Journal of Management
1989, Vol. 15, No. 2, 335-344

.05: A Case of the Tail Wagging the Distribution

Kerry S. Sauley
Arthur G. Bedeian
Louisiana State University

After a brief review of the origins of the .05 criterion for judging statistical significance, this article challenges the management discipline's rigid adherence to conventional levels of significance for differentiating reliable from unreliable results. Arguing instead that there is no right or wrong level of significance, it contends that the selection of a significance level should be treated as one more research parameter. In this respect, the appropriateness of a specific level of significance should be based on such considerations as sample size and research design rather than a priori declarations of .05, .01, or whatever.

It is customary in the management literature to report a "level of statistical significance" when testing a hypothesis. This significance level refers to the probability of committing a Type I error (i.e., the probability of wrongly rejecting a true null hypothesis) and, in part, to the probability of committing a Type II error (i.e., the probability of accepting a wrong null hypothesis). As traditionally viewed, reliance on conventional (i.e., .05, .01, or .001) significance levels offers certain advantages. For instance, they provide a criterion for publication decisions, they assist in evaluating and accumulating results from different studies, and, in situations where it is difficult to estimate the consequences associated with making a mistake in accepting or rejecting a hypothesis, they offer a rule of thumb for judging research results (Rudner, 1953).

It is seemingly often overlooked, however, that for conventional significance levels to be used to make scientific inferences, they must be uniform from application to application—a quality that exists only in the case of equally valid and equally powerful studies. What often occurs, for example, is that a .05 level of significance is really .10 or .15 or .20 depending on a study's validity and statistical powerfulness. Because studies are seldom equally valid and equally powerful, the actual value of a significance criterion is likewise seldom uniform. This lack of uniformity is even further compounded in situations requiring the evalu-

The helpful comments of Hubert S. Feild and Chester A. Schriesheim on an earlier draft manuscript are gratefully acknowledged.

Address all correspondence to Kerry S. Sauley and Arthur G. Bedeian, Department of Management, Louisiana State University, Baton Rouge, Louisiana 70803.

ation of a series of statistical tests. In such situations, the probability of family-wise Type I errors increases as the number of significance tests increases. Alternatives to control for the effects of multiple-significance tests on Type I errors are reviewed in Feild and Armenakis (1974), among others.

The purpose of this article is not to debate whether statistical inferences should be made or to advise caution in situations requiring use of multiple tests of significance, but to challenge the management discipline's rigid adherence to conventional levels of significance for differentiating reliable from unreliable results. Both sociology (Morrison & Henkel, 1969) and psychology (Cowles & Davis, 1982) have undergone a parallel commentary.

Judging by current practice, there is every indication that many management researchers are unaware of the complex issues that surround the setting and reporting of significance levels. At present, the levels of .05, .01, and .001 are almost universally selected, irrespective of research problem, sample size, or research design. Of these three, the .05 criterion is unquestionably the most sacred (Skipper, Guenther, & Nass, 1967).

Viewed from a practical perspective, fixed adherence to the .05 level as the maximum acceptable probability for determining statistical significance has had the consequence of differentiating reliable from unreliable results and, perhaps, even more pragmatically, publishable from unpublishable research (Lewis & Lewis, 1980; Rosenthal, 1979). Convention notwithstanding, there is little, if any, justification for such a rigid state of affairs. Indeed, Lykken (1968) has argued that statistical significance is the "least important attribute of an experiment" (p. 158). It is his position that the value of any research should be determined by subjectively evaluating such concerns as the coherence of the underlying theory, degree of experimental control, and sophistication of the measuring techniques.

Origins

It is generally held that the .05 level of significance originated in 1925 when Sir Ronald A. Fisher developed the popular analysis of variance (ANOVA) statistical procedure. The following statement in his book *Statistical Methods for Research Workers* is often cited as the first mention of the $p = .05$ criterion.

> The value for which $p = .05$, or 1 in 20, is 1.96 or nearly 2; it is convenient to take this point as a limit in judging whether a deviation is to be considered significant or not. Deviations exceeding twice the standard deviation are thus formally considered significant. (p. 46)

Commenting on a similar, but slightly later statement by Fisher (1926, p. 504), in a paper on agricultural field experiments, Cochran (1976) states that, "students sometimes ask 'How did the 5% significance level or Type I error come to be used as a standard?' I am not sure, but this is the first comment known to me on the choice of 5%." Fisher (1926) goes on to acknowledge that other levels of significance may be selected:

> If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty (the 2 per cent. point), or one in a

> hundred (the 1 per cent. point). Personally, the writer prefers to set a low standard of significance at the 5 per cent. point, and ignore entirely all results which fail to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance. (p. 504)

In Cochran's (1976) judgement, "Fisher sounds fairly casual about the choice of 5% for the significance level, as the words 'convenient' and 'prefers' have indicated" (p. 17). Based on Fisher's statements, however, Cowles and Davis (1982) believe that there is "no doubt" that he accepted the .05 level as the critical cutoff point for determining statistical significance.

