each specified difference, the probability of a Type II error is smaller than that for the conventional two-tailed test of significance. Thus, the proposed procedure is uniformly more powerful than the conventional procedure.

Hodges and Lehmann (1954) proposed a modification of the traditional Student test, converting it from a two-sided test of the null hypothesis to 2 one-sided tests. Kaiser (1960) proposed combining the 2 one-sided tests into a single test, but one with two directional alternative hypotheses, $\mu_A < \mu_B$ and $\mu_A > \mu_B$. (For further discussion, see Bohrer, 1979; Bulmer, 1957; and Harris, 1997.) However, the unrealistic null hypothesis of zero mean difference in the population is included in these proposals, in contrast to the formulation above. By acknowledging the fiction of the null hypothesis and following the implications from "every null hypothesis is false," our formulation yields, for any sample size and any value of α, a test with greater reported sensitivity to detect the direction of a difference between two population means.

Note that those accustomed to make "tests of hypotheses" at .05 would, using the procedures set forth here, do the same arithmetic but would describe their results as a "test of significance" at one half of .05, that is, at .025.

Alternatively, to maintain at .05 the probability of acting as if the parametric difference is in one direction when, in fact, it is in the other, the investigator would employ the .10 tabled value of α.

In summary, then:

- Prefer confidence intervals when they are available.
- Recognize that point hypotheses, while mathematically convenient, are never fulfilled in practice.
- When performing a simple test of significance, seek one of three outcomes, as described above.

References

Abelson, R. P. (1997). A retrospective on the significance test ban of 1999 (If there were no significance tests, they would be invented). In L. A. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 117–144). Mahwah, NJ: Erlbaum.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 1–29.

Bechhofer, R. E. (1954). A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics, 25,* 16–39.

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. *Journal of the American Statistical Association, 33,* 526–536.

Berkson, J. (1942). Tests of significance considered as evidence. *Journal of the American Statistical Association, 37,* 325–335.

Bohrer, R. (1979). Multiple three-decision rules for parametric signs. *Journal of the American Statistical Association, 74,* 432–437.

Bulmer, M. G. (1957). Confirming statistical hypotheses. *Journal of the Royal Statistical Society, Series B, 19,* 125–132.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist, 49,* 997–1003.

Hagen, R. L. (1997). In praise of the null hypothesis statistical test. *American Psychologist, 52,* 15–24.

Harris, R. J. (1997). Reforming significance testing via three-valued logic. In L. A. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 145–174). Mahwah, NJ: Erlbaum.

Hodges, J. L., & Lehmann, E. L. (1954). Testing the approximate validity of statistical hypotheses. *Journal of the Royal Statistical Society, Series B, 16,* 261–268.

Jones, L. V. (1952). Tests of hypotheses: One-sided vs. two-sided alternatives. *Psychological Bulletin, 49,* 43–46.

Jones, L. V. (1955). Statistics and research design. *Annual Review of Psychology, 6,* 405–430.

Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review, 67,* 160–167.

Lykken, D. (1968). Statistical significance in psychological research. *Psychological Bulletin, 70,* 151–159.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34,* 103–115.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. A. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 65–116). Mahwah, NJ: Erlbaum.

Rozeboom, W. W. (1960). The fallacy of the null hypothesis significance test. *Psychological Bulletin, 57,* 416–428.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1,* 115–129.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. A. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 37–64). Mahwah, NJ: Erlbaum.

Tukey, J. W. (1960). Conclusions vs. decisions. *Technometrics, 2,* 423–433. [Also in L. V. Jones (Ed.) (1986). *The collected works of John W. Tukey: Vol. III. Philosophy and principles of data analysis: 1949–1964* (pp. 127–142). Monterey, CA: Wadsworth & Brooks/Cole.]

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6,* 100–116.

Tukey, J. W. (1993). Where should multiple comparisons go next? In F. M. Hoppe (Ed.), *Multiple comparisons, selection, and applications in biometry* (pp. 187–208). New York: Dekker.

Wainer, H. (1999). One cheer for null hypothesis significance testing. *Psychological Methods, 6,* 212–213.

Wilkinson, L., & The Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics, 24,* 42–69.

Received November 30, 1999
Revision received August 18, 2000
Accepted August 21, 2000

■