DETERMINANCY VERSUS RISK: A CRITIQUE OF CONTEMPORARY STATISTICAL METHODOLOGY IN SOCIOLOGY

L. Keith Miller
University of Kansas

The contention of this paper is that contemporary sociology has been misled into relying on tests of significance to the virtual exclusion of measures of association. A glance at any recent issue of the *American Sociological Review* will show that the bulk of the research is analyzed by the computation and interpretation of levels of significance. Further, even if a measure of association is computed it is usually analyzed in terms of a test of significance rather than in terms of its own numerical value. This reliance means that our discipline annually discovers and reaffirms the probable existence of many interesting relationships; however, it also means that the strengths of these relationships are not known. What are the consequences of this methodological policy? Gouldner states that there is a

> fairly widespread tendency among sociologists . . . to rest content with a demonstration that some sociological variable "makes a difference." If a variable can be shown to control even the smallest proportion of the variance in a problematic pattern, it is all too readily regarded as a memorable contribution to sociology and all too ceremoniously ushered into its theoretical hall of fame. . . . Unless sustained interest is manifested in the *degree* of variance which a variable controls, and, unless, further, we can identify sociological variables that certifiably control *substantial* proportions of variance in specified patterns of human behavior, sociology will remain scientifically immature and practically ineffectual.

One consequence of this policy is to help sociology find a large body of impotent relationships. To the extent that sociological theory is based

on and includes weak relationships, it must itself be weak. If such impotency is indeed a consequence of our contemporary methodology, then one must conclude that it should be either modified or abandoned.

Has our methodology led us to accept and dwell upon weak relationships? Is a "significant" result good evidence that a researcher has indeed stumbled upon a reasonably strong relationship? After all, if it had not been significant this would strongly suggest that there is no relationship. Hence, tests of significance divide relationships objectively into the two categories of "some relationship" and "no relationship". Unfortunately, "some" might mean either a very strong or a very weak relationship, so that this by no means settles the question. The remainder of the paper will approach this question from two directions. First, the logic behind tests of significance will be examined and it will be shown that there is no necessary relationship between level of significance and strength of relationship. Second, application of a standardized measure of the strength of recently published "significant" relationships reveals that tests of significance discover relationships which are distinctly weak.

An Evaluation of Statistical Risk

The probability calculated in tests of significance is a function of two independent factors — the determinancy of the relationship and the number of observations. Consequently, the inference that a "significant" relationship exists between two variables may arise from the combination of a very weak relationship and a large sample, a very strong relationship and a small sample, or, of course, a strong relationship and a large sample. On the other hand, the inference that a significant relationship does not exist between two variables may arise from a strong relationship in a too small sample, a weak relationship in a sample not quite large enough, or, of course, a weak relationship in a small sample.

Everyone knows these facts, yet, they are widely ignored. We always conclude that our hypothesis is strongly confirmed if we obtain a 1 per cent level of significance, and conclude that a 10 per cent level is reason for emphatically rejecting our hypothesis. Yet, these conclusions simply are not justified. What is justified is the conclusion that there is at least some relationship carrying only a small risk in the one case, and, by convention, too large a risk in the other. In fact, the hypothesis may have received very strong support in the 10 per cent risk case — that is, the data may indicate that there is a very high degree of determinancy — but the sample size may be so small that the finding must be adjudged unreliable, and the indicated conclusion risky. On the other hand, the hypothesis may have received only the weakest support in the

1 per cent case, as indicated by little determinancy, but the sample size may be so excessively large — not unusual in sociological research — that the trivial relationship is highly reliable.

In other words, the fact that there are two independent factors which determine the level of significance in a given case has been ignored, the level of significance being equated with the determinancy of relationship. As a result, though we have seemed to have been piling up an impressive amount of knowledge about the determinants of various social patterns, it is possible that all we have piled up is a collection of weak relationships imbedded in huge samples.

But, some hardy optimist may object, that in spite of our pre-occupation with tests of significance and our erroneous equation of significance level with determinancy, it is quite likely that many, even most, of our findings point to strong and consequently important relationships. After all, determinancy *is* one of the components of significance level. In order to evaluate our optimistic friend's suggestion, let us develop a standardized yardstick to measure the strength of the relationship between two variables. We may, then, apply this to the "significant" findings reported in our journal articles in recent years and discover how important these significant findings actually are. First let us imagine a yardstick which varies between zero and one corresponding to the strength of the relationship between two variables. Further, suppose that this yardstick is a ratio scale so that a relationship whose determinancy is .25 is only half as strong as one that is .50. This yardstick which we shall dub the "coefficient of determinancy", may be interpreted as measuring the percentage of variation in one variable which is predictable from another variable. In other words, this yardstick tells us the percentage of one variable which may be statistically determined by another variable.

