Hypothesis Testing in Relation to Statistical Methodology

Cherry Ann Clark

Review of Educational Research, Vol. 33, No. 5, Statistical Methodology. (Dec., 1963), pp. 455-473.

Stable URL:
http://links.jstor.org/sici?sici=0034-6543%28196312%2933%3A5%3C455%3AHTIRTS%3E2.0.CO%3B2-D

Review of Educational Research is currently published by American Educational Research Association.

CHAPTER I

Hypothesis Testing in Relation to Statistical Methodology

CHERRY ANN CLARK

THE SHORTCOMINGS in the methodology of statistical hypothesis testing used in educational and psychological research have been emphasized repeatedly in recent behavioral science and statistical literature (Binder, 1963; Edwards, Lindman, and Savage, 1963; Grant, 1962; Lubin, 1962; McNemar, 1960; Mowrer, 1960; Nunnally, 1960; Rozeboom, 1960; Savage, 1957). This chapter reviews the salient points of the criticisms. Since this issue of the REVIEW marks the first time an entire chapter has been devoted to the statistical methodology of hypothesis testing, a brief account of several theories of statistical inference is included to provide a background for evaluating the rationales of significance tests compared with other methods for statistical inferences, as well as the modifications in the function of the null hypothesis in testing statistical hypotheses. The problems in statistical inference which have been associated with the widespread use of significance tests are reviewed. The limitations in and the effectiveness of significance tests as methods for informative inference are described. The applications of interval estimation contrasted with significance tests in the investigation of hypotheses and models and in the development of a body of empirical data are summarized.

Hypothesis Testing and Statistical Inference

Hypothesis testing is a central and complex problem in the methodology of science. It involves comparing the deductions or the predictions from scientific hypotheses with observational data to eliminate unsatisfactory hypotheses and to give support to satisfactory ones.

A primary objective of hypothesis testing methodology is the formulation of rules or criteria, sometimes called decision procedures, to use in determining whether the data should be construed as rejecting or accepting the hypothesis under investigation. Decision procedures are judged by such criteria as consistency, relevance, completeness, and effectiveness. Decision rules are formulated to provide objective, reliable, and valid solutions to important problems in the analysis of data (Buehler, 1959; Hotelling, 1958; Jeffreys, 1957, 1961; Tukey, 1962). The hypotheses to be tested must be examined for their logical validity or consistency, their heuristic value, and their amenability to empirical test. The testing of

hypotheses requires that appropriate controls be employed in making observations and in analyzing the data. Systematic errors in the experimental procedures must be minimized if decision rules are to be efficient (Tukey, 1954, 1960a, 1962).

Statistical hypothesis testing is a special instance of the general scientific method of testing hypotheses. As in the general method, statistical hypothesis testing combines deductive and inductive methods in intricate ways. For the most part, statistical hypothesis testing has been concerned with determining which one of a dichotomous set of mutually exclusive and exhaustive hypotheses is to be rejected and which one accepted at a specified level of risk in making an erroneous conclusion or decision. Probability theory provides the deductive foundations for theories of statistical inference (Hotelling, 1958). Natural phenomena which are presumed to be subject to random fluctuations, either because of a stochastic process characterizing their behavior or because of random errors in observing or measuring them, provide the inductive basis for statistical inference. Both enumerative and eliminative methods of induction are used in treating the data and in formulating statistical generalizations.

The question of whether statistical methods based on eliminative or enumerative rules of induction constitute the best foundation for statistical inference is widely disputed (Rozeboom, 1960; Savage and others, 1962). The argument about whether scientific knowledge advances primarily by accumulation of positive instances of phenomena to support a hypothesis or by negative instances to refute a hypothesis is the focus of some of the criticism of the classical theories of statistical inference—namely, the Fisher and Neyman-Pearson theories (Savage, 1961). Another argument in the current controversy on the foundations of statistics focuses on the question of whether there is a sound logical basis for assigning a direct probability to a set of propositions or hypotheses. The frequency or objective school has denied the legitimacy of such a procedure. Probability, for this school, refers to the relative frequency of events within a class of random recurrences (Hotelling, 1958), not to the amount of support for or the degree of belief in a hypothesis. On the other hand, the subjective probability school, among others, has argued that a fruitful and consistent way to proceed with statistical inferences is by assigning a degree of belief to a statement based upon available information and then by altering the amount of uncertainty about the statement in the light of additional experimental evidence by the use of Bayes's theorem (Lindley, 1953, 1961; Rozeboom, 1960; Savage, 1961).

Another aspect of the controversy is how the likelihood function should be incorporated into statistical theory and practice. One of the serious limitations in the classical theories of significance tests is that they are not dependent upon the likelihood function (Birnbaum, 1962).

Statistical hypotheses are more restricted than are scientific hypotheses, just as statistical inferences are more restricted than scientific inferences (Bolles, 1962; Cox, 1958b; Tukey, 1960a). Statistical hypotheses con-

cern the behavior of observable random variables, whereas scientific hypotheses treat the phenomena of nature and man. A null hypothesis is a particular statistical hypothesis which refers to the theoretical probability distribution governing the random variable(s) under investigation and is the hypothesis under test (Kendall and Stuart, 1961). Statistical inference, broadly defined, deals with statements about statistical populations made from given observations with a measured degree of uncertainty (Cox, 1958b). Statistical inference proceeds from observations to conclusions about the populations sampled. Scientific inference, on the other hand, often argues from descriptive facts about populations to an abstraction about the system of phenomena under investigation. A statistical inference frequently is instrumental in formulating a general inference, but usually it constitutes only a small part of the uncertainty connected with the scientific inference (Cox, 1958b; Tukey, 1960a).

Decision procedures used in making statistical inferences, including those that aid in the selection of statistical design and analysis (Cox, 1958a; Savage, 1961), require that the investigator use his judgment in diverse ways. How such judgment should be allowed to affect the use of the decision procedures is a crucial issue in the present controversy on the foundations of statistical inference (Savage and others, 1962; Tukey, 1954, 1960a, 1962). Much of the criticism of the rationales for significance tests has been aimed at the inconsistencies in the performance of the decision procedures (Berkson, 1942; Birnbaum, 1962; Cox, 1958b; Edwards, Lindman, and Savage, 1963; Pratt, 1961a, b; Savage and others, 1962).

