



Some Statistical Problems in Research Design

Leslie Kish

American Sociological Review, Vol. 24, No. 3. (Jun., 1959), pp. 328-338.

Stable URL:

<http://links.jstor.org/sici?sici=0003-1224%28195906%2924%3A3%3C328%3ASSPIRD%3E2.0.CO%3B2-A>

American Sociological Review is currently published by American Sociological Association.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/about/terms.html>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/journals/asa.html>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.

from a facet design based on social psychological consideration. "Blind" analyses are now known to be essentially incapable of revealing ordered structures, such as the simplex, even when the order is strikingly apparent to the eye and is the most parsimonious way of viewing the data.

(b) At least two different rotations of axes are meaningful for our case: equations (3)–(6) and equations (7)–(10). In each case, our factors are "named" in advance. A third rotation—to principal axes—may also prove meaningful eventually, when the appropriate psychology is worked out, as discussed in the references of footnotes 4 and 10. But our structural conclusions depend on none of these: the additivity of distance functions is all we need and this requires no particular location of reference axes.

(c) Even after eliminating unique factors, we have left as many common-factors as observed variables, as in equations (7)–(10), or even *more* common-factors than observed variables,

as in equation (3)–(6). New light is cast by radex theory on the communality problem of factor analysis: the assumption that parsimony is equivalent to smallness of number of common-factors has been shown to be unfounded.¹³

(d) Using variances as distance functions, instead of standard deviations, implies using a non-Euclidean metric for the variables. The variables lie in one dimension—along a straight line, and in the semantic scale order—when our non-Euclidean metric is used; but they define an n-dimensional space—where n is the number of common-factors—when the Euclidean metric is used.

The orthogonality conditions between α , β , and γ lay no restrictions on their covariances with t_i in pattern (7)–(10); t_i may be oblique or orthogonal to each of these latter factors, without affecting the additivity of distances.

¹³ *Ibid.*; and Louis Guttman, "To What Extent Can Communalities Reduce Rank?" *Psychometrika*, 23 (December, 1958), pp. 297–308.

SOME STATISTICAL PROBLEMS IN RESEARCH DESIGN *

LESLIE KISH

University of Michigan

Several statistical problems in the design of research are discussed: (1) The use of statistical tests and the search for causation in survey research are examined; for this we suggest separating four classes of variables: explanatory, controlled, confounded, and randomized. (2) The relative advantages of experiments, surveys, and other investigations are shown to derive respectively from better control, representation, and measurement. (3) Finally, three common misuses of statistical tests are examined: "hunting with a shot-gun for significant differences," confusing statistical significance with substantive importance, and overemphasis on the primitive level of merely finding differences.

STATISTICAL inference is an important aspect of scientific inference. The statistical consultant spends much of his time in the borderland between statistics and the other aspects, philosophical and substantive, of the scientific search for explanation. This marginal life is rich both in direct experience and in discussions of fundamentals; these have stimulated my concern with the problems treated here.

I intend to touch on several problems dealing with the interplay of statistics with the more general problems of scientific in-

ference. We can spare elaborate introductions because these problems are well known. Why then discuss them here at all? We do so because, first, they are problems about which there is a great deal of misunderstanding, evident in current research; and, second, they are *statistical* problems on which there is broad agreement among research statisticians—and on which these statisticians generally disagree with much in the current practice of research scientists.¹

¹ Cf. R. A. Fisher, *The Design of Experiments*, London: Oliver and Boyd, 6th edition, 1953, pp. 1–2: "The statistician cannot evade the responsibility for understanding the processes he applies or recommends. My immediate point is that the questions involved can be disassociated from all that is strictly technical in the statistician's craft, and *when so detached*, are questions only of the right use of human reasoning powers, with which all intel-

* This research has been supported by a grant from the Ford Foundation for Development of the Behavioral Sciences. It has benefited from the suggestions and encouragement of John W. Tukey and others. But the author alone is responsible for any controversial opinions.

Several problems will be considered briefly, hence incompletely. The aim of this paper is not a profound analysis, but a clear elementary treatment of several related problems. The footnotes contain references to more thorough treatments. Moreover, these are not *all* the problems in this area, nor even necessarily the most important ones; the reader may find that his favorite, his most annoying problem, has been omitted. The problems selected are a group with a common core, they arise frequently, yet they are widely misunderstood.

STATISTICAL TESTS OF SURVEY DATA

That correlation does not prove causation is hardly news. Perhaps the wittiest statements on this point are in George Bernard Shaw's preface to *The Doctor's Dilemma*, in the sections on "Statistical Illusions," "The Surprises of Attention and Neglect," "Stealing Credit from Civilization," and "Biometrika." (These attack, alas, the practice of vaccination.) The excellent introductory textbook by Yule and Kendall² deals in three separate chapters with the problems of advancing from correlation to causation. Searching for causal factors among survey data is an old, useful sport; and the attempts to separate true explanatory variables from extraneous and "spurious" correlations have taxed scientists since antiquity and will undoubtedly continue to do so. Neyman and Simon³ show that beyond common sense, there are some technical skills involved in tracking down spurious correlations. Econometricians and geneticists have developed great interest and skill in the prob-

ligent people, who hope to be intelligible, are equally concerned, and on which the statistician, as such, speaks with no special authority. The statistician cannot excuse himself from the duty of getting his head clear on the principles of scientific inference, but equally no other thinking man can avoid a like obligation."

