NONPARAMETRIC STATISTICS*

SIDNEY SIEGEL

The Pennsylvania State University [1]

In the development of modern techniques of statistical inference, the first tests to gain prominence and wide use were those which make a good many assumptions, and rather stringent ones, about the nature of the population from which the observations were drawn. The title of these tests—*parametric*—suggests the central importance of the population, or its parameters, in their use and interpretation. These tests also use the operations of arithmetic in the manipulation of the scores on which the inference is to be based, and therefore they are useful only with observations which are numerical. The t and F tests are the most familiar and widely used of the parametric tests, and the Pearson product-moment correlation coefficient and its associated significance test are the most familiar parametric approaches to assessing association.

More recently, *nonparametric* or "distribution-free" statistical tests have gained prominence. As their title suggests, these tests do not make numerous or stringent assumptions about the population. In addition, most nonparametric tests may be used with non-numerical data, and it is for this reason that many of them are often referred to as "ranking tests" or "order tests." Many nonparametric tests use as their data the ranks of the observations, while others are useful with data for which even ordering is impossible, i.e., classificatory data.

Some nonparametric methods, such as the χ^2 tests, the Fisher exact probability test, and the Spearman rank correlation, have long been among the standard tools of the statisticians. Others are relatively new, and therefore have not yet gained such widespread use. At present, however, nonparametric tests are available for all the common experimental designs.

The purpose of the present paper is not to discuss the rationale and application of the various nonparametric tests. That has been done, at various levels of sophistication and with varying degrees of comprehensiveness, in other sources (2, Chap. 17; 3; 6, Chap. 16; 7; 8; 9; 10; 11; 13). Rather, the purpose is to discuss, at a non-technical level, certain issues which have arisen in connection with the use of nonparametric tests. In particular, issues are discussed which are relevant to the choice among alternative tests, parametric and nonparametric, applicable to the same experimental design.

* EDITOR'S NOTE: Articles of an expository, nontechnical nature similar to this one are desired in the "newer" fields in statistics for possible publication in *The American Statistician*.

[1] At the Center for Advanced Study in the Behavioral Sciences, Stanford, California, for the 1957-1958 year.

Power

When alternative statistical tests are available to treat data from a given research design, as is very often the case, it is necessary for the researcher to employ some rationale in choosing among them. The criterion most often suggested is that the researcher should choose the most powerful test.

The *power* of a test is defined as the probability that the test will reject the null hypothesis when in fact it is false and should be rejected. That is,

Power = 1 — probability of a Type II error

Thus, a statistical test is considered a good one if it has small probability of rejecting H_0 the null hypothesis when H_0 is true, but a large probability of rejecting H_0 when H_0 is false.

However, there are considerations other than power which enter into the choice of a statistical test. One must consider the nature of the population from which the sample was drawn, and the kind of measurement which was employed in the operational definitions of the variables of the research. These matters also enter into determining which statistical test is optimal for analyzing a particular set of data.

It is suggested here that the choice among statistical tests which might be used with a given research design should be based on these three criteria:

1. The statistical model of the test should fit the conditions of the research.
2. The measurement requirement of the test should be met by the measures used in the research.
3. From among those tests with appropriate statistical models and appropriate measurement requirements, that test should be chosen which has greatest power-efficiency.

The Statistical Model

When we have asserted the nature of the population and the manner of sampling in the research, we have established a statistical model on the basis of which we may conduct a statistical test. The validity of the conclusion based on the statistical test depends on whether or not the conditions of the statistical model underlying the test are met. That is, the conclusion based on a statistical test carries a qualifier: "If the model used was correct, then we may conclude that . . ."

Sometimes the researcher is able to determine whether the conditions of a particular statistical model are met in his research, but more often he simply has to assume that they are met. Thus the conditions of a statistical

model of a test are often called the "assumptions" of the test.

The model underlying the common parametric tests, the t and F tests, imposes these conditions: (a) the observations must be independent, (b) the observations must be drawn from normally distributed populations, (c) these populations must have the same variance, or, in special cases, they must have a known ratio of variances, and (d) in the case of the analysis of variance, the means of these normal and homoscedastic populations must be linear combinations of effects due to columns and/or rows—the effects must be additive. In addition, as we shall note further below, the nature of the t and F tests also imposes a measurement requirement: a test on the means imposes the requirement that the measures must be additive, i.e., numerical.

As we have already noted in the discussion of their title, the nonparametric tests are not based on a statistical model which specifies such restrictive conditions. In addition to assuming that the observations are independent, some nonparametric tests assume that the variable under study has underlying continuity. Moreover, the measurement requirement of nonparametric tests is weaker; as will be shown, most nonparametric tests require either ranking or classificatory measurement.