Development of Theories of Statistical Inference and Significance Tests

The use of significance tests preceded the development of any systematic theories of statistical inference. Barnard (Savage and others, 1962) and Lehmann (1959) have mentioned their use by two early probability theorists, Daniel Bernoulli and Pierre Simon de Laplace. The principal components of significance tests were contained in their mathematical studies on the departures of planetary orbits from specific mathematical models. The essential components were the theoretical or mathematical hypotheses and equations and the observations on the extent of the departures of the planetary planes from the expected values. The significant departures were those that were considered unlikely on a particular hypothesis. When Bernoulli found that the values he computed were uniformly improbable on the hypotheses, he concluded that the observations provided evidence for rejecting the initial hypotheses. He did not consider an alternative hypothesis, nor did he have a particular degree of departure in mind before he made his computations.

Karl Pearson's work on goodness of fit tests and the chi square distribution and W. S. Gossett's derivation of the student distribution, a small

sample exact-sampling distribution, were important influences on subsequent developments in the theory of significance tests.

Contributions of Fisher to Statistical Inference

Sir Ronald Fisher has been called the founder of modern statistical theory. He propounded many basic concepts and methods and developed the first complete theory of statistical inference (Hotelling, 1951; Pearson, 1962; Yates, 1951). Among his contributions to the theory of significance tests were a systematic rationale for critical ratio tests based on the Central Limit Theorem and the Law of Large Numbers; exact small sample tests for which he derived many theoretical sampling distributions; and randomization or permutation tests to be used in complex experimental designs when the assumptions underlying other significance tests procedures are not fulfilled. He developed interval estimation methods, which culminated in his theory of fiducial inference with its fiducial intervals. He introduced an extensive formal theory of experimental and statistical design and related methods of analysis, including the analyses of variance and covariance. He set forth the method of maximum likelihood as a foundation for statistical inference and as a basis for developing many statistical techniques. He discussed the concept of sufficient statistics, which has been used extensively in modern statistical theories. He argued that inverse probability has no place in modern science or statistical theory (Fisher, 1956, 1960).

He originated the term *null hypothesis.* He presented significance tests of null hypotheses as examples of a logical disjunction. In this view, the null hypothesis could never be proved by experimentation. On the contrary, the null hypothesis exists only to be rejected by a sufficiently sensitive experiment. Rarely is an experimenter interested in accepting the null hypothesis, an exception being in determining the uniformity of experimental procedures (Fisher, 1960). Fisher discussed the combination of estimation procedures with significance tests after the initial phases of experimentation have been completed.

Hotelling (1951) and Yates (1951), in evaluating Fisher's impact upon statistical practice in the social sciences, suggested that his books have not had an altogether positive influence. Psychologists, for example, have been content to publish very limited reports on one or at most several significance tests based on null hypotheses of no difference (Grant, 1962; Nunnally, 1960; Rozeboom, 1960). Such statistical tests have very little informative value. They often are known to be false prior to experimentation (Savage, 1954).

Contributions of Neyman and Pearson to Statistical Inference

Jerzy Neyman and Egon Pearson collaborated for more than a decade on the development of many aspects of statistical theory and methods.

Some of their work was an extension of Fisher's contributions, and some of it was a reaction to Fisher's proposals. They expanded and systematized the rationale of significance tests. They argued for the necessity of carefully considering the alternative to the null hypothesis in making any test of significance. They introduced the notions of errors of the first and second kinds, the power of a test in discriminating between the hypothesis under test and the alternative, and the comparison of the power functions of different tests to aid in the selection of the most satisfactory statistical test. They formulated an extensive rationale for the comparison and derivation of different tests in terms of their "nice properties," such as the uniformly most powerful test, most powerful tests, and unbiased tests. They expanded Fisher's maximum likelihood method into the test criterion called the Neyman-Pearson lemma, which they used to develop tests of simple statistical hypotheses. They gave considerable attention to the likelihood ratio in the development of likelihood ratio tests, including both univariate and multivariate tests. They pointed out the great complexity of the problems in deriving suitable test criteria for complex hypotheses as contrasted with simple hypotheses. They, like Fisher, demonstrated the wide variety of statistical tests to be derived from the maximum likelihood method.

They posited that the size of the sample should be set before experimentation is begun. The decision rules for tests of significance within their theory are based upon hypothetical repetitions of similar size random samples from the population(s) under investigation. Determination of the size of the sample is a function both of the size of the critical region and of the sensitivity or power of the test in detecting departures from the null hypothesis. They conceptualized the error rate in significance tests as the probability of obtaining a given or more extreme level of significance in an extended series of random fixed sample size experiments. They reasoned that a significance test is a decision procedure to be used as an aid in deciding which one of two actions to take in the face of uncertainty (Lehmann, 1959).

In their exposition of the theory of significance tests, Neyman and Pearson incorporated the spirit of probabilistic reasoning for a two-valued problem. Theoretically, the null hypothesis and the alternative hypothesis have equal status: the data are to determine which hypothesis is more probable. The error of the first kind is considered the more serious; it is determined by the size of the critical region or the level of significance. The error of the second kind cannot be controlled directly by the experimenter, but is a function of the size of the critical region and of the characteristics of the power function of the test in relation to the specific alternative hypothesis under consideration. In actuality, the null hypothesis is rarely found to be acceptable in the Neyman-Pearson formulation, for as the sample size increases, a sufficiently sensitive test rejects the null hypothesis at a decreasing level of significance (Kendall and Stuart, 1961; Pratt, 1961b). This fact has been of grave

concern to a number of statisticians (Savage and others, 1962). This fact, in addition to certain other inconsistencies in the performance of the decisions rule, has evoked sharp criticism of this classical method for statistical inference (Pratt, 1961b).

