primary concern of the social scientist ought to be the development of such theories.

We distinguish three major uses of probability in science: in intrinsic statistical hypotheses; in auxiliary statistical hypotheses, including the theory of errors of observation, sampling from finite populations, and randomized experimentation; and thirdly in inductive probability policies, particularly the theory of tests of significance.

We have tried to show that the hypothetical character of the risk probabilities associated with the level of significance and the pragmatic ambiguities of the rationale for choosing any particular level of significance seriously undermine its value in the evaluation of statistical hypotheses.

It is our belief that the great reliance placed by many sociologists on tests of significance is chiefly an attempt to provide scientific legitimacy to empirical research without adequate theoretical significance. Lest we be misunderstood we wish to emphasize the point that the construction of deductive theory is developmental, involving an accumulation of research experience. But it also involves formal reasoning, and the task of the scientist is to expose his thinking to a formal reconstruction, to formalize it as well as he can, so that it can contribute to the construction of a deductive system of some size and scope.

17. The Sacredness of .05: A Note Concerning the Uses of Statistical Levels of Significance in Social Science

JAMES K. SKIPPER, JR., ANTHONY L. GUENTHER, and GILBERT NASS

Decisions regarding the uses of statistical levels of significance have typically been rendered by social scientists with specialized training and considerable expertise. However, a strong case can be made for more involvement in such issues by researchers whose activities include tests of significance in an applied context. The issue of whether inference should be made at all has been rather thoroughly debated (Selvin, 1957 [Chapter 9]; Gold, 1958 [Chapter 10]; and Kihl, 1959 [Chapter 15]), but there are indications that the use of present "recommended" levels of significance is due for reassessment. This note attempts to state the problem, explore the issues involved, and suggest an alternative to current policy.

The choice of a statistical level of significance, that is, establishing the probability of rejecting the null hypothesis when in fact it is true, apparently demands little psychic energy on the part of researchers. Casual examination of the literature discloses that the common, arbitrary, and virtually sacred levels of .05, .01, and .001 are almost universally selected regardless of the nature and type of problem. Of these three, .01 is perhaps most sacred.

Reprinted from The American Sociologist, 2 (February, 1967), 16-18, by permission of the authors and the publisher. Copyright © 1967 by the American Sociological Association.

1. For a discussion of the more general question of the current usefulness of mathematics and statistics in social science, the recent comments by Parsons (1964), Sibley (1965), Selvin (1965), and Etzioni (1965) are of special interest.
Although statistically-inclined methodologists do not often explicitly recommend use of these arbitrary levels, their positions frequently suggest that these levels are conventional. Prominent textbooks bear this out:

But the agony of making a fresh selection of the level of significance in each instance would be a painful business. Hence, the statistical worker finds welcome relief in the .05 convention, which prescribes that the null hypothesis is automatically to be rejected whenever the probability of being wrong in that decision is 5 percent or less (Mueller and Schuessler, 1961: 395 [Skipper et al., italics]).

By locating the observed findings in the theoretical distribution of all possible findings, the investigator determines the probability of obtaining the finding by chance if the null hypothesis is actually correct. If this probability turns out to be small enough (less than a predetermined level such as .05 or .01), he then decides to reject the null hypothesis (Riley, 1963: 638).

In the social sciences, it is more or less conventional to reject the null hypothesis when the statistical analysis indicates that the observed difference would not occur more than 5 times out of 100 by chance alone (Sellitz et al., 1961: 418 [Skipper et al., italics]).

A convention frequently followed is to state the result significant if the hypothesis is rejected with $\alpha = .05$ and highly significant if it is rejected with $\alpha = .01$ (Dixon and Massey, 1957: 91 [Skipper et al., italics]).

The current obsession with .05, it would seem, has the consequence of differentiating significant research findings and those best forgotten, published studies from unpublished ones, and renewal of grants from termination. It would not be difficult to document the joy experienced by a social scientist when his F ratio or t value yields significance at .05, nor his horror when the table reads “only” .10 or .06. One comes to internaliz the difference between .05 and .06 as “right” vs. “wrong,” “credible” vs. “embarrassing,” “success” vs. “failure.” Tradition notwithstanding, there seems to be little justifiable reason for such a state of affairs. We find it hard to believe that social scientists simply wish to avoid the inconvenience of selecting significance level as one of the parameters of the problem under investigation.