Support for the belief that use of the .05 level of significance is subjective, or what Yule and Kendall (1950) referred to as a "matter of personal taste" (p. 472), however, is readily available. Winer (1962), for example, states that "The frequent use of the .05 and .01 levels of significance is a matter of convention having little scientific or logical basis" (p. 13). Similarly, Camilleri (1962) says that "the particular level of significance chosen for an investigation is not a logical consequence of the theory of statistical inference." Agreeing with Yule and Kendall (1950), he concludes, "We are free to choose whatever level seems appropriate" (p. 176).

Earlier Statements

Although Fisher (1925) is credited with the first specific mention of the .05 level of significance and his 1926 paper concerning agricultural field experiments seems to contain the first specific discussion of significance levels, the notion of odds against an hypothesis being true is clearly much older. Indeed, John Venn (1988), of Venn diagram fame, used a remarkably similar criterion some 40 years earlier and is believed to be the first writer to use the term *significant* in a statistical sense: "When we are dealing with statistics, we ought to be able not merely to say vaguely that the difference does or does not seem significant to us, but we ought to have some test as to what difference would be significant" (pp. 147-148).

Like statements can be found in other sources. William Gosset (1908), writing in the famous article in which he developed the *t*-test, (using the *non de plume* "Student") commented that "three times the probable error in the normal curve, for most purposes, would be considered significant" (p. 13). McGaughy (1924) used the term "critical ratio" to describe the value $x/PE = 3$, where x represents a difference and *PE* probable error. McCall (1923) applied the term "experimental coefficient" for the ratio $x/2.786$. Other examples could be easily cited (e.g., Flugel, 1925; Peters, 1933; Wood & Stratton, 1910).

What seems evident, however, is that Fisher's selection of the .05 criterion may not have been as innovative as some authorities have assumed. Prevailing misconceptions aside, there was little new in Fisher's choice of the .05 level of significance. Indeed, Cowles and Davis (1982) argue that Fisher cannot be given credit for originating today's "recommended" levels of significance, but rather can be credited with expressing the value of a significance level in terms of its standard deviation rather than its probable error.

Subjective Probability

Knowing how the .05 level of significance seemingly came about begs the question of why this criterion was judged appropriate by early researchers (Cowles & Davis, 1982). Addressing this point, Walker (1929) suggests that the concept of *subjective probability* may offer the answer. She writes:

> Any numerical definition of "practical certainty" must be more or less arbitrary, inasmuch as certainty is a subjective matter, and odds which seem practically certain to one man in a given situation might well be unconvincing to another man in another situation. However, $x = 3\sigma$ is a very convenient limit and most people would be convinced by odds of 369 to 1. (p. 56)

The mental processes on which subjective probability depend have been explored by Alberoni (1962a, 1962b). Based on a stream of experiments, he concluded that individuals arrive at the idea of *chance* when they are unable to find a cause or regular pattern to explain the reason for a certain event. Once the notion of chance is accepted, individuals form a system of expectations for the future. If reality confirms these expectations, they continue to consider the event as being due to chance; if reality contradicts the system of expectations, a cause is introduced to explain the discrepancy. The point at which chance is abandoned as an explanation depends on each individual's discrepancy tolerance. Alberoni (1962) refers to this point as the "threshold of dismissal of the idea of chance" (p. 262). This process is operationally parallel to the notion of "covariation analysis" as encountered in attribution research (Kelley, 1967).

The fundamental question, however, is whether (in terms of Alberoni's "threshold") individuals view discrepancies that occur 5% of the time or less to be rare events (i.e., significant). Or, stated differently, are they prepared to introduce a cause other than chance to explain such discrepancies (Cowles & Davis, 1982)? A partial answer to this question can be found in the work of Rosenthal and Gaito (1963), who investigated the degree of confidence a group of researchers placed in various levels of significance (p) ranging from .001 to .50. For 84% of the researchers, the .05 level had "cliff characteristics" evidenced by a relatively steeper loss of confidence in moving from the .05 to the .10 level than was the case at either higher or lower levels of significance. Whether this effect was due to subjective probabilities or the researchers' training is, however, uncertain.