The Neyman-Pearson theory indicated the relationship between tests of significance and the method of interval estimation called confidence intervals (Berkson, 1942; Birnbaum, 1961; Bulmer, 1957; Kendall and Stuart, 1961; Lehmann, 1959; Natrella, 1960; Pratt, 1961a). A test of significance is concerned with only those values specified by the hypotheses under consideration, while confidence intervals by their width and by the level of confidence indicate the multitude of statistical hypotheses which are acceptable as well as how acceptable each hypothesis is by its location within the confidence interval. The Neyman-Pearson rationale showed the practical and theoretical equivalence of tests of significance accompanied by the power function or the operating characteristic curve and confidence intervals (Natrella, 1960). Pratt (1961a, b) reminded statisticians of one important shortcoming of the theory of confidence intervals: the probability is either one or zero that the value of interest is contained within the confidence interval; moreover, the formulation is based upon the frequency interpretation of probability, which does not allow for a straightforward degree of confidence in the likelihood of the obtained statistic.

Educational and psychological investigators have rarely used effectively the Neyman-Pearson formulation either for significance testing or for interval estimation.

Contributions of Wald to Statistical Inference

In the short but exceptionally productive time that Wald contributed to statistics, he made many important modifications in the theory of statistical hypothesis testing. He recognized the desirability of incorporating into statistical theory and practice provision for handling sequentially selected samples of varying sizes of items rather than limiting statistical design and analysis to fixed sample sizes, as required by the Neyman-Pearson method. He devised the sequential probability ratio test to analyze sequentially sampled items in industrial situations, such as quality control. Sampling is continued until a desired degree of precision is obtained as a basis either for accepting the null hypothesis or for rejecting it and accepting the alternative. The dichotomous decision problem of earlier significance tests is expanded into a trichotomous decision problem: either to accept or reject a statistical hypothesis or to continue sampling (Savage, 1954; Schlaifer, 1959). Two constants are selected for the test procedure to give the desired weight to the probability of each of the two kinds of errors.

In his *Statistical Decision Functions,* Wald (1950) gave explicit consideration to the statistician's role as a decision maker by his detailed

specifications of the various factors of the decision situation. Previously, Neyman and Pearson had given indirect consideration to the possible losses consequent to a wrong decision by the different values to be selected for the two kinds of errors. Instead of basing the decision rule upon the specification of error rates as Fisher and Neyman and Pearson had done, Wald gave formal recognition to the assessment of losses involved in making wrong decisions (Bahadur and Robbins, 1950; Lindley, 1953, 1961).

The formal mathematical structure of statistical decision theory is based upon the theory of games (Luce and Raiffa, 1957). It requires the statistical consumer or investigator to be able to list a set of possible actions, among which is a preferred action. The action preferred depends on what the "true state of nature" is, in other words, upon the unknown value of some parameter. The preferences among the actions are assessed in terms of the losses attached to each possible action to be taken in the presence of each state of nature (the circumstance which may be referred to as the loss function). The decision maker can obtain information about the state of nature at some specified cost for the observations made; he must balance cost against possible losses in the face of insufficient information. The decision function, then, depends upon the given loss function, the set of possible actions, the possible states of nature, the prior information available about the state of nature, and the cost of making observations (Savage, 1954).

Point estimation, interval estimation, and hypothesis testing are subsumed within a single theory, known as statistical hypothesis testing. The problem of choosing a design and an analysis is integrated within the theory and is closely associated with the factors determining the choice of the statistical decision function.

Another noteworthy contribution of Wald was his demonstration that there is not a unique decision procedure which is equally effective in all conditions. He showed that in those cases where prior information is available, a Bayes solution is admissible and optimum. There are situations, he pointed out, in which a Bayes solution is not part of an admissible strategy. He also raised the question of formulating a class of admissible hypotheses, and he thereby clarified some of the issues connected with composite hypotheses in the Neyman-Pearson formulation.

The theory of testing statistical hypotheses provides a decision procedure for working with multiple hypotheses rather than merely with two hypotheses, as in the theory of significance tests. For example, given three hypotheses, including a null hypothesis and two alternative hypotheses at different distances on either side of the null hypothesis, an optimum statistical decision function often is available to select the most probable of the three hypotheses on the available data (Lehmann, 1959).

The mathematical complexity of statistical decision theory and the great amount of information which is required to use it have discouraged its application in behavioral science research. Its theoretical formulations

have helped to clarify a number of problems in the conventional use of significance tests (Kaiser, 1960).

Contributions of Bayesian Statistical Theory to Statistical Inference

Bayesian statistical theory, for the most part, is not so dependent on significance tests for informative statistical inferences as is classical statistical theory, namely the theories of Fisher and Neyman and Pearson. With rare exceptions, the British-American school of statistics has been firmly entrenched in the objective or frequency interpretation of probability and has eschewed the use of Bayes's theorem or the likelihood function as bases for formulating statistical inferences. The Continental school, on the other hand, has been concerned with giving formal mathematical treatment to the subjective theory of probability or with developing statistical methods based on the use of Bayes's theorem (Savage, 1954).

One British scientist working somewhat independently of the British-American school, his work unrecognized until recently by the majority of American statisticians, has made many important technical advances in the use of Bayes's theorem. Jeffreys (1957, 1961) has treated the problem of significance tests in a unique manner. Tests of significance are used to assess whether the hypothesized parameter(s) can be considered to account adequately for the obtained data. If the approximation of the posterior probability distribution to the null hypothesis, which reflects the prior probability distribution, is not satisfactory, then another parameter is introduced into the model. The process of adjusting the value of the null hypothesis is continued until a satisfactory approximation is attained, that is, within the limits of random sampling variation and errors of measurement. Jeffreys has emphasized the importance of refining measurement methods for the advancement of scientific knowledge. For Jeffreys one of the objectives of scientific investigation has been the verification of satisfactory null hypotheses. Perusal of the computations required in the use of Bayes's theorem for *t*-tests of significance would dismay the average social scientist accustomed to *t*-tests without consideration of prior probability distributions.