Next let us decide upon what we consider to be a "strong" relationship on this scale. Sociologists, when asked this question, give answers ranging anywhere from 30 per cent to as high as 90 per cent determinancy. The modal value seems to be around 50 per cent. And, one must admit that this would, indeed, be a strong relationship. However, since our yardstick is the square of the correlation coefficient (under those conditions for which correlations are appropriate), some will argue for lowering of this figure. A correlation of .30, and, thus, a yardstick value of 9 per cent is considered respectable. It seems reasonable to argue that this is the very lowest level at which we might consider a relationship to have even tolerable importance. Let us assume, then, that a relationship which determines 10 per cent of the variation is weak but tolerable, that a relationship that determines 30 per cent is quite good, and that one that determines 50 per cent is very good.

Let us apply our yardstick to some relationships, which have been subjected to statistical manipulation and reported as significant. Table I reports the results of just such an examination of 99 tests of significance reported in the first three issues of the 1961 *American Sociological Review.* The yardstick may be directly applied to any data in a form usable by the major statistical tests, Chi-square, t-test, analysis of variance, and correlation analysis. For purposes of comparison the results are reported separately for each type of test.

We find that 41 Chi-squares have found significant relationships which determine only 8 per cent of the data. Furthermore, the single highest one accounts for only 25 per cent of the data. That is, the single strongest relationship discovered by Chi-square analysis left over 75 per cent of the significant datum unexplained. Perhaps, the results from most of this research, say the 66 per cent which did not explain at least 10 per cent of the data, should have led to a rejection of the theories and hypotheses presented in the articles on the grounds that they are so weak that they are almost completely unimportant.

Research analyzed by other popular methods seems to fare only slightly better. Nine analyses of variance account for only 16 per cent of the variation. Of the nine, however, only one third do not meet the minimal level of 10 per cent. The 49 correlation coefficients single out variables which explain on the average under 12 per cent of the variation. In this case over half do not attain the minimal level.

Since all of these figures are based on a ratio scale with the same meaning, we can compute an overall average of the strength of relationships which have been discovered in these statistically significant results. This computation shows that all of the research reported in the first half of 1961, and which has been adjudged statistically significant, accounts for less than 11 per cent of the variation in the social patterns studied. It is difficult not to draw the same conclusion for this data as was drawn for the Chi-square tests; much of it should have been used to conclude that the theories and hypotheses under consideration were of very little, or no importance.

Before concluding that tests of significance by themselves are totally insufficient, the cautious will want to examine the data still further. Our optimist might suggest that those relationships singled out by a high significance level may be stronger than those singled out by a low significance level, thus allowing us to differentiate between these relationships. This suggestion may be easily checked by comparing the strength of relationships judged significant at each of the acceptable levels to discover if their strengths vary systematically as a function of significance. Table II indicates that they do not. While the average relationship significant at the .1 per cent level determines about 13 per cent of the data, the average relationship significant at the 5 per cent level determines 15 per cent. Furthermore, those at the 1 per cent level

determine only 4 per cent and those at the 2 per cent level 7 per cent of the data. The fact that our yardstick is not related to level of significance is simply a further illustration that significance is highly dependent upon sample size.

Conclusions

The foregoing discussion lends force to Gouldner's criticism of our contemporary methodology. Tests of significance lump the weak, moderate and strong relationships together in one "significant" basket, having only the virtue of excluding the nullest of null relationships. The coefficient of determinancy appears to be a much better measure of relatedness. Nevertheless, there are several objections which may be raised to using the coefficient of determinancy.

While it may be plausible to argue that the low level of the average determinancy in the significant relationships studied in this paper is due to the fairly primitive stage of our scientific development, the indiscriminate lumping of weak and strong relationships cannot be so excused. After all, some of the relationships included in average are quite high, at least a few over 30 per cent. Surely it is of great importance to be able to distinguish these relationships from the quite numerous ones which determine less than 2 per cent.