² G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics*, London: Griffin, 11th edition, 1937, Chapters 4, 15, and 16.

³ Jerzy Neyman, *Lectures and Conferences on Mathematical Statistics and Probability*, Washington, D. C.: Graduate School of Department of Agriculture, 1952, pp. 143-154. Herbert A. Simon, "Spurious Correlation: A Causal Interpretation," *Journal of the American Statistical Association*, 49 (September, 1954), pp. 467-479; also in his *Models of Man*, New York: Wiley, 1956.

lems of separating the explanatory variables.⁴

The researcher designates the explanatory variables on the basis of substantive scientific theories. He recognizes the evidence of other *sources of variation* and he needs to separate these from the explanatory variables. Sorting all sources of variation into four classes seems to me a useful simplification. Furthermore, no confusion need result from talking about sorting and treating "variables," instead of "sources of variation."

I. The *explanatory* variables, sometimes called the "experimental" variables, are the objects of the research. They are the variables among which the researcher wishes to find and to measure some specified relationships. They include both the "dependent" and the "independent" variables, that is, the "predictand" and "predictor" variables.⁵ With respect to the aims of the research all other variables, of which there are three classes, are extraneous.

II. There are extraneous variables which are *controlled*. The control may be exercised in either or both the selection and the estimation procedures.

⁴ See the excellent and readable article, Herman Wold, "Causal Inference from Observational Data," *Journal of the Royal Statistical Society (A)*, 119 (Part 1, January, 1956), pp. 28-61. Also the two-part technical article, M. G. Kendall, "Regression, Structure and Functional Relationship," *Biometrika*, 38 (June, 1951), pp. 12-25; and 39 (June, 1952), pp. 96-108. The interesting methods of "path coefficients" in genetics have been developed by Wright for inferring causal factors from regression coefficients. See, in Oscar Kempthorne *et al.*, *Statistics and Mathematics in Biology*, Ames, Iowa: The Iowa State College Press, 1954; Sewall Wright, "The Interpretation of Multi-Variate Systems," Chapter 2; and John W. Tukey, "Causation, Regression and Path Analysis," Chapter 3. Also C. C. Li, "The Concept of Path Coefficient and Its Impact on Population Genetics," *Biometrics*, 12 (June, 1956), pp. 190-209. I do not know whether these methods can be of wide service in current social science research in the presence of numerous factors, of large unexplained variances, and of doubtful directions of causation.

⁵ Kendall points out that these latter terms are preferable. See his paper cited in footnote 4, and M. G. Kendall and W. R. Buckland, *A Dictionary of Statistical Terms*, Prepared for the International Statistical Institute with assistance of UNESCO, London: Oliver and Boyd, 1957. I have also tried to follow in IV below his distinction of "variate" for random variables from "variables" for the usual (nonrandom) variable.

III. There may exist extraneous uncontrolled variables which are *confounded* with the Class I variables.

IV. There are extraneous uncontrolled variables which are treated as *randomized* errors. In "ideal" experiments (discussed below) they are actually randomized; in surveys and investigations they are only assumed to be randomized. Randomization may be regarded as a substitute for experimental control or as a form of control.

The aim of efficient design both in experiments and in surveys is to place as many of the extraneous variables as is feasible into the second class. The aim of randomization in experiments is to place all of the third class into the fourth class; in the "ideal" experiment there are no variables in the third class. And it is the aim of controls of various kinds in surveys to separate variables of the third class from those of the first class; these controls may involve the use of repeated cross-tabulations, regression, standardization, matching of units, and so on.

The function of statistical "tests of significance" is to test the effects found among the Class I variables against the effects of the variables of Class IV. An "ideal" experiment here denotes an experiment for which this can be done through randomization without any possible confusion with Class III variables. (The difficulties of reaching this "ideal" are discussed below.) In survey results, Class III variables are confounded with those of Class I; the statistical tests actually contrast the effects of the random variables of Class IV against the explanatory variables of Class I confounded with unknown effects of Class III variables. In both the ideal experiment and in surveys the statistical tests serve to separate the effects of the random errors of Class IV from the effects of other variables. These, in surveys, are a mixture of explanatory and confounded variables; their separation poses severe problems for logic and for scientific methods; statistics is only one of the tools in this endeavor. The scientist must make many decisions as to which variables are extraneous to his objectives, which should and can be controlled, and what methods of control he should use. He must decide where and how to introduce statistical tests of hypotheses into the analysis.

As a simple example, suppose that from a probability sample survey of adults of the United States we find that the level of political interest is higher in urban than in rural areas. A test of significance will show whether or not the difference in the "levels" is large enough, compared with the sampling error of the difference, to be considered "significant." Better still, the confidence interval of the difference will disclose the limits within which we can expect the "true" population value of the difference to lie.⁶ If families had been sent to urban and rural areas respectively, after the randomization of a true experiment, then the sampling error would measure the effects of Class IV variables against the effects of urban *versus* rural residence on political interest; the difference in levels beyond sampling errors could be ascribed (with specified probability) to the effects of urban *versus* rural residence.