The merits of Jeffreys' work have been recognized by Lindley (1953, 1961), Raiffa and Schlaifer (1961), Savage (1954, 1961), and Savage and others (1962). His work and that of the French and Italian probability theorists have influenced the surgence of Bayesian statistics, especially Bayesian statistical decision theory in the United States (Raiffa and Schlaifer, 1961; Roberts, 1962; Schlaifer, 1959).

In his compendious discussion concerning the foundations of statistics of 10 years ago, Savage (Savage, 1954; Edwards, 1956) argued that various aspects of the classical theories of statistics could be reinforced and given a consistent logical framework relevant to the behavior of the

scientist and the decision maker by adopting the personal interpretation of probability. At that time, he critically evaluated many of the conventional practices in the use of statistical methods, especially the widespread use of extreme null hypotheses which are known to be false prior to experimentation. Subsequently, Lindley (1961) and Raiffa and Schlaifer (1961) have built upon the work of Jeffreys a still incomplete structure of statistical theory and methods using Bayes's theorem. Edwards, Lindman, and Savage (1963) have presented to psychologists some of the fundamental ideas and techniques of Bayesian statistics.

Bayesian statisticians have pointed out that classical statistical methods do not make full use of available information and that they do not provide concise and relevant answers to the questions of greatest importance to investigators. Furthermore, in many situations the decision procedures do not meet the basic criterion of consistency (Pratt, 1961b).

Classical and Bayesian formulations have points in common. Both have the following elements: alternative hypotheses or acts, possible parameter values or states, the possibility of sampling (experimentation) to obtain information about the parameter value or state, the sampling or experimental results in the form of descriptive statistics, and either loss structures or error rates. Both seek the best decision rule which will minimize loss or error in the decisions or conclusions made following experimentation.

The classical decision maker chooses an act on the basis of the outcome of a sample which is conditional upon the parameter and the type of sample and experiment. Within the decision rules for significance tests, the so-called conditionality of the experiment and of the sample is not adequately represented; the statistician must use his judgment to give appropriate weight to the conditions when he interprets the obtained level of significance (Cox, 1958b).

The Bayesian statistician pursues a different course. He begins by making probabilistic statements about the parameter under investigation. The probabilistic statements are in the form of a prior distribution. Then, dependent upon how precise he wishes to be in his final probabilistic statements about the value of the parameter, he sets a limit, or an upper bound, in the form of an expected value on the cost or worth of experimental data for selecting a terminal act or decision. From this point of view he decides whether it is worthwhile to sample. As sampling information becomes available, the Bayesian modifies his prior distribution in the light of sample evidence and thereby obtains a posterior distribution. By using some specified decision rule, he may decide that he has sufficient information to come to a decision or a conclusion; or he may decide that his uncertainty has not been reduced enough and that, therefore, he must continue sampling until he has a suitable basis for a terminal act or conclusion (Raiffa and Schlaifer, 1961; Roberts, 1962).

The Bayesian analysis formalizes many aspects of statistical practice which classical methods leave to the judgment of the investigator. For the orthodox contention that there is often little objective basis for arriv-

ing at prior probability distributions, Bayesians have countered that in at least a number of situations the prior distribution assumes little weight in the final or posterior distribution (Edwards, Lindman, and Savage, 1963; Savage and others, 1962). Several interesting solutions to this problem have been indicated by Raiffa and Schlaifer (1961).

Edwards, Lindman, and Savage (1963), Lindley (1961), and Savage and others (1962) have discussed some of the advantages of Bayesian significance tests over classical tests. The first-mentioned authors have discussed the importance in many educational and psychological research problems of taking possible losses or gains and prior information into account. For example, classical *t*-tests and *F*-tests have been used to study the differential effects on several groups of several methods of instruction. In neither approach has it been practical to take into account considerations of previously available information about the differential effectiveness of one method versus another; nor has it been customary to give attention to various kinds of losses, economic or learning, which might be associated with the adoption of one method rather than another. The authors have showed how Bayesian procedures can provide appropriate answers to such problems.

The Likelihood Principle and Statistical Inference

The likelihood principle, like Bayes's theorem, has been recognized as a part of the armamentarium of statistics; but, as with Bayes's theorem, it has been given little attention as a primary tool for statistical inference. The likelihood function is represented by the distribution function of the random sample variables corresponding to values of given parameters and is determined by the observed outcome of a random variable in any specified experiment (Birnbaum, 1962). Barnard, Fisher, and Birnbaum (Birnbaum, 1962) have argued that the likelihood function alone provides a suitable and sufficient basis for interpreting experimental data. The likelihood function can be interpreted without reference to the structure of the experiment. In contrast with the classical methods of significance tests and confidence levels, its use avoids the problem of having to modify the interpretation of the significance and confidence levels in the light of the experimental frame of reference. The likelihood function clearly asserts the irrelevance of possible experimental outcomes which have not been observed during any given experiment. One of the main objections to the use of the tail areas of the probability distribution for significance tests is just that unobserved experimental possibilities are irrelevant for the formulation of inferences about the actual observations. The likelihood function and Bayes's theorem are not dependent upon the sequence of or the stopping point in sampling, whereas orthodox significance tests and confidence intervals cannot be used when sampling is stopped arbitrarily.