Most textbooks in research methods and statistics recommend that the levels of significance associated with tests on the data be set in advance of actual data collection. In part, this decision is one of minimizing the probability of committing Type I or Type II error. (Type I error occurs when one rejects the null hypothesis which in fact is true. Type II error is committed when one fails to reject the null hypothesis which in fact is false.) Since the two types of errors are inversely related to each other, it is impossible to minimize both of them at the same time without increasing the sample size.

2. The use of the arbitrary .05 level of significance seems to have originated with the statistician R. A. Fisher, who used this level with experimental situations in agriculture and biology. Subsequently it was adopted by social scientists. See R. A. Fisher (1955a), and the discussion by Quinn McNemar (1955: 64).

Therefore, it is the nature of the problem under study which ought to dictate which type of error is to be minimized.

Some social scientists feel that too much emphasis has been placed upon the level of significance of a test at the expense of power considerations. Representative of these is the managing editor of Biometrika who comments:

The frequent use of the .05 and .01 levels of significance is a matter of convention having little scientific or logical basis. When the power of tests is likely to be low under these levels of significance, and when Type I and Type II errors are of approximately equal importance, the .30 and .20 levels of significance may be more appropriate than the .05 and .01 levels (Winer, 1962: 13).

It is appropriate to consider at this point an example illustrating the dependence of optimal alpha level upon the problem investigated. A social scientist may face the problem of whether to recommend a group of college seniors having high I.Q.'s but poor academic indexes, for graduate school. Controlling other relevant variables, he sets up $H_0$: There will be no difference in rate of achievement in graduate school between individuals with (a) high I.Q.'s but low grades, and (b) high I.Q.'s and outstanding grades. A decision must be made regarding the circumstances under which $H_0$ may be rejected. If a low level of alpha, say, .001 is established, the probability of unfairly discriminating in favor of the good students is minimized. On the other hand, $\alpha = .001$ implies a much greater probability of sending students to graduate school who could not do the work than would be the case with, e.g., $\alpha = .25$.

In our opinion, there is no “right” or “wrong” level here—the decision must be made in full consideration of parameters inherent in the problem itself. It is doubtful that setting a priori levels of .05, .01, or what have you settles the matter.

The issue becomes clearer when we compare the above problem with another. A decision may be needed, for example, whether to recommend that military personnel in combat zones are more effective when fighting alone or when accompanied by a “buddy.” Again, controlling other relevant variables, $H_0$ states: There will be no difference in the combat effectiveness of military personnel fighting alone, as opposed to fighting with a “buddy.” What alpha level is optimal for this hypothesis? In contrast with the first problem, a higher level (say, .20) would be justified, that is, we feel it is more desirable to risk rejection of the null hypothesis when it is true. In terms of the example, it is more desirable to place combat personnel with buddies when in fact they are no more effective fighters this way.

In the first example (college students with poor grades), it seems best to minimize Type I error, while in the second case (combat effectiveness) it is more important to minimize Type II error. Whether one agrees with the values we have employed in interpreting these examples or not is of little
importance. The point is that blind adherence to the .05 level denies any consideration of alternative strategies, and it is a serious impediment to the interpretation of data.

There are those, of course, who may feel that the cases we have cited are essentially moot problems, since they believe social scientists should deal only in "pure" and not "applied" research. This view contends that one should design, execute, and report upon his research, but not serve to implement his findings. Only in his role as "moral man" should the scientist-as-layman take responsibility for practical decisions affecting the community. Where one stands on the "pure or applied" issue is also of little consequence for this report. Either view implies a responsible consideration of the various factors underlying decisions of significance levels.