Actually, however, residences are not assigned at random. Hence, in survey results, Class III variables may account for some of the difference. If the test of significance rejects the null hypothesis of no difference, *several* hypotheses remain in addition to that of a simple relationship between urban *versus* rural residence and political interest. Could differences in income, in occupation, or in family life cycle account for the difference in the levels? The analyst may try to remove (for example, through cross-tabulation, regression, standardization) the effects due to such variables, which are extraneous to his expressed interest; then he computes the difference, between the urban and rural residents, of the levels of interest now free of several confounding variables. This can be followed by a proper test of significance—or, preferably, by some other form of statistical inference, such as a statement of confidence intervals.

Of course, other variables of Class III may remain to confound the measured relationship between residence and political interest.

⁶ The sampling error measures the chance fluctuation in the difference of levels due to the sampling operations. The computation of the sampling error must take proper account of the actual sample design, and not blindly follow the standard simple random formulas. See Leslie Kish, "Confidence Intervals for Complex Samples," *American Sociological Review*, 22 (April, 1957), pp. 154-165.

The separation of Class I from Class III variables should be determined in accord with the nature of the hypothesis with which the researcher is concerned; finding and measuring the effects of confounding variables of Class III tax the ingenuity of research scientists. But this separation is beyond the functions and capacities of the statistical tests, the tests of null hypotheses. Their function is not explanation; they cannot point to causation. Their function is to ask: "Is there anything in the data that needs explaining?"—and to answer this question with a certain probability.

Agreement on these ideas can eliminate certain confusion, exemplified by Selvin in a recent article:

Statistical tests are unsatisfactory in non-experimental research for two fundamental reasons: it is almost impossible to design studies that meet the conditions for using the tests, and the situations in which the tests are employed make it difficult to draw correct inferences. The basic difficulty in design is that sociologists are unable to randomize their uncontrolled variables, so that the difference between "experimental" and "control" groups (or their analogs in nonexperimental situations) are a mixture of the effects of the variable being studied and the uncontrolled variables or correlated biases. Since there is no way of knowing, in general, the sizes of these correlated biases and their directions, there is no point in asking for the probability that the observed differences could have been produced by random errors. The place for significance tests is after all relevant correlated biases have been controlled. . . . In design and in interpretation, in principle and in practice, tests of statistical significance are inapplicable in nonexperimental research.⁷

Now it is true that in survey results the explanatory variables of Class I are confounded with variables of Class III; but it does not follow that tests of significance should not be used to separate the random variables of Class IV. Insofar as the effects found "are a mixture of the effects of the variable being studied and the uncontrolled

⁷ Hanan C. Selvin, "A Critique of Tests of Significance in Survey Research," *American Sociological Review*, 22 (October, 1957), p. 527. In a criticism of this article, McGinnis shows that the separation of explanatory from extraneous variables depends on the type of hypothesis at which the research is aimed. Robert McGinnis, "Randomization and Inference in Sociological Research," *American Sociological Review*, 23 (August, 1958), pp. 408-414.

variables;" insofar as "there is no way of knowing, in general, the sizes" and directions of these uncontrolled variables, Selvin's logic and advice should lead not only to the rejection of statistical tests; it should lead one to refrain altogether from using survey results for the purposes of finding explanatory variables. *In this sense*, not only tests of significance but any comparisons, any scientific inquiry based on surveys, any scientific inquiry other than an "ideal" experiment, is "inapplicable." That advice is most unrealistic. In the (unlikely) event of its being followed, it would sterilize social research—and other nonexperimental research as well.

Actually, much research—in the social, biological, and physical sciences—must be based on nonexperimental methods. In such cases the rejection of the null hypothesis leads to several alternate hypotheses that may explain the discovered relationships. It is the duty of scientists to search, with painstaking effort and with ingenuity, for bases on which to decide among these hypotheses.

As for Selvin's advice to refrain from making tests of significance until "after all relevant" uncontrolled variables have been controlled—this seems rather far-fetched to scientists engaged in empirical work who consider themselves lucky if they can explain 25 or 50 per cent of the total variance. The control of all relevant variables is a goal seldom even approached in practice. To postpone to that distant goal all statistical tests illustrates that often "the perfect is the enemy of the good."⁸

⁸ Selvin performs a service in pointing to several common mistakes: (a) The mechanical use of "significance tests" can lead to false conclusions. (b) Statistical "significance" should not be confused with substantive importance. (c) The probability levels of the common statistical tests are not appropriate to the practice of "hunting" for a few differences among a mass of results. However, Selvin gives poor advice on what to do about these mistakes; particularly when, in his central thesis, he reiterates that "tests of significance are inapplicable in nonexperimental research," and that "the tests are applicable only when all relevant variables have been controlled." I hope that the benefits of his warnings outweigh the damages of his confusion.