Use of Significance Tests for Statistical Inference

The test of significant difference has been called the prototype of modern experimental statistics (Tukey, 1960a). It is a qualitative rather than a quantitative procedure for statistical analysis and inference which is used to answer questions such as this: Dare we conclude that the difference is not zero? A classical significance test assays whether the two statistical hypotheses A and B are equal or whether A is less than or greater than B. The failure to obtain a significant difference does not warrant the conclusion that A is equal to B until careful consideration is given to the specific experimental situation, the available evidence, and the assessment of the consequences for such a decision or conclusion (Tukey, 1960a). The precision of the experimental comparison, the power of the statistical test, and the theoretical and the empirical closeness of A to B must be examined before the lack of statistical significance is interpreted as a positive finding (Cox, 1958a; Tukey, 1960a). The effectiveness of confidence intervals in showing the probable relationship between A and B indicates their superiority over significance tests for informative inference. Statisticians have given a great deal of attention to the use and the misuse of significance tests (Anscombe, 1961; Bahadur and Robbins, 1950; Good, 1958; Kish, 1959; Lindley, 1958; Savage, 1957; Selvin, 1957; Sterling, 1959, 1960; Williams, 1959). There has been general agreement that significance tests have been used too frequently in social science research (Grant, 1962; Harrington, 1961; Lubin, 1962; Nunnally, 1960; Rozeboom, 1960; Savage, 1957; and Wilson, 1961) and at the expense of more appropriate procedures, such as confidence and fiducial intervals (Birnbaum, 1961; Grant, 1962; Kish, 1959; Lindley, 1958; and Savage and others, 1962). Significance tests without power functions do not answer such questions as these: How far is the sample statistic from the null hypothesis? How much credence should be given to the null or the alternative hypotheses?

The situations in which classical significance tests can be used are limited. Anscombe (1961) has pointed out the usefulness of significance tests as a fundamental method for testing theoretical hypotheses or models. For this purpose, the null hypothesis is given some credence in the light of theoretical and experimental consideration. The investigator wishes to determine in a new sample whether the null hypothesis is tenable, the observed departures being no more than those expected as a result of random sampling variations. When the departures are excessive, the investigator considers whether another parameter should be introduced into the statistical model or whether some other modification in the model should be made. Another use for significance tests occurs when an experimenter wishes to test the adequacy of a particular statistical design, an analysis or technique, a particular stochastic process, or an experimental procedure. He is interested in determining whether the methods used produce the desired precision or uniformity with due allowance for random sampling

errors (Anscombe, 1961). Significance tests are also applicable when an investigator is concerned with testing a particular statistical hypothesis, when a set of specific alternatives has not yet been conceptualized (Savage and others, 1962). When a research worker wishes to verify a prediction that an experimental result has been in a specific direction, one-tailed significance tests are appropriate (Anscombe, 1961).

Significance tests can be informative in later stages of research when estimation is included in the statistical procedures. The null hypothesis of no difference has been judged to be no longer a sound or fruitful basis for statistical investigation. Both the null and the alternative hypotheses are formulated to include estimation of relevant experimental variables (Bush, 1963; Grant, 1962; Nunnally, 1960; Tukey, 1960b).

Significance tests can be divided into two groups: those for which a set of alternative hypotheses can be defined, such as tests of specific parameter values, and those for which a set of mutually exclusive and exhaustive hypotheses are not available, such as tests of randomness. The former almost always have comparable interval estimation procedures—at least, such procedures can be derived—whereas the latter pose serious problems for the derivation of interval estimation procedures. When comparable interval estimation procedures are not available, then an investigator has no choice but to use significance tests, as in the use of some distribution-free methods (Kendall and Stuart, 1961; Savage and others, 1962).

There are many statistical problems for which both significance and interval estimation procedures are available. Similar answers are obtained if the procedures are used effectively and appropriately (Berkson, 1942; Birnbaum, 1961; Kendall and Stuart, 1961; Natrella, 1960; Pratt, 1961a). Statisticians have agreed that interval estimation does give a more intuitively understandable summary of the data than do significance tests with power functions (Kish, 1959; Natrella, 1960).

Among statistical problems which often can be treated by either testing or interval estimation procedures are (a) simple preference problems, wherein a family of simple hypotheses is ranked in order of credibility on the basis of the data, and (b) composite preference problems, wherein a family of composite hypotheses is ranked on the basis of the data (Lehmann, 1959; Savage and others, 1962). These problems conventionally have been classified under the rubrics of hypothesis testing, point and interval estimation, and discrimination (Savage and others, 1962).

Problems in Statistical Hypothesis Testing

There are many unsolved problems in the theory and the practice of statistical hypothesis testing. Some of the theoretical problems have been mentioned at the beginning of this chapter. In the following paragraphs some of the practical problems are summarized.

Decision Theory and Conclusion Theory

Tukey (1960a) has suggested that some of the misunderstanding surrounding the use of significance tests can be clarified by distinguishing those statistical problems concerned with the necessity of making a choice between two alternatives from those concerned with the accumulation of evidence in support of a theory or set of working hypotheses. Decision theory is typically concerned with problems in which economic losses or rewards, a set of possible actions, and their consequences in various states of nature can be defined. Decision theory counsels the statistical consumer about how to choose wisely among available strategies (Bahadur and Robbins, 1950; Flanagan, 1958).

Conclusion theory, on the other hand, is concerned with evaluating the adequacy of evidence. Conclusions need not be made if the evidence is deemed inadequate. A conclusion is accepted relative to the conditions of an experiment until compelling evidence to the contrary is found. Decision procedures choose between two hypotheses in terms of minimizing the risk for an action, while conclusion procedures often are concerned with controlling errors of both the first and the second kinds at suitably low levels in choosing between two hypotheses (Anscombe, 1961; Lindley, 1961; Tukey, 1960a).

Rejection or Acceptance of the Null Hypothesis

Much of the controversy among statisticians and behavioral scientists about the theory and the practice of statistical hypothesis testing has centered upon whether the null hypothesis can reasonably be accepted. Grant (1962) has argued that experiments oriented toward the acceptance of null hypotheses are inappropriate. Binder (1963) has countered that the Neyman-Pearson formulation admits the validity of accepting the null hypothesis, although, unless some adjustment in significance level is made, the rationale is prone to reject the null hypothesis when sample sizes are large or when powerful tests are used (Pratt, 1961b). It is the ready rejection of the null hypothesis by classical methods that is the focus of much of the Bayesian criticism (Edwards, Lindman, and Savage, 1963; Savage and others, 1962).