.05 Reconsidered

Over the past three decades, controversy has surrounded the selection of levels of significance. Selvin's (1957) classic article, in particular, prompted a variety of replies defending the use of significance tests. Selvin and his fellow critics not only took exception to tests of significance, but the selection of significance levels. Camilleri (1962), for instance, maintained that a particular level of significance only has meaning in practical situations, such as a quality control department, where the cost of accepting or rejecting an item can be expressed in terms of profit or loss. Because no such criteria exist in the social sciences, "we are free to choose whatever level seems appropriate" (p. 151). Taking a similar po-

sition, Morrison and Henkel (1969) argue that "to insist on the .05 to .01 level is, then, to talk about the science of business, not the business of science" (p. 137). They further contend that unquestioned use of conventional levels of significance reduces probability to a false dichotomy of significant or not significant, thereby reducing research to a rote decision that is inappropriate in scientific work. Morrison and Henkel (1969) conclude with an attack on sociology that arguably applies to contemporary management research:

> If, indeed, .05 (or any other level) is 'sacred' ... then what we do in sociology [read management] surely is much more akin to religion than science and we might as well forget empirical work and get on with the development of more rituals. (p. 137)

The point of these criticisms is quite direct. Briefly put, there is no right or wrong level of significance. Blind adherence to the .05 level of significance as the crucial value for differentiating publishable from unpublishable research cannot be justified. As Skipper et al. (1967) suggest, the selection of a significance level by a researcher should be treated as one more research parameter. Rather than being set at a priori levels of .05, .01, or whatever, the appropriateness of a specific level of significance should be based upon considerations such as discussed below and summarized in Table 1:

Table 1
Sample Considerations in Selecting an Appropriate Level of Statistical Significance

Consideration	Significance Level: Less Conservative	Significance Level: More Conservative
1. *Sample Size*		
With large samples		√
With small samples	√	
2. *Effect Size*		
With large effect size		√
With small effect size	√	
3. *Measurement Error* (ME)		
With large ME		√
With small ME	√	
4. *Null Hypothesis/Practical Consequences of Rejecting*		
Practical consequences are high		√
Practical consequences are low	√	
5. *Coherence of Underlying Theory*		
Plausibility of alternative hypothesis is high		√
Plausibility of alternative hypothesis is low	√	
6. *Degree of Experimental Control*		
High amount of experimental control	√	
Low amount of experimental control		√
7. *Robustness*		
Test assumptions have been met	no change	
Test assumptions have not been met/ but test is robust to violation		√
Test assumptions have not been met and test isn't robust to violation		√+

1. *Statistical power.* The power of a significance test is the probability that it will result in the conclusion that a phenomenon exists. Formulae for determining statistical power are available in Kirk (1978) and elsewhere. Measurement error aside, statistical power is a function of three factors: (a) effect size (i.e., the *degree* to which a phenomenon actually exists in a population), (b) sample size, and (c) level of significance. As effect size increases, power increases, other things being equal. The same relationship exists between sample size and power, as well as between level of significance and power.

A common problem for both effect size and sample size is that significance tests are relatively insensitive to the strength of the relationships between variables. As regards effect size, if a researcher expects a large effect, a conservative level of significance should be selected based on the rationale that if a large difference between two groups is expected, but only a small difference is obtained, the null hypothesis should not be rejected. With respect to sample size, with a large enough n, any relationship (or effect) between two variables will be found to be statistically "significant," but may not be of any practical or substantive importance. To illustrate, a correlation of .10 is significant at the .01 level when sample size equals 1,000 or more cases—a finding of little practical import. Likewise, with a small enough sample size, any relationship between two variables will be found to be statistically "nonsignificant," regardless of its practical or substantive importance. To wit, when sample size equals 10 cases, a correlation of .75 is nonsignificant at the .01 level—an instance of too little power.

Thus, when either sample size or anticipated effect size are large, a researcher should typically select a more conservative level of significance (e.g., .01 vs. .05). Conversely, when either sample size or anticipated effect size are small, a researcher should typically select a less conservative level of significance (e.g., .10 vs. .05).

2. *Measurement error.* A second consideration in selecting a level of significance involves measurement error. As typically defined, measurement error is a source of either systematic bias or random errors that tend to obscure the degree of lawfulness that can be detected among underlying concepts (Nunnally, 1978). Sources of measurement error include errors due to coding, interviewer fatigue, ambiguous instructions, memory lapses, and the like. When a large amount of measurement error exists (e.g., an inaccurate measuring procedure), this unreliability inflates standard errors of estimates with the result that the likelihood of finding significant differences between means of different treatment groups is considerably reduced.