I noticed three misleading references in the article. (a) In the paper which Selvin appears to use as supporting him, Wold (*op. cit.*, p. 39) specifically disagrees with Selvin's central thesis, stat-

EXPERIMENTS, SURVEYS, AND OTHER
INVESTIGATIONS

Until now, the theory of sample surveys has been developed chiefly to provide descriptive statistics—especially estimates of means, proportions, and totals. On the other hand, experimental designs have been used primarily to find explanatory variables in the analytical search of data. In many fields, however, including the social sciences, survey data must be used frequently as the analytical tools in the search for explanatory variables. Furthermore, in some research situations, neither experiments nor sample surveys are practical, and other investigations are utilized.

By "experiments" I mean here "ideal" experiments in which all the extraneous variables have been randomized. By "surveys" (or "sample surveys"), I mean probability samples in which all members of a defined population have a known positive probability of selection into the sample. By "investigations" (or "other investigations"), I mean the collection of data—perhaps with care, and even with considerable control—without either the randomization of experiments or the probability sampling of surveys. The differences among experiments, surveys, and investigations are not the consequences of statistical techniques; they result from different methods for introducing the variables and for selecting the population elements (subjects). These problems are ably treated in recent articles by Wold and Campbell.⁹

ing that "The need for testing the statistical inference is no less than when dealing with experimental data, but with observational data other approaches come to the foreground." (b) In discussing problems caused by complex sample designs, Selvin writes that "Such errors are easy enough to discover and remedy" (p. 520), referring to Kish (*op. cit.*). On the contrary, my article pointed out the seriousness of the problem and the difficulties in dealing with it. (c) "Correlated biases" is a poor term for the confounded uncontrolled variables and it is not true that the term is so used in literature. Specifically, the reference to Cochran is misleading, since he is dealing there only with errors of measurement which may be correlated with the "true" value. See William G. Cochran, *Sampling Techniques*, New York: Wiley, 1953, p. 305.

⁹ Wold, *op. cit.*; Donald T. Campbell, "Factors Relevant to the Validity of Experiments in Social Settings," *Psychological Bulletin*, 54 (July, 1957), pp. 297-312.

In considering the larger ends of any scientific research, only part of the total means required for inference can be brought under objective and firm control; another part must be left to more or less vague and subjective—however skillful—judgment. The scientist seeks to maximize the first part, and thus to minimize the second. In assessing the ends, the costs, and the feasible means, he makes a strategic choice of methods. He is faced with the three basic problems of scientific research: measurement, representation, and control. We ignore here the important but vast problems of measurement and deal with representation and control.

Experiments are strong on control through randomization; but they are weak on representation (and sometimes on the "naturalism" of measurement). Surveys are strong on representation, but they are often weak on control. Investigations are weak on control and often on representation; their use is due frequently to convenience or low cost and sometimes to the need for measurements in "natural settings."

Experiments have three chief advantages: (1) Through randomization of extraneous variables the confounding variables (Class III) are eliminated. (2) Control over the introduction and variation of the "predictor" variables clarifies the *direction* of causation from "predictor" to "predictand" variables. In contrast, in the correlations of many surveys this direction is not clear—for example, between some behaviors and correlated attitudes. (3) The modern design of experiments allows for great flexibility, efficiency, and powerful statistical manipulation, whereas the analytical use of survey data presents special statistical problems.¹⁰

The advantages of the experimental method are so well known that we need not dwell on them here. It is the scientific method *par excellence*—when feasible. In many situations experiments are not feasible and this is often the case in the social sciences; but it is a mistake to use this situation to separate the social from the physical and biological sciences. Such situations also occur frequently in the physical sciences (in meteorology, astronomy, geology), the biological sciences, medicine, and elsewhere.

¹⁰ Kish, *op. cit.*

The experimental method also has some shortcomings. First, it is often difficult to choose the "control" variables so as to exclude *all* the confounding extraneous variables; that is, it may be difficult or impossible to design an "ideal" experiment. Consider the following examples: The problem of finding a proper control for testing the effects of the Salk polio vaccine led to the use of an adequate "placebo." The Hawthorne experiment demonstrated that the design of a proposed "treatment *versus* control" may turn out to be largely a test of *any* treatment *versus* lack of treatment.¹¹ Many of the initial successes reported about mental therapy, which later turn into vain hopes, may be due to the hopeful effects of *any* new treatment in contrast with the background of neglect. Shaw, in "The Surprises of Attention and Neglect" writes: "Not until attention has been effectually substituted for neglect as a general rule, will the statistics begin to show the merits of the particular methods of attention adopted."

There is an old joke about the man who drank too much on four different occasions, respectively, of scotch and soda, bourbon and soda, rum and soda, and wine and soda. Because he suffered painful effects on all four occasions, he ascribed, with scientific logic, the common effect to the common cause: "I'll never touch soda again!" Now, to a man (say, from Outer Space) ignorant of the common alcoholic content of the four "treatments" and of the relative physiological effects of alcohol and carbonated water, the subject is not fit for joking, but for further scientific investigation.