Compared to the models underlying parametric tests, the models underlying nonparametric tests are far less restrictive, and therefore the conclusions based on nonparametric inference are more general. When a parametric test, say the t test, is used for inference, we must preface our conclusions with a statement like "If the observations are truly numerical, and are drawn from normally distributed populations which are equal in variance, then we may conclude that . . .", whereas when a nonparametric test is used for inference, we may say, "Regardless of the nature of the underlying populations, we may conclude that . . ."

By the criterion of generality, then, the nonparametric tests are preferable to the parametric. By the single criterion of power, however, the parametric tests are superior, precisely because of the strength of their assumptions; with data for which the strong and extensive assumptions and requirements associated with the parametric tests are valid, these tests are most likely of all tests to reject H_0 when H_0 is false.

Attracted by the power of parametric tests, and seeking to justify their use of these tests with their data, researchers have developed certain approaches in an attempt to determine whether the assumptions of the parametric tests are valid for their data.

For example, in connection with the assumption that the scores are drawn from a normally distributed population, it is common practice to test the normality of the distribution of the scores in the sample by use of say the χ^2 goodness of fit test. If this test does not lead to the rejection of H_0, the researcher concludes that he may safely use tests whose statistical models pose the condition that the population must be normally distributed. At least two objections to this procedure may be raised: (a) it involves an attempt to "prove" the null hypothesis that the sample is from a normally distributed population—the statistical test is employed in order to enable the researcher to *accept* that H_0, and (b) ambiguous and difficult situations arise when the obtained probability of deviations from normality as large as those observed in the sample is close to the arbitrarily set significance level.

Similar objections may be raised to comparable attempts to justify the homoscedasticity assumption by attempting to "prove" the null hypothesis that the variances of the two or more samples do not differ.

When the investigator's test of his data indicates that the obtained sample of scores could well have been drawn from a population which is not normal, his earnest wish to justify the use of the most powerful test leads him to alter the distribution of scores. By a mathematical operation on the original scores, he "transforms" them so that the normality assumption becomes tenable. The question which must be raised in connection with such an attempt is this: Will the process of "normalizing" the distribution by altering the numerical values of the scores cause a distortion of the experimental effect under investigation? This is a question which the investigator may or may not be able to answer. If the process of transforming the scores has the effect of diminishing the experimental effect, then the investigator has placed himself in a paradoxical situation. The steps he has taken in order to justify the use of a statistical test which has maximum capacity to reject H_0 when it should be rejected are steps which have reduced the sensitivity of the measurement and have thereby reduced the likelihood that H_0 will be rejected when it should be. That is, his efforts to gain power paradoxically result in a loss of power.

When the research involves the comparison of scores in two or more samples and when a test of their variances renders the homoscedasticity assumption questionable, the procedure of transforming scores in order to justify that assumption is open to the same objection.

When we have reason to believe that the conditions of the parametric model are met in the data under analysis, then we should certainly choose a parametric statistical test, such as a t or F test, for analyzing those data, because of the power of parametric tests. But if the assumptions of the test are not met, then it is difficult if not impossible to say what is really the power of the parametric test. It is even difficult to estimate the extent to which a probability statement

about the hypothesis is meaningful when that probability statement results from the inappropriate application of the test. Although some empirical evidence has been gathered to show that slight deviations in meeting the assumptions underlying parametric tests may not have radical effects on the obtained probability figure, there is as yet no general agreement as to what constitutes a "slight" deviation.

Measurement

The computation of any statistic involves performing certain manipulations of the research data. To compute a mean, for example, we perform the arithmetic manipulations of addition and division on the scores.

The manipulations to which observations may meaningfully be subjected depend on the sort of measurement which the observations represent. For example, if the observations represent non-numerical measurement, the computation of a mean to represent the central tendency of the observations introduces distortion.

Different statistical tests require different kinds of manipulations of the research observations, and therefore different statistical tests are useful in making inferences from data representing measurement of different strengths or levels.

Levels of measurement. In general, we can clearly define at least four distinct levels at which measurement may be achieved (12), when measurement is understood to mean the process of assigning symbols to observations in some consistent manner.

Measurement is weakest when the objects in the universe are simply partitioned into mutually exclusive classes. This system of classes constitutes a *nominal* or *classificatory* scale. Each class may be represented by a letter, a name, a number, or some other symbol. The only relation which holds in the nominal scale is the relation of equivalence, which holds between entities in the same class. We use nominal scaling in identifying the fields of scholarly endeavor: we assign a scholar to a class in a nominal scale when we say he is a "physicist," "linguist," "biochemist," or "historian." Often numbers are used as the symbols in nominal scaling. The numbers on automobile plates and postal zone numbers are examples.