Among statisticians, Anscombe (1961), Good (1958), Lindley (1958, 1961), Savage and others (1962), and Sterling (1960) have recommended that hypothesis testing should be directed toward testing plausible null hypotheses and not toward testing implausible ones that can be rejected by a single observation.

Statistical Significance and Substantive Significance

That "significance levels do not signify" (Savage and others, 1962) has been widely recognized as a primary problem in the use of significance

tests. Among the proposed solutions to the problem have been the inclusion of a distance function to indicate how deviant the data are from the null hypothesis (Bulmer, 1957), the use of a form of variance in analysis of variance to show how well the independent variables account for the obtained results (Bolles and Messick, 1958; Gaito, 1958), the substitution of interval estimation for significance tests to give a quantitative representation of the spread of the obtained values (Birnbaum, 1961; Kish, 1959; Natrella, 1960), the use of Bayesian methods (Edwards, Lindman, and Savage, 1963), and the use of the simple likelihood function (Birnbaum, 1962).

Significance tests as methods for informative inference have been criticized severely (Birnbaum, 1962; Edwards, Lindman, and Savage, 1963; Lubin, 1962; Rozeboom, 1960; Savage and others, 1962). Mere inspection of the data often reveals the untenability of the null hypothesis. When there are grounds for believing that random sampling variations alone do not account for the data, a significance test is not the most appropriate method to use (Williams, 1959). Significance tests do not summarize the evidence (Birnbaum, 1962; Lindley, 1958; Tukey, 1962; Wilson, 1961). Much caution must be used in interpreting significance levels, for the obtained level of significance is not consistent from sample to sample or among comparable tests (Pratt, 1961b). The obtained level of significance is dependent upon conditions which do not modify the obtained level (Birnbaum, 1962). There are many conditions which vitiate the accuracy of the obtained significance levels, even for robust test procedures (Tukey, 1962). Often graphical representation indicates the presence of disturbing conditions, but social scientists do not use such a procedure regularly (Tukey, 1962). Classical significance procedures are not sensitive to the size of the error (Wilson, 1961); often it is the size of the error that an investigator wishes to know to assess his procedures.

An artificial level of significance is not a sufficient or an efficient indication of the presence or absence of evidence; yet editorial policies have chosen to ignore this fact in their insistence on extreme levels of significance (Melton, 1962). Publication policies have posed other problems for the interpretation of test results. Sterling (1959) has commented upon the prevalence of Type I errors among psychological publications associated with the suppression of nonsignificant results as a consequence of the inclination of editors to publish mainly articles with statistically significant findings. Cohen (1962) has found that articles on social and abnormal psychology have not been appropriately concerned with the power of the designs and tests used.

Significance Tests and Statistical Assumptions

The assumptions underlying all phases of statistical methods constitute pervasive problems in statistical practice (Savage, 1957; Savage and others, 1962). Rarely can an investigator be certain that the assumptions are ade-

quately fulfilled in any particular problem (Tukey, 1960a, 1962). Among the important problems regarding assumptions in ordinary statistical practice are the following: (a) the appropriateness of the probability model to characterize the population (Savage, 1957; Neyman, 1960); (b) the inappropriate use of small sample methods before large sample methods have been used to make estimates of variance (Nunnally, 1960); (c) the robustness or sensitivity of statistical methods to violations of such assumptions as normality, independence, homogeneity, additivity (Boneau, 1960); (d) the selection of test procedures with due awareness for the performance and the requirements of the procedures and the nature of the data (Binder, 1959; Boneau, 1960; Kendall and Stuart, 1961; Lehmann, 1959); (e) the use of transformations of data to meet the assumptions of the more powerful statistical methods versus the use of the less powerful distribution-free and nonparametric methods (Boneau, 1962; Lehmann, 1959; Savage, 1957). Always the investigator must remember that different tests based on different assumptions when applied to varying sets of data lead to different results (Tukey, 1962). Often the precision promised is not the precision delivered (Tukey, 1954, 1962).

Another basic assumption underlying classical statistical methods is that of randomization. Psychologists have been criticized for not being sufficiently concerned with random sampling procedures, the definition of target populations, and the limitations in the representative nature of most samples studied (Kish, 1959; McGinnis, 1958; Nunnally, 1960; Stevens, 1960). McGinnis (1958) and Kish (1959) have discussed the validity of significance tests in surveys versus well-controlled experiments.

Data Analysis and Statistical Hypothesis Testing

Tukey (1962) commented that much of current statistical practice has emphasized the development of elaborate methods at the expense of attention to the analysis of data. Instead of concentrating upon probing the nature of the data and upon the questions that concern scientists, statistical practitioners have been prone to commit errors of the third kind—that is, giving exact answers to the wrong questions—which is perhaps the most serious of the three kinds of errors (Schlaifer, 1959). The routine use of significance tests has perpetrated many errors of the third kind, for, as mentioned above, significance tests do not provide the information that scientists need, and, furthermore, they are not the most effective method for analyzing and summarizing data. Tukey (1954, 1962) recommended that attention should be given to adapting methods to problems rather than to forcing problems into particular methods.

Statistics and Experimentation

Statistical methods have been widely used in the sciences in which strict experimental controls are either not feasible or desirable. The usefulness

of statistical methods for designing and analyzing investigations has been widely recognized; however, there have been dissenters, among them Hogben (Hogben, 1957; Stevens, 1960), who have maintained that the most appropriate solution for the many problems associated with the theory and practice of statistics is the replacement of statistical methods by well-controlled experimentation. Lindley (1958), in his response to Hogben's criticisms, contended that statistical methodology requires extensive modifications in order to play an effective part in science, but that statistics will doubtlessly continue to have an important role in all kinds of research.

The problems of the relative emphasis which should be given to considerations of statistical versus experimental design does not admit an easy solution.

Statistics and Psychological Theory

If statistics is to make significant contributions to the testing of scientific hypotheses, then the methodology of statistical hypothesis testing must incorporate theoretical formulations of the subject matter under investigation. Statistical inferences are maximally informative if they are supported by other findings and if they are enmeshed in an expanding network of hypotheses and constructs.