Recommendations for selecting a significance level in cases where measurement error is relatively high are difficult, at best. At the very least, a more conservative level of significance should be selected to avoid interpreting random variation as significant. The role of replication is key in such situations because a pattern of results is typically more meaningful than a single, statistically significant finding. Indeed, as Keppel (1982) puts it, "the ultimate test of any significant finding *is* its repeatability" (p. 74).

A final point is that no significance level will really compensate for unreliable measures. Although correction for attenuation formulas may help in this regard

somewhat, if measures are unreliable then no strong statements can be made about the effects being investigated.

3. *Null hypothesis*. Consideration in selecting a level of significance should also be given to the practical consequences of rejecting a null hypothesis. With respect to the practical consequences of Type I errors, a more conservative level of significance should be selected as the cost of rejecting a null hypothesis increases. For example, if the hypothesis being considered were to the effect that a toxic workplace chemical was not present in harmful quantity, we would demand a high degree of confirmation before accepting the hypothesis because the consequences of a mistake are quite high. On the other hand, if, on the basis of a sample, our hypothesis stated that a run of plastic was not defective, the degree of confirmation required would not be as high. Thus, the degree of confidence required before accepting a hypothesis should be related to the consequences of making a mistake (Rudner, 1953).

4. *Coherence of underlying theory.* A level of significance should be selected in relation to a body of knowledge or theory rather than in isolation. For example, when a hypothesis goes against theory, empirical research, or perhaps even common sense, a much more conservative level of significance should be selected in order to avoid a Type I error that would lead research into a fruitless area and be costly in terms of wasted time, money, and effort. Therefore, when alternative hypotheses are highly plausible, a much more conservative level of significance (e.g., .01) should be selected than when the plausibility of alternative hypotheses is low (e.g., .05). Similarly, if the plausibility of alternative hypotheses is largely unknown (due, for instance, to an absence of underlying theory), a more conservative level of significance should also be selected to avoid capitalizing on chance. Contrary to 'conventional wisdom,' this would particularly be the case in ''exploratory'' research whose results provide the basis for further research that may be fruitless if those results are unsound. Indeed, as Lindquist (1953) advises, ''if we always set a high level of significance for our tests at the exploratory level, we may be quite sure that we will not follow many completely false leads, and at the same time, we will have some assurance that the true leads which we ignore (because of Type II errors) are probably among the less promising ones'' (p. 69).

5. *Degree of experimental control*. The degree of experimental control in a research design should also be considered in selecting a level of significance. As experimental control increases, both the error rate and likelihood of alternative explanations or what Campbell and Stanley (1966) have dubbed ''threats to internal validity'' decrease. Hence, when a research design has a high degree of experimental control, a researcher is more free to select a less conservative level of significance. One would thus select a lower level of significance (e.g., .01 vs. .05) in a quasi-experimental design than in a ''true'' experimental design based upon the above considerations.

6. *Robustness*. The robustness of a statistical test is another criterion that should be examined in selecting a level of significance. If the assumptions underlying a test of significance are met, then it can be assumed that researchers are testing their hypotheses at the stated a priori significance level (e.g., they are in-

deed testing a hypothesis, stated a priori at the .01 level, at the .01 level). However, if assumptions underlying a test of significance are violated, researchers may be testing their hypotheses at a level lower than that stated (e.g., they may be testing their hypotheses at the .10 level instead of at the stated .05 level). This problem is further compounded if a test of significance is not robust to violations of its underlying assumptions. As a result, when assumptions of a test are violated, a more conservative level of significance should be selected. For example, when researchers violate the assumptions of a test and are testing a hypothesis at the .05 level, they should select a more conservative level, such as .02 or .01, in order both to compensate for the violation and to actually test the hypothesis at the .05 level. Of course, it should be realized that if a test's assumptions are violated sufficiently, then its outcome may well be without interpretive value. Obviously, in such instances, the selection of a more conservative significance level will not compensate for the violation of a test's underlying assumptions.