When the objects in the various classes of a scale stand in some kind of relation to one another, the scale is an *ordinal* or *ranking* scale. The fundamental difference between a nominal and an ordinal scale is that the ordinal scale incorporates not only the relation of equivalence ($=$) but also the relation "greater than" ($>$). Given a group of equivalence classes, if the relation $>$ holds between some but not all pairs of classes, we have a partially ordered scale. If the relation $>$ holds for all pairs of classes so that a complete rank ordering of classes arises, we have an ordinal scale. In the academic world, professorial positions stand in an ordinal relation to one another: professor $>$ associate professor $>$ assistant professor $>$ instructor. The use of numbers to represent an ordinal relation is illustrated by Civil Service job classifications (GS12 $>$ GS11 $>$ GS10) and by street addresses.

When the distances between any two classes on a scale are known numerically, the classes fall on an *interval* scale. An interval scale has all the characteristics of an ordinal scale, and in addition has a common and constant unit of measurement which assigns a real number to all pairs of objects in the ordered set. On an interval scale, the ratio of any two intervals is independent of the unit of measurement and of the zero point, both of which are arbitrary. Our two scales to measure temperature, the Fahrenheit and centigrade scales, are both examples of interval scales.

When a scale has all the characteristics of an interval scale and in addition has a true zero point as its origin, it is called a *ratio* scale. On a ratio scale, the ratio of any two scale points is independent of the unit of measurement. We measure mass or weight in a ratio scale. The scale of ounces and pounds has a true zero point, as does the scale of grams. The ratio between any two weights is independent of the unit of measurement. For example, if we should determine the weights of two different objects not only in pounds but also in grams, we would find that the ratio of the two pound weights would be identical to the ratio of the two gram weights.

Permissible operations and appropriate statistics. The purpose of the preceding discussion of levels of measurement is to remind the reader that at different times we use typographically identical numbers to represent observations and coding procedures of widely varying strengths. The manipulations which may meaningfully be performed on a set of numbers depend on the strength of measurement which the numbers represent. It is clear, for example, that while it is certainly meaningful to add two weights (when we combine the contents of a two-pound box of candy with the contents of a one-pound box we will indeed have three pounds of candy), there is no comparable simple or useful meaning to the sum of two automobile license numbers or the sum of two street addresses.

Each of the four levels of measurement has certain appropriate manipulations associated with it. In order to be able to make certain operations with numbers that have been assigned to operations, the structure of our method of mapping numbers (assigning scores) must be *isomorphic* to some numerical structure which includes these operations. If two systems are isomorphic, their structures are the same in the relations and operations they allow.

In a *nominal* scale, the information may be equally well represented by any set of symbols, as long as the

equivalence relation is preserved. That is, *the nominal scale is unique up to a one-to-one transformation:* the symbols designating the various classes in the scale may be changed or even exchanged, if this is done consistently and completely. Thus, a postal area may be rezoned, with blocks formerly in zone 5 now falling in zone 12, and those formerly in zone 8 now falling in zone 5, etc., and no information will be lost in the rezoning if it is accomplished consistently and throughly. With nominal data, the meaningful statistics are those whose information would be unchanged by a one-to-one transformation: frequency counts, the mode, etc. Under certain conditions, we can test hypotheses regarding the distribution of cases among classes by using statistical tests which use frequencies in unordered categories, i.e., enumerative data. The χ^2 tests are of this type. The most common measure of association for classificatory data is the contingency coefficient. All of these are nonparametric statistics.

The information in an *ordinal* scale may be equally well represented by any ordered set of symbols. That is, *the ordinal scale is unique up to a monotonic transformation*—any order-preserving transformation does not diminish the information it encodes. At present, a corporal wears two stripes and a sergeant wears three. These insignia denote that sergeant $>$ corporal. This relation would be as well expressed if the corporal wore four stripes and sergeant wore seven. The statistic most appropriate for describing the central tendency of scores in an ordinal scale is the median, for the median is not affected by changes of any scores which are above or below it as long as the number of scores above and below remains the same. With ordinal data, hypotheses can be tested by using that large group of nonparametric statistical tests which are sometimes called "order tests" or "ranking tests." Correlation coefficients based on rankings, e.g., those of Spearman and Kendall, are appropriate.

Some ranking tests assume that there is a continuum underlying the observed ranks. Such an assumption is frequently quite tenable even though the grossness of our measuring devices obscures the underlying continuity. For example, although we may classify college students at the end of their college careers in only three ranks—graduating with honors, graduating, and failing to graduate—underlying this ranking is a continuum of achievement in college. Data based on such a ranking could appropriately be subjected to a test which carried the assumption of underlying continuity.