The constitution of a set of mutually exclusive and exhaustive statistical hypotheses requires careful consideration of relevant theoretical alternatives in order to select heuristically valuable alternatives (Savage and others, 1962). Recognition of the limitations in the theoretical formulations and in the statistical methods should motivate prudence against unwarranted statistical and scientific inferences (Jeffreys, 1957).

Bush (1963) remarked that theories are not substantiated by the goodness of fit between the data and a hypothetical statistical model but rather by the estimation of significant parameters in the statistical model so that information about the functions of experimental and theoretical variables can be accumulated. Estimation is essential for the testing of accurate and fruitful theories.

Conclusion

While the so-called crisis in statistics (Hogben, 1957; Lindley, 1958; Stevens, 1960) has not been resolved, it has highlighted many of the shortcomings in such frequently used methods as significance tests, and it has evoked constructive thought about the very foundations of statistical inference (Savage and others, 1962) as well as about the rationale of statistical methods (Tukey, 1954, 1962).

If educational and psychological research workers heed the admonitions and the recommendations of statisticians, a change in many aspects of

statistical practices may be anticipated. At least investigators will be wary of the routine application of significance tests as the main basis for statistical inference.

Bibliography

ANSCOMBE, F. J. "Examination of Residuals." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Berkeley: University of California Press, 1961. Vol. 1, pp. 1-35.

BAHADUR, RAGHU RAJ, and ROBBINS, HERBERT. "The Problem of the Greater Mean." *Annals of Mathematical Statistics* 21: 469-87; December 1950.

BERKSON, JOSEPH. "Tests of Significance Considered as Evidence." *Journal of the American Statistical Association* 37: 325-35; September 1942.

BINDER, ARNOLD. "Considerations of the Place of Assumptions in Correlational Analysis." *American Psychologist* 14: 504-10; August 1959.

BINDER, ARNOLD. "Further Considerations on Testing the Null Hypothesis and the Strategy and Tactics of Investigating Theoretical Models." *Psychological Review* 70: 107-15; January 1963.

BIRNBAUM, ALLAN. "Confidence Curves: An Omnibus Technique for Estimation and Testing Statistical Hypotheses." *Journal of the American Statistical Association* 56: 246-49; June 1961.

BIRNBAUM, ALLAN. "On the Foundations of Statistical Inference." *Journal of the American Statistical Association* 57: 269-326; June 1962.

BOLLES, ROBERT C. "The Difference Between Statistical Hypotheses and Scientific Hypotheses." *Psychological Reports* 11: 639-45; December 1962.

BOLLES, ROBERT, and MESSICK, SAMUEL. "Statistical Utility in Experimental Inference." *Psychological Reports* 4: 223-27; June 1958.

BONEAU, C. ALAN. "The Effects of Violations of Assumptions Underlying the t Test." *Psychological Bulletin* 57: 49-64; January 1960.

BONEAU, C. ALAN. "A Comparison of the Power of the U and t Tests." *Psychological Review* 69: 246-56; May 1962.

BUEHLER, ROBERT J. "Some Validity Criteria for Statistical Inferences." *Annals of Mathematical Statistics* 30: 845-63; December 1959.

BULMER, M. G. "Confirming Statistical Hypotheses." *Journal of the Royal Statistical Society* (Series B, Methodological) 19: 125-32; No. 1, 1957.

BUSH, ROBERT R. "Estimation and Evaluation." *Handbook of Mathematical Psychology*. (Edited by R. Duncan Luce, Robert R. Bush, and Eugene Galanter.) New York: John Wiley & Sons, 1963. Vol. 1, Chapter 8, pp. 429-69.

COHEN, JACOB. "The Statistical Power of Abnormal-Social Psychological Research: A Review." *Journal of Abnormal and Social Psychology* 65: 145-53; September 1962.

COX, D. R. *Planning of Experiments*. New York: John Wiley & Sons, 1958. 308 pp. (a)

COX, D. R. "Some Problems Connected with Statistical Inference." *Annals of Mathematical Statistics* 29: 357-72; June 1958. (b)

EDWARDS, WARD. "Savage Statistics." *Contemporary Psychology* 1: 14-15; January 1956.

EDWARDS, WARD; LINDMAN, HAROLD; and SAVAGE, LEONARD J. "Bayesian Statistical Inference for Psychological Research." *Psychological Review* 70: 193-242; May 1963.

FISHER, SIR RONALD A. *Statistical Methods and Scientific Inference*. New York: Hafner Publishing Co., 1956. 175 pp.

FISHER, SIR RONALD A. *The Design of Experiments*. Seventh edition. New York: Hafner Publishing Co., 1960. 248 pp.

FLANAGAN, JOHN C. "The Dollar Dimension in Testing Decisions." *Contemporary Psychology* 3: 164-66; June 1958.

GAITO, JOHN. "The Bolles-Messick Coefficient of Utility." *Psychological Reports* 4: 595-98; December 1958.

GOOD, I. J. "Significance Tests in Parallel and in Series." *Journal of the American Statistical Association* 53: 799-813; December 1958.

GRANT, DAVID A. "Testing the Null Hypothesis and the Strategy and Tactics of Investigating Theoretical Models." *Psychological Review* 69: 54-61; January 1962.

HARRINGTON, GORDON M. "Statistics' Logic." *Contemporary Psychology* 6: 304-305; September 1961.

HOGBEN, LANCELOT. *Statistical Theory: The Relationship of Probability, Credibility, and Error.* New York: W. W. Norton & Co., 1957. 510 pp.

HOTELLING, HAROLD. "The Impact of R. A. Fisher on Statistics." *Journal of the American Statistical Association* 46: 35-46; March 1951.

HOTELLING, HAROLD. "The Statistical Method and the Philosophy of Science." *American Statistician* 12: 9-14; December 1958.

JEFFREYS, HAROLD. *Scientific Inference.* Second edition. New York: Cambridge University Press, 1957. 236 pp.