It may also be plausible to argue that merely citing the coefficient of determinancy does not allow us to make a "decision" about our relationship; there is no criterion to guide us. Indeed, this is one seductive property of tests of significance: you either accept or reject the hypothesis. With the coefficient, however, there is a uniform gradation with only the barest guideposts of weak, moderate and strong relationships along the way. Actually, of course, we are still making a decision, only it is a more complicated decision. Within the framework of determinancy analysis we estimate the strength of the relationship and then decide whether it is strong enough to warrant further attention and inclusion in our theories.

But how is this decision made? It should be clear that we can no longer make a purely "statistical" decision about a relationship. Rather, our decision must be based upon theoretical considerations. For example, we might ask "what is the state of our theory in this area? " If it is nil, that is if we know virtually nothing about what are important determinants of social patterns in a particular area, then we should want a rather high degree of determinancy so that the theory can really be said to explain something. If the area is already well mapped, however, then a relationship exhibiting a rather small determinancy will be of considerable significance. Thus if we find a relationship which accounts for 5 per cent of the variation in group anomie and we know nothing about the determinants of anomie, we should not be very ready

to get excited. If, on the other hand, we already can account for 80 per cent of the variation in anomie, another 5 per cent is quite significant and might lead to a vast modification of theory. At any rate, the question of significance in a non-statistical sense becomes infinitely more complicated within the framework suggested by this paper.

Certainly other objections will be forthcoming to this framework but perhaps these two are of some importance. If so it may be fair for me to stop anticipating the reactions of the community and to state my own conclusions. They are that the coefficient of determinancy should routinely accompany the use of tests of significance and that the coefficient should be the chief criterion for determining whether the data support any particular hypothesis. The maximal usefulness of the coefficient is attendant upon the solution of several problems. First, a general definition of the coefficient must be developed which applies to the widest possible range of measurement and design. Unbiased estimates of this coefficient must then be developed along with tests of significance. Fortunately all of the elements of a solution to this problem already exist. In fact the data in Table I are based upon these unbiased estimates and the usual tests of significance apply unaltered. Furthermore, the same computations are used to derive the results. The only work that remains is to state the general definition and tie these strands together. Second, the general definition alluded to above, while capable of operational definition for both nominal and equal interval scales, does not apply to ordinal scales. The parameters estimated in the usual "non-parametric" tests of significance do not relate to the general coefficient of determinancy. Consequently, either a new statistic must be invented or some other way must be invented to bring ordinal scales into the fold. I feel that this latter tack is the most fruitful. I would argue that any ordinal scale can be transformed into a scale along which an appropriate population will be normally distributed. This being the case, all of the assumptions necessary for the use of parametric statistics hold and they may be applied. In short, a future paper shall argue that ordinal scales may be transformed so that they lose their obnoxious properties.

The general definition ordinarily assumes that the variables are measured without error. A number of techniques have been developed which permit the correction for attenuation. It may be possible that one reason for the low average determinancy in Table I of this paper was a result of attenuation due to measurement error. If this is the case, this subject deserves concerted attention so that we may compensate for error and get a more just idea of the power of our theories.

There is one hitch to the above points: they apply only to the relationships between two variables and a narrow range of multiple variable cases. In studies resulting from a careful experimental design, unbiased estimates of the coefficient of determinancy are available

regardless of the number of variables. However, in studies wherein multiple correlation techniques are typically used, the situation is almost hopeless. In this case, quite general in field and survey research, the strength of the relationships are hopelessly entangled with the strength of their interactions. This asymmetry between field and experimental studies also deserves considerable attention. It raises a rather serious question about the relationship between field and experimental research. At any rate, these problems must be tackled before determinancy analysis can attain its full usefulness. The first problem is solved; the second and third should not be very difficult; the fourth problem will probably require a good deal of work. The promise of the method would seem to justify our attention.