If a variate is truly continuously distributed, then the probability of a tie between two observations is zero. However, tied ranks and tied scores frequently occur in research data. Tied scores are almost invariably a reflection of the lack of sensitivity of our measuring instruments, which fail to distinguish small differences which in fact do exist between the tied observations—they "exist" in the sense that a more sensitive measuring instrument would distinguish them. Therefore, even when ties are observed it may not be unreasonable to assume that a continuous distribution underlies the observations. Most nonparametric techniques incorporate a correction for tied observations.

In the case of an *interval* scale, any change in the numbers associated with the positions of the objects on the scale must preserve not only the ordering of the objects but also the relative differences between them. That is, *an interval scale is unique up to a linear transformation.* For example, although for a given heat the readings on our two temperature scales, centigrade and Fahrenheit, would differ, both scales contain the same amount and the same kind of information—they are linearly related. Although the two scales have a different zero point and a different unit of measurement, a reading on one scale can be transformed to the equivalent reading on the other by the linear transformation $F = 9/5\ C + 32$, where $F =$ number of degrees on the Fahrenheit scale and $C =$ number of degrees on the centigrade scale. It can be shown that the ratios of temperature differences (intervals) are independent of the unit of measurement and of the zero point. Some readings of the same heat on the two scales are:

Centigrade	0	10	30	100
Fahrenheit	32	50	86	212

Notice that the ratio of the differences between temperature readings on one scale is equal to the ratio between the equivalent differences on the other scale. For example, on the centigrade scale the ratio of the differences between 30 and 10, and 10 and 0, is $30 - 10/10 - 0 = 2$. For the comparable readings on the Fahrenheit scale, the ratio is $86 - 50/50 - 32 = 2$.

The interval scale is a quantitative (numerical) scale and statistics which are obtained by treating scores as numbers (such as the mean, the standard deviation, the Pearson product-moment correlation coefficient, etc.) may appropriately be used to represent data based on interval scaling. Most parametric statistical tests, including the t and F tests, are applicable to such data.

A ratio scale is unique up to multiplication by a positive constant. That is, the ratios between any two numerical observations on the scale are preserved when the scale values are all multiplied by a positive constant, and thus such a transformation does not alter the information encoded in the scale. Any statistical test is usable when ratio measurement has been achieved. In addition to those statistics previously mentioned as being appropriate for data in an interval scale, with a ratio scale one may meaningfully use such statistics as the geometric mean and the coefficient of variation—

Table 1—Four Levels of Measurement and the Statistics Appropriate to Each Level

Scale	Defining Relations	Examples of Appropriate Statistics	Appropriate Statistical Tests
Nominal	(1) Equivalence	Mode Frequency Contingency coefficient	Nonparametric statistical tests
Ordinal	(1) Equivalence (2) Order	Median Percentile Spearman r_s Kendall τ Kendall W	
Interval	(1) Equivalence (2) Order (3) Ratio of intervals	Mean Standard deviation Pearson product-moment correlation Multiple product-moment correlation	Nonparametric and parametric statistical tests
Ratio	(1) Equivalence (2) Order (3) Ratio of intervals (4) Ratio of values	Geometric mean Coefficient of variation	

statistics which require knowledge of the true zero point.

Table 1 summarizes the discussion which has been presented concerning the relation between the strength of measurement represented by the data and the statistics and statistical tests which are appropriate. Of course this presumes that the assumptions of the tests' statistical models are satisfied.

Power-Efficiency

The researcher may find that the test which suits the level of measurement he has achieved and whose statistical model is appropriate to the conditions of his research is not the most powerful test available. Confronted by the dilemma posed by the contradictory outcomes of the criteria of power and appropriateness, the researcher may resolve the conflict by choosing the more appropriate test and then enlarging his sample in order to increase the power of that test. The assertion that a test with greater generality is usually weaker in the test of H_0 than is a test restricted by many assumptions is generally true for any given sample size. But it may very well not be true in a comparison of two statistical tests which are applied to two samples of unequal size. That is, with samples of say 30, test A may be more powerful than test B. But test B may be more powerful with a sample of 30 than test A is with a sample of 20.

The concept of *power-efficiency* is concerned with the amount of increase in sample size which is necessary to make test B as powerful as is test A with a given sample size. If test A is the most powerful known test of its type (when used with data which meet the conditions of its statistical model) and if test B is another test suitable for the same research design which is just as powerful with N_b cases as test A is with N_a cases, then

$$\text{Power-efficiency of test B} = 100 \frac{N_a}{N_b} \text{ per cent}$$

For example, if test B requires a sample of $N_b = 30$ cases to have the same power that test A has with $N_a = 27$ cases, then test B has power-efficiency of (100) 27/30 percent, i.e., its power-efficiency is 90 per cent. This means that in order to equate the power of test A with the power of test B (when all the conditions of both tests are met by the data, and when test A is the more powerful), we need to draw 10 cases for test B for every 9 cases drawn for test A.

Relative to the *t* and *F* test