JEFFREYS, HAROLD. *Theory of Probability.* Third edition. New York: Oxford University Press, 1961. 447 pp.

KAISER, HENRY F. "Directional Statistical Decisions." *Psychological Review* 67: 160-67; May 1960.

KENDALL, MAURICE G., and STUART, ALAN. *The Advanced Theory of Statistics.* Revised edition. New York: Hafner Publishing Co., 1961. Vol. 2, "Inference and Relationship," 676 pp.

KISH, LESLIE. "Some Statistical Problems in Research Design." *American Sociological Review* 24: 328-38; June 1959.

LEHMANN, E. L. *Testing Statistical Hypotheses.* New York: John Wiley & Sons, 1959. 369 pp.

LINDLEY, D. V. "Statistical Inference." *Journal of the Royal Statistical Society* (Series B, Methodological) 15: 30-76; No. 1, 1953.

LINDLEY, D. V. "Professor Hogben's 'Crisis'—A Survey of the Foundations of Statistics." *Applied Statistics* 7: 186-98; November 1958.

LINDLEY, D. V. "The Use of Prior Probability Distributions in Statistical Inference and Decisions." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1961. Vol. 2, pp. 453-68.

LUBIN, ARDIE. "Statistics." *Annual Review of Psychology.* (Edited by Paul R. Farnsworth, Olga McNemar, and Quinn McNemar.) Palo Alto, Calif.: Annual Reviews, 1962. Vol. 13, pp. 345-70.

LUCE, R. DUNCAN, and RAIFFA, HOWARD. *Games and Decisions: Introduction and Critical Survey.* New York: John Wiley & Sons, 1957. 509 pp.

MCGINNIS, ROBERT. "Randomization and Inference in Sociological Research." *American Sociological Review* 23: 408-14; August 1958.

MCNEMAR, QUINN. "At Random: Sense and Nonsense." *American Psychologist* 15: 295-300; May 1960.

MELTON, ARTHUR W. "Editorial." *Journal of Experimental Psychology* 64: 553-57; December 1962.

MOWRER, O. HOBART. *Learning Theory and the Symbolic Processes.* New York: John Wiley & Sons, 1960. 473 pp.

NATRELLA, MARY G. "The Relation Between Confidence Intervals and Tests of Significance." *American Statistician* 14: 20-22, 38; February 1960.

NEYMAN, JERZY. "Indeterminism in Science and New Demands on Statisticians." *Journal of the American Statistical Association* 55: 625-39; December 1960.

NUNNALLY, JUM. "The Place of Statistics in Psychology." *Educational and Psychological Measurement* 20: 641-50; Winter 1960.

PEARSON, E. S. "Some Thoughts on Statistical Inference." *Annals of Mathematical Statistics* 33: 394-403; June 1962.

PRATT, JOHN W. "Length of Confidence Intervals." *Journal of the American Statistical Association* 56: 549-67; September 1961. (a)

PRATT, JOHN W., reviewer. "*Testing Statistical Hypotheses* by E. L. Lehmann." *Journal of the American Statistical Association* 56: 163-67; March 1961. (b)

RAIFFA, HOWARD, and SCHLAIFER, ROBERT. *Applied Statistical Decision Theory.* Boston: Division of Research, Harvard Business School, Harvard University, 1961. 356 pp.

ROBERTS, HARRY V., reviewer. "*Applied Statistical Decision Theory* by Howard Raiffa and Robert Schlaifer." *Journal of the American Statistical Association* 57: 199-202; March 1962.

ROZEBOOM, WILLIAM W. "The Fallacy of the Null-Hypothesis Significance Test." *Psychological Bulletin* 57: 416-28; September 1960.

SAVAGE, I. RICHARD. "Nonparametric Statistics." *Journal of the American Statistical Association* 52: 331-44; September 1957.
SAVAGE, LEONARD J. *The Foundations of Statistics.* New York: John Wiley & Sons, 1954. 294 pp.
SAVAGE, LEONARD J. "The Foundations of Statistics Reconsidered." *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability.* Berkeley: University of California Press, 1961. Vol. 1, pp. 575-86.
SAVAGE, LEONARD J., and OTHERS. *The Foundations of Statistical Inference.* New York: John Wiley & Sons, 1962. 112 pp.
SCHLAIFER, ROBERT. *Probability and Statistics for Business Decisions.* New York: McGraw-Hill Book Co., 1959. 732 pp.
SELVIN, HANAN C. "A Critique of Tests of Significance in Survey Research." *American Sociological Review* 22: 519-27; October 1957.
STERLING, THEODOR D. "Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance—or Vice Versa." *Journal of the American Statistical Association* 54: 30-34; March 1959.
STERLING, THEODOR D. "What Is So Peculiar About Accepting the Null Hypothesis?" *Psychological Reports* 7: 363-64; October 1960.
STEVENS, S. S. "The Predicament in Design and Significance." *Contemporary Psychology* 5: 273-76; September 1960.
TUKEY, JOHN W. "Unsolved Problems of Experimental Statistics." *Journal of the American Statistical Association* 49: 706-31; December 1954.
TUKEY, JOHN W. "Conclusions vs. Decisions." *Technometrics* 2: 423-33; November 1960. (a)
TUKEY, JOHN W. "Where Do We Go from Here?" *Journal of the American Statistical Association* 55: 80-93; March 1960. (b)
TUKEY, JOHN W. "The Future of Data Analysis." *Annals of Mathematical Statistics* 33: 1-67; March 1962.
WALD, ABRAHAM. *Statistical Decision Functions.* New York: John Wiley & Sons, 1950. 179 pp.
WILLIAMS, E. J. *Regression Analysis.* New York: John Wiley & Sons, 1959. 214 pp.
WILSON, KELLOGG V. "Subjectivist Statistics for the Current Crisis." *Contemporary Psychology* 6: 229-31; July 1961.
YATES, F. "The Influence of *Statistical Methods for Research Workers* on the Development of the Science of Statistics." *Journal of the American Statistical Association* 46: 19-34; March 1951.