TABLE I DETERMINANCY (Dx) of SIGNIFICANT RELATIONSHIPS REPORTED IN ARTICLES PUBLISHED IN THE FIRST THREE ISSUES OF THE *AMERICAN SOCIOLOGICAL REVIEW* of 1961.*

Original Statistic	Nb of Studies	Nb of Tests	Dx	Range in Dx Low	Range in Dx High	% where Dx > .10
Chi-square	5	41	.083[a]	.009	.250	34.1%
F-ratio	4	9	.159[b]	.003	.578	66.7%
Correlation	2	49	.117[c]	.001	.518	46.7%
All tests	12	99	.107	.001	.578	42.4%

*The data in the table are based on all articles reporting statistical tests of significance which appeared in the February, April, and June issues of the *American Sociological Review.* Due to its complicated structure the analysis of variance reported in Anderson, T. R. and Egeland, J. A. "Spatial Aspects of Social Area Analysis," was not included. In order to provide a somewhat firmer basis from which to generalize the lone correlation study in this period was augmented by the first non-demographic correlational study appearing prior to this time: Martin, J. G. and Westie, F. R. "The Tolerant Personality," A.S.R.: 1959, (24 August), pp. 521-528. This addition balanced the Seeman article in which the average r^2 was .028 since its average was .210. The overall average of about 12 per cent is probably not unrepresentative of correlation studies in sociology. At any rate our inclusion is biased opposite to our main argument.

[a] An unbiased estimate of the "coefficient of constraint" forms the

basis for these figures. This is the ratio of uncertainty controlled by the independent variable, termed "transmitted information" by Shannon, to the uncertainty in the dependent variable, again measured by the usual logarithmic information measure. This measure was chosen because it is identical in structure and meaning to the other measures used in this paper and applying it to normally distributed equal interval scales would produce results identical to an analysis of variance. For further information consult: McGill, W. J. "Isomorphism in Statistical Analysis" and G. A. Miller, "Note on the Bias of Information Measures," both in Quastler, H. (ed.) *Information Theory in Psychology.* Glencoe, Ill.: The Free Press, 1955; and also Senders, V. *Measurement and Statistics.* N.Y.: The Oxford University Press, 1958, pp. 89-98.

[b] These figures are based on unbiased estimators of the variance in the dependent variable attributable to the independent variable and to the total variance of the dependent variable. These estimators were derived from what is called "components of variance analysis." The percentages are simply the ratio of these two terms. The best introduction to this ratio is given by Haggard in his book on *Intraclass Correlation and the Analysis of Variance.* N.Y.: The Dryden Press, 1958.

[c] These figures are based on biased estimators of the variance in the dependent variable which is attributable to the independent variable and the total variance in the dependent variable. The ratio of these variances is given by the square of the correlation coefficient. Unfortunately, unbiased estimates could not be derived from the information made available in the articles. However, they may also be derived from component of variance analysis. This is discussed in Acton's book, *Analysis of Straight Line Data.* N.Y.: John Wiley and Sons, 1959.

TABLE II THE RELATIONSHIP BETWEEN RISK AND STATISTICAL DETERMINANCY FOR RELATIONSHIPS SIGNIFICANT BEYOND THE 5 PER CENT LEVEL.

Null hypothesis rejected at	5%	2%	1%	.1%	Total
Percent Determinancy	.15,4	.07,0	.03,9	.13,3	.10,7
N	41	10	29	19	99

VERSTEHEN AND EXPLANATION

Don Marquis
University of Kansas

The use of *Verstehen* in sociology has had a checkered history. Indeed one might think that the "operation called *Verstehen*" had been laid to rest twenty years ago by an oft-cited article of that title by Theodore Abel. (1) Abel concludes that *Verstehen* has to do with relating the behavior of others to "our personal experiences". (1, 685) Accordingly:

> Primarily the operation of *Verstehen* does two things: It relieves us of a sense of apprehension in connection with behavior that is unfamiliar or unexpected and it is a source of "hunches", which can help us in the formulation of hypotheses. (1, 687)

Professor Ernest Nagel, at the conclusion of a discussion of "meaningful" explanations of human behavior, concludes that:

> (The social scientist's) ability to enter into relations of empathy with the human actors in some social process may indeed be heuristically important in his efforts to *invent* suitable hypotheses which will explain the process. Nevertheless, his empathetic identification with these individuals does not, by itself, constitute knowledge. (2, 484)

Professor Murray Wax has defended the use of *Verstehen* in sociology in a recent article. Wax argues that Abel's criticisms are misguided because Abel has misunderstood what *Verstehen* is. According to Wax:

> Following the theoretical lead of Weber (cited above), he (Abel) places the emphasis (of *Verstehen*) upon the imputation of motive which, and here he is quite correct, he sees as a difficulty in this kind of case, but he misses the point that the true level of *Verstehen* involved here is far deeper and more primitive. (3, 326)

Not only does the "real" *Verstehen* not involve imputation of motive, but, according to Wax, it also does not involve "interpersonal