Thus, the advantages of experiments over surveys in permitting better control are only

¹¹ F. J. Roethlisberger and W. J. Dickson, *Management and the Worker*, Cambridge: Harvard University Press, 1939. Troubles with experimental controls misled even the great Pavlov into believing *temporarily* that he had proof of the inheritance of an acquired ability to learn: "In an informal statement made at the time of the Thirteenth International Physiological Congress, Boston, August, 1929, Pavlov explained that in checking up these experiments it was found that the apparent improvement in the ability to learn, on the part of successive generations of mice, was really due to an improvement in the ability to teach, on the part of the experimenter." From B. G. Greenberg, *The Story of Evolution*, New York: Garden City, 1929, p. 327.

relative, not absolute.¹² The design of proper experimental controls is not automatic; it is an art requiring scientific knowledge, foresight in planning the experiment, and hindsight in interpreting the results. Nevertheless, the distinction in control between experiments and surveys is real and considerable; and to emphasize this distinction we refer here to "ideal" experiments in which the control of the random variables is complete.

Second, it is generally difficult to design experiments so as to represent a specified important population. In fact, the questions of sampling, of making the experimental results representative of a specified population, have been largely ignored in experimental design until recently. Both in theory and in practice, experimental research has often neglected the basic truth that causal systems, the distributions of relations—like the distributions of characteristics—exists only within specified universes. The distributions of relationships, as of characteristics, exist only within the framework of specific populations. Probability distributions, like all mathematical models, are abstract systems; their application to the physical world must include the specification of the populations. For example, it is generally accepted that the statement of a value for mean income has meaning only with reference to a specified population; but this is not generally and clearly recognized in the case of regression of assets on income and occupation. Similarly, the *statistical* inferences derived from the experimental testing of several treatments are restricted to the population(s) included in the experimental design.¹³ The clarification of the population sampling aspects of experiments is now being tackled vigorously by Wilk and Kempthorne and by Cornfield and Tukey.¹⁴

¹² Jerome Cornfield, "Statistical Relationships and Proof in Medicine," *American Statistician*, 8 (December, 1954), pp. 19–21.

¹³ McGinnis, *op. cit.*, p. 412, points out that usually "it is not true that one can uncover 'general' relationships by examining some arbitrarily selected population. . . . There is no such thing as a completely general relationship which is independent of population, time, and space. The extent to which a relationship is constant among different populations is an empirical question which can be resolved only by examining different populations at different times in different places."

¹⁴ Martin B. Wilk and Oscar Kempthorne, "Some

Third, for many research aims, especially in the social sciences, contriving the desired "natural setting" for the measurements is not feasible in experimental design. Hence, what social experiments give sometimes are clear answers to questions the meanings of which are vague. That is, the artificially contrived experimental variables *may* have but a tenuous relationship to the variables the researcher would like to investigate.

The second and third weaknesses of experiments point to the advantages of surveys. Not only do probability samples permit clear statistical inferences to defined populations, but the measurements can often be made in the "natural settings" of actual populations. Thus in practical research situations the experimental method, like the survey method, has its distinct problems and drawbacks as well as its advantages. In practice one generally cannot solve simultaneously all of the problems of measurement, representation, and control; rather, one must choose and compromise. In any specific situation one method may be better or more practical than the other; but there is no over-all superiority in all situations for either method. Understanding the advantages and weaknesses of both methods should lead to better choices.

In social research, in preference to both surveys and experiments, frequently some design of controlled investigation is chosen—for reasons of cost or of feasibility or to preserve the "natural setting" of the measurements. Ingenious adaptations of experimental designs have been contrived for these controlled investigations. The statistical framework and analysis of experimental designs are used, but not the randomization of true experiments. These designs are aimed to provide flexibility, efficiency, and, especially, some control over the extraneous variables. They have often been used to improve con-

siderably research with controlled investigations.

These designs are sometimes called "natural experiments." For the sake of clarity, however, it is important to keep clear the distinctions among the methods and to reserve the word "experiment" for designs in which the uncontrolled variables are randomized. This principle is stated clearly by Fisher,¹⁵ and is accepted often in scientific research. Confusion is caused by the use of terms like "ex post facto experiments" to describe surveys or designs of controlled investigations. Sample surveys and controlled investigations have their own justifications, their own virtues; they are not just second-class experiments. I deplore the borrowing of the prestige word "experiment," when it cloaks the use of other methods.

Experiments, surveys, and investigations can all be improved by efforts to overcome their weaknesses. Because the chief weakness of surveys is their low degree of control, researchers should be alert to the collection and use of auxiliary information as controls against confounding variables. They also should take greater advantage of changes introduced into their world by measuring the effects of such changes. They should utilize more often efficient and useful statistics instead of making tabular presentation their only tool.