Even if one accepts the pure research orientation there are still important reasons for choosing optimal alpha levels. Blalock, for example, points out that usually the only real decision facing a social scientist is whether to publish or suppress his findings (Blalock, 1960: 135). In this circumstance he suggests a rule of thumb: "The researcher should lean over backwards to prove himself wrong or to obtain results that he usually does not want to obtain" (Blalock, 1960: 125). He warns, however, that this too may involve the decision to minimize chances of making Type I or Type II error depending on the type of problem. It is inadvisable, therefore, in any type of problem, to state a priori the "most appropriate" level of significance. The risk of so doing may be evinced by the paucity of criteria or guidelines for establishing levels of significance for different classes of problems.

Blalock's remarks are most appropriate for research which is designed to test hypotheses. There is, of course, much social science research which seeks to develop hypotheses rather than attempt verification. The rubrics pilot study and exploratory research are well known, and frequently provide for the collection and analysis of data not "explained" in advance by an explicit theoretical framework. Past experience has proven this to be an effective strategy and often such serendipitous findings outweigh the original purpose of the research (see Merton, 1957: 103-108). In cases where the goal is generation rather than testing hypotheses, it would seem advisable to tailor level of significance to the open-ended character of research design. Again, an arbitrary .05 level may not be optimal.

With the increasing involvement of social scientists in the very structures and processes they study, e.g. complex organizations, ethnic relations, and political activity, it is expected that published research findings will be disseminated to audiences of laymen untrained in the interpretation of data. When the researcher places an asterisk (or maybe even two asterisks) after his obtained F, Chi-square, or t value, and thereby decrees significance, this may constitute the basis for social action by statistically unsophisticated decision makers. On the other hand, failure to achieve significance at the prearranged level may lead to immediate dismissal of the hypothesized relationship.

As social scientists increasingly publish research findings in scholarly journals having a wider circulation than the scientific community of their own particular discipline, this problem becomes of increasing importance. Journals of practicing professions and occupations often reach an audience who is relatively unsophisticated in scientific methodology and techniques of statistical analysis. For example, in the area of medical sociology a study concerning interaction and communication in a hospital, say between nurses and patients, published in a sociological journal might reach anywhere from 2,000-12,000 subscribers, mostly sociologists and other social scientists. The same findings published in the American Journal of Nursing would come to the attention of more than 160,000 registered nurses, a portion of whom might try to make practical use of findings which are termed "significant." It is not at all improbable that this may be characteristic of many other areas of specialization.

Finally, we would like to suggest that a more suitable procedure for setting and reporting levels of significance might be: (1) For the social scientist to reflect upon the arbitrary nature of .05, .01, and .001 levels now in common use, and to recognize that selection of one of these is not a panacea for the interpretation of evidence. Different classes of research problems (or components thereof) may require different levels of alpha. (2) For the social scientist to recall that even well educated lay groups are inexpert in the interpretation of statistical significance. The demand by such groups upon professional researchers must be met by a response that is ethical and communicative. (3) For social scientists reporting research findings (especially in journals indigenous to their discipline) to do away with arbitrary levels of significance and the calling of one test result "significant" and another "not significant." We recommend a procedure whereby the actual level of significance associated with each research finding be stated (Siegel, 1956: 9). In other words, if a
18. Common Misinterpretations of Significance Levels in Sociological Journals

THOMAS J. DUGGAN and CHARLES W. DEAN

Periodically, the uses and misuses of probability statistics in social and behavioral science research have been reviewed. For instance, Lewis and Burke (1949) pointed to several misuses of the Chi-square test and an article by Selvin (1957 [Chapter 9]) stimulated discussion on the general question of using statistics in social surveys. Most recently, Skipper, Guenther, and Nass (1967 [Chapter 17]) reviewed the discussion of substantive interpretation associated with significant levels. While such discussions have served to clarify some of the technical requirements and have corrected some of the misunderstandings often associated with the use of statistical tests, one crucial matter has received relatively little attention. This concerns the substantive interpretation of significance tests and the consequences of such interpretations.

The frequently used Chi-square test and the interpretations given to data analyzed by this statistic will serve to illustrate the problem. This statistic can be used to test goodness of fit or independence although it is the latter which is more frequently used in reporting research. Since this is a test of the independence of variables, significant values of Chi-square are often taken to indicate a dependence or relationship between variables. In interpreting such relationships, there are two serious problems which are often overlooked. The first concerns the strength of relationship and the second, the form of relationship.