Discussion and Conclusion

The selection of a particular significance level should not be seen as a panacea for interpreting evidence. Critical thinking is necessary for verifying research propositions. Thus, *both* practical and theoretical significance should be of concern. Authors, as well as editors and reviewers, must take the responsibility for deciding which research is good and which is not, without letting the p value of a significance test *solely* determine this decision (Lykken, 1968). When *all* the parameters inherent in a research problem are not fully considered prior to selecting an appropriate significance level, the academic enterprise is, at least in part, a form of "scientific roulette" (Bakan, 1966), in which lucky players win (i.e., have their papers published). As long as this practice goes unchallenged, it will remain easy to document the joy experienced by Skipper et al.'s (1967) mythical researcher when her f ratio yielded significance at .05, as well as her horror when the table of significance reached "only" .10 or .06.

In sum, the purpose of this note has been to challenge the management discipline's rigid adherence to conventional levels of significance for differentiating reliable from unreliable results. It has been suggested that the frequent use of the .05 level as the maximum acceptable probability for determining statistical significance is too often a matter of custom rather than independent thought. Indeed, it may be argued that the precision and empirical concreteness associated with establishing a discrete borderline between accepting and rejecting a hypothesis are illusory. Adherence to such a rigid strategy may be appropriate for some designs, but doubtlessly detracts from interpretive power in others.

The reliance that management researchers place on conventional levels of significance, as if they were a reliable arbiter of truth, belies the complexity surrounding the setting and reporting of significance levels. A more rational approach calls for fully considering all the parameters inherent in a research problem prior to selecting an appropriate significance level. In this sense, there is no right or wrong level, but a need for management researchers to exercise scientific judgment.

References

Alberoni, F. (1962a). Contribution to the study of subjective probability. I. *Journal of General Psychology, 66,* 241-264.

Alberoni, F. (1962b). Contribution to the study of subjective probability: Prediction. II. *Journal of General Psychology, 66,* 265-285.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66,* 423-437.

Camilleri, S.F. (1962). Theory, probability, and induction in social research. *American Sociological Review, 27,* 170-178.

Campbell, D.T., & Stanley, J.C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

Cochran, W.G. (1976). Early development of techniques in comparative experimentation. In D.B. Owen (Ed.), *On the history of statistics and probability* (pp. 2-25). New York: Dekker.

Cowles, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist, 37,* 553-558.

Feild, H.S., & Armenakis, A.A. (1974). On use of multiple tests of significance in psychological research. *Psychological Reports, 35,* 427-431.

Fisher, R.A. (1925). *Statistical methods of research workers.* Edinburgh: Oliver & Boyd.

Fisher, R.A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture, 33,* 503-513.

Flugel, J.C. (1925). A quantitative study of feeling and emotion in everyday life. *British Journal of Psychology, 15,* 318-355.

Kelley, H. (1967). Attribution theory in social psychology. In D. Levine (Ed.), *Nebraska symposium on motivation* (pp. 192-238). Lincoln: University of Nebraska Press.

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

Kirk, R.E. (1978). *Introductory statistics.* Monterey, CA: Brooks/Cole.

Lewis, G.H., & Lewis, J.F. (1980). The dog in the night-time: Negative evidence in social research. *British Journal of Sociology, 31,* 544-558.

Lindquist, E.F. (1953). *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.

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

McCall, W.A. (1923). *How to experiment in education.* New York: Macmillan.

McGaughy, J.R. (1924). *The fiscal administration of city school systems.* New York: Macmillan.

Morrison, D.E., & Henkel, R.E. (1969). Significance tests reconsidered. *American Sociologist, 4,* 131-140.

Nunnally, J.C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

Peters, C.C. (1933). Note on a misconception of statistical significance. *American Journal of Sociology, 39,* 231-236.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

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

Rudner, R. (1953). The scientist *qua* scientist makes value judgments. *Philosophy of Science, 20,* 1-6.

Selvin, H.C. (1957). A critique of tests of significance in survey research. *American Sociological Review, 22,* 519-527.

Skipper, J.K., Jr., Guenther, A.L., & Nass, G. (1967). The sacredness of .05: A note concerning the uses of statistical levels of significance in social science. *American Sociologist, 2,* 16-18.

Student [Gossett, W.S.]. (1908). The probable error of a mean. *Biometrika, 6,* 1-25.

Venn, J. (1888). Cambridge anthropometry. *Journal of the Anthropological Institute, 18,* 140-154.

Walker, H.M. (1931). *Studies in the history of statistical method.* Baltimore: Williams & Wilkins.

Winer, B.J. (1962). *Statistical principles in experimental design.* New York: McGraw-Hill.

Wood, T.B., & Stratton, F.J.M. (1910). The interpretatin of experimental results. *Journal of Agriculture Science, 3,* 417-440.

Yule, G.U., & Kendall, M.G. (1950). *An introduction to the theory of statistics* (14th ed.). London: Griffin.