On the other hand, experiments and controlled investigations can often be improved by efforts to specify their populations more clearly and to make the results more representative of the population. Often more should be done to broaden the area of inference to more important populations. Thus, in many situations the deliberate attempts of the researcher to make his sample more "homogeneous" are misplaced; and if common sense will not dispel the error, reading Fisher may.¹⁶ When he understands this,

Aspects of the Analysis of Factorial Experiment in a Completely Randomized Design," *Annals of Mathematical Statistics*, 27 (December, 1956), pp. 950-985; and "Fixed, Mixed and Random Models," *Journal of the American Statistical Association*, 50 (December, 1955), pp. 1144-1167. Jerome Cornfield and John W. Tukey, "Average Values of Mean Squares in Factorials," *Annals of Mathematical Statistics*, 27 (December, 1956), pp. 907-949.

¹⁵ Fisher, *op. cit.*, pp. 17-20. "Controlled investigation" may not be the best term for these designs. "Controlled observations" might do, but "observation" has more fundamental meanings.

¹⁶ *Ibid.*, pp. 99-100. Fisher says: "We have seen that the factorial arrangement possesses two advantages over experiments involving only single factors: (i) Greater *efficiency*, in that these factors are evaluated with the same precision by means of only a quarter of the number of observations that would otherwise be necessary; and (ii) Greater

the researcher can view the population base of his research in terms of efficiency—in terms of costs and variances. He can often avoid basing his research on a comparison of one sampling unit for each “treatment.” If he cannot obtain a proper sample of the entire population, frequently he can secure, say, four units for each treatment, or a score for each.¹⁷

Suppose, for example, that thorough research on one city and one rural county, discloses higher levels of political interest in the former. It is presumptuous (although common practice) to present this result as evidence that urban people in *general* show a higher level. (Unfortunately, I am not beating a dead horse; this nag is pawing daily in the garden of social science.) However, very likely there is a great deal of variation in political interest among different cities, as well as among rural counties; the results of the research will depend heavily on which city and which county the researcher picked as “typical.” The research would have a broader base if a city and a rural county

comprehensiveness in that, in addition to the 4 effects of single factors, their 11 possible interactions are evaluated. There is a third advantage which, while less obvious than the former two, has an important bearing upon the utility of the experimental results in their practical application. This is that any conclusion, such as that it is advantageous to increase the quantity of a given ingredient, has a wider inductive basis when inferred from an experiment in which the quantities of other ingredients have been varied, than it would have from any amount of experimentation, in which these had been kept strictly constant. The exact standardisation of experimental conditions, which is often thoughtlessly advocated as a panacea, always carries with it the real disadvantage that a highly standardized experiment supplies direct information only in respect of the narrow range of conditions achieved by standardisation. Standardisation, therefore, weakens rather than strengthens our ground for inferring a like result, when, as is invariably the case in practice, these conditions are somewhat varied.”

¹⁷ For simplicity the following illustration is a simple contrast between two values of the “explanatory” variable, but the point is more general; and this aspect is similar whether for true experiments or controlled observations. Incidentally, it is poor strategy to “solve” the problem of representation by obtaining a good sample, or complete census, of some small or artificial population. A poor sample of the United States or of Chicago usually has more over-all value than the best sample of freshman English classes at X University.

would have been chosen in each of, say, four different situations—as different as possible (as to region, income, industry, for example); or better still in twenty different situations. A further improvement would result if the stratification and selection of sampling units followed a scientific sample design.

Using more sampling units and spreading them over the breadth of variation in the population has several advantages. First, some measure of the variability of the observed effect may be obtained. From a probability sample, statistical inference to the population can be made. Second, the base of the inference is broadened, as the effect is observed over a variety of situations. Beyond this lies the combination of results from researches over several distinct cultures and periods. Finally, with proper design, the effects of several potentially confounding factors can be tested.

These points are brought out by Keyfitz in an excellent example of controlled investigation (which also uses sampling effectively): “Census enumeration data were used to answer for French farm families of the Province of Quebec the question: Are farm families smaller near cities than far from cities, other things being equal? The sample of 1,056 families was arranged in a 2⁶ factorial design which not only controlled 15 extraneous variables (income, education, etc.) but incidentally measured the effect of 5 of these on family size. A significant effect of distance from cities was found, from which is inferred a geographical dimension for the currents of social change.”¹⁸ The mean numbers of children per family were found to be 9.5 near and 10.8 far from cities; the difference of 1.3 children has a standard error of 0.28.

SOME MISUSES OF STATISTICAL TESTS

Of the many kinds of current misuses this discussion is confined to a few of the most common. There is irony in the circumstance that these are committed usually by the more statistically inclined investigators; they are avoided in research pre-

¹⁸ Nathan Keyfitz, “A Factorial Arrangement of Comparisons of Family Size,” *American Journal of Sociology*, 53 (March, 1953), p. 470.

sented in terms of qualitative statements or of simple descriptions.

First, there is "hunting with a shot-gun" for significant differences. Statistical tests are designed for distinguishing results at a pre-determined level of improbability (say at $P = .05$) under a specified null hypothesis of random events. A rigorous theory for dealing with individual experiments has been developed by Fisher, the Pearsons, Neyman, Wold, and others. However, the researcher often faces more complicated situations, especially in the analysis of survey results; he is often searching for interesting relationships among a vast number of data. The keen-eyed researcher hunting through the results of one thousand random tosses of perfect coins would discover and display about fifty "significant" results (at the $P = .05$ level).¹⁹ Perhaps the problem has become more acute now that high-speed computers allow hundreds of significance tests to be made. There is no easy answer to this problem. We must be constantly aware of the nature of tests of null hypotheses in searching survey data for interesting results. After finding a result improbable under the null hypothesis the researcher must not accept blindly the hypothesis of "significance" due to a presumed cause. Among the several alternative hypotheses is that of having discovered an improbable random event through sheer diligence. Remedy can be found sometimes by a reformulation of the statistical aims of the research so as to fit the available tests. Unfortunately, the classic statistical

¹⁹ William H. Sewell, "Infant Training and the Personality of the Child," *American Journal of Sociology*, 53 (September, 1952), pp. 150-159. Sewell points to an interesting example: "On the basis of the results of this study, the general null hypothesis that the personality adjustments and traits of children who have undergone varying training experiences do not differ significantly cannot be rejected. Of the 460 chi square tests, only 18 were significant at or beyond the 5 per cent level. Of these, 11 were in the expected direction and 7 were in the opposite direction from that expected on the basis of psychoanalytic writings. . . . Certainly, the results of this study cast serious doubts on the validity of the psychoanalytic claims regarding the importance of the infant disciplines and on the efficacy of prescriptions based on them" (pp. 158-159). Note that by chance alone one would expect 23 "significant" differences at the 5 per cent level. A "hunter" would report either the 11 or the 18 and not the hundreds of "misses."

tests give clear answers only to some simple decision problems; often these bear but faint resemblance to the complex problems faced by the scientist. In response to these needs the mathematical statisticians are beginning to provide some new statistical tests. Among the most useful are the new "multiple comparison" and "multiple range" tests of Tukey, Duncan, Scheffé,²⁰ and others. With a greater variety of statistical statements available, it will become easier to choose one without doing great violence either to them or to the research aims.

Second, statistical "significance" is often confused with and substituted for substantive significance. There are instances of research results presented in terms of probability values of "statistical significance" alone, without noting the magnitude and importance of the relationships found. These attempts to use the probability levels of significance tests as measures of the strengths of relationships are very common and very mistaken. The function of statistical tests is merely to answer: Is the variation great enough for us to place some confidence in the result; or, contrarily, may the latter be merely a happenstance of the specific sample on which the test was made? This question is interesting, but it is surely *secondary*, auxiliary, to the main question: Does the result show a relationship which is of substantive interest because of its nature and its magnitude? Better still: Is the result consistent with an assumed relationship of substantive interest?

The results of statistical "tests of significance" are functions not only of the magnitude of the relationships studied but also of the numbers of sampling units used (and the efficiency of design). In small samples significant, that is, meaningful, results may fail to appear "statistically significant." But if the sample is large enough the most insignificant relationships will appear "statistically significant."

Significance should stand for meaning and refer to substantive matter. The statistical

²⁰ John W. Tukey, "Comparing Individual Means in the Analysis of Variance," *Biometrics*, 5 (June, 1949), pp. 99-114; David B. Duncan, "Multiple Range and Multiple F Tests," *Biometrics*, 11 (March, 1955), pp. 1-42; Henry Scheffé, "A Method for Judging All Contrasts in the Analysis of Variance," *Biometrika*, 40 (June, 1953), pp. 87-104.

tests merely answer the question: Is there a big enough relationship here which *needs* explanation (and is not merely chance fluctuation)? The word *significance* should be attached to another question, a substantive question: Is there a relationship here *worth* explaining (because it is important and meaningful)? As a remedial step I would recommend that statisticians discard the phrase "test of significance," perhaps in favor of the somewhat longer but proper phrase "test against the null hypothesis" or the abbreviation "TANH."

Yates, after praising Fisher's classic *Statistical Methods*, makes the following observations on the use of "tests of significance":

Second, and more important, it has caused scientific research workers to pay undue attention to the results of the tests of significance they perform on their data, particularly data derived from experiments, and too little to the estimates of the magnitude of the effects they are investigating.

Nevertheless the occasions, even in research work, in which quantitative data are collected solely with the object of proving or disproving a given hypothesis are relatively rare. Usually quantitative estimates and fiducial limits are required. Tests of significance are preliminary or ancillary.

The emphasis on tests of significance, and the consideration of the results of each experiment in isolation, have had the unfortunate consequence that scientific workers have often regarded the execution of a test of significance on an experiment as the ultimate objective. Results are significant or not significant and this is the end of it.²¹

For presenting research results statistical estimation is more frequently appropriate than tests of significance. The estimates should be provided with some measure of sampling variability. For this purpose confidence intervals are used most widely. In large samples, statements of the standard errors provide useful guides to action. These problems need further development by theoretical statisticians.²²

The responsibility for the current fashions should be shared by the authors of statistical

textbooks and ultimately by the mathematical statisticians. As Tukey puts it:

Statistical methods should be tailored to the real needs of the user. In a number of cases, statisticians have led themselves astray by choosing a problem which they could solve exactly but which was far from the needs of their clients. . . . The broadest class of such cases comes from the choice of significance procedures rather than confidence procedures. It is often much easier to be "exact" about significance procedures than about confidence procedures. By considering only the most null "null hypothesis" many inconvenient possibilities can be avoided.²³

Third, the tests of null hypotheses of zero differences, of no relationships, are frequently weak, perhaps trivial statements of the researcher's aims. In place of the test of zero difference (the nullest of null hypotheses), the researcher should often substitute, say, a test for a difference of a specific size based on some specified model. Better still, in many cases, instead of the tests of significance it would be more to the point to measure the magnitudes of the relationships, attaching proper statements of their sampling variation. The magnitudes of relationships cannot be measured in terms of levels of significance; they can be measured in terms of the difference of two means, or of the proportion of the total variance "explained," of coefficients of correlations and of regressions, of measures of association, and so on. These views are shared by many, perhaps most, consulting statisticians—although they have not published full statements of their philosophy. Savage expresses himself forcefully: "Null hypotheses of no difference are usually known to be false before the data are collected; when they are, their rejection or acceptance simply reflects the size of the sample and the power of the test, and is not a contribution to science."²⁴

Too much of social research is planned and presented in terms of the mere existence of some relationship, such as: individuals

²¹ Frank Yates, "The Influence of *Statistical Methods for Research Workers* on the Development of the Science of Statistics," *Journal of the American Statistical Association*, 46 (March, 1951), pp. 32-33.

²² D. R. Cox, "Some Problems Connected with Statistical Inference," *Annals of Mathematical Statistics*, 29 (June, 1958), pp. 357-372.

²³ John W. Tukey, "Unsolved Problems of Experimental Statistics," *Journal of the American Statistical Association*, 49 (December, 1954), p. 710. See also D. R. Cox, *op. cit.*, and David B. Duncan, *op. cit.*

²⁴ Richard J. Savage, "Nonparametric Statistics," *Journal of the American Statistical Association*, 52 (September, 1957), pp. 332-333.

high on variate x are also high on variate y . The *exploratory* stage of research may be well served by statements of this order. But these statements are relatively weak and can serve *only* in the primitive stages of research. Contrary to a common misconception, the more advanced stages of research should be phrased in terms of the quantitative aspects of the relationships. Again, to quote Tukey:

There are normal sequences of growth in immediate ends. One natural sequence of immediate ends follows the sequence: (1) Description, (2) Significance statements, (3) Estimation, (4) Confidence statement, (5) Evaluation. . . . There are, of course, other normal sequences of immediate ends, leading mainly through various decision procedures, which are appropriate to development research and to operations research, just as the sequence we have just discussed is appropriate to basic research.²⁵

At one extreme, then, we may find that the contrast between two "treatments" of a labor force results in a difference in productivity of 5 per cent. This difference may appear "statistically significant" in a sample of, say, 1000 cases. It may also mean a difference of millions of dollars to the company. However, it "explains" only about one per

cent of the total variance in productivity. At the other extreme is the far-away land of completely determinate behavior, where every action and attitude is explainable, with nothing left to chance for explanation.

The aims of most basic research in the social sciences, it seems to me, should be somewhere between the two extremes; but too much of it is presented at the first extreme, at the primitive level. This is a matter of over-all strategy for an entire area of any science. It is difficult to make this judgment off-hand regarding any specific piece of research of this kind: the status of research throughout the entire area should be considered. But the superabundance of research aimed at this primitive level seems to imply that the over-all strategy of research errs in this respect. The construction of scientific theories to cover broader fields—the persistent aim of science—is based on the synthesis of the separate research results in those fields. A coherent synthesis cannot be forged from a collection of relationships of unknown strengths and magnitudes. The necessary conditions for a synthesis include an *evaluation* of the results available in the field, a coherent interrelating of the *magnitudes* found in those results, and the construction of models based on those magnitudes.

²⁵ Tukey, *op. cit.*, pp. 712-713.

THE GEOMETRIC INTERPRETATION OF AGREEMENT

W. S. ROBINSON

University of California, Los Angeles

This paper refers to a prior article by the author in which the statistical idea of "agreement" as contrasted with correlation was developed and shown to be measured appropriately by the intraclass correlation coefficient or a simple modification of it. Section I of this paper is entirely expository, and simplifies the original algebraic argument for the non-mathematical reader by discussing some properties of agreement with reference to a series of simple graphs. Section II, by adding a geometric counterpart to the original algebraic definition of agreement, spells out further implications of agreement as a concept.

SOME time ago I published a paper on "The Statistical Measurement of Agreement,"¹ in which I developed the idea of agreement as contrasted with correlation from first principles, and showed that the intraclass correlation coefficient or a simple

modification of it was an appropriate measure of agreement. The argument of the paper was algebraic and somewhat compact.

The original characterization of agreement was as follows: "Agreement requires that paired values be identical, while correlation requires only that paired values be linked by a linear relationship, or, if one defines correlation more broadly, that the paired values

¹ W. S. Robinson, *American Sociological Review*, 22 (February, 1957), pp. 17-25.

This document was created with Win2PDF available at <http://www.daneprairie.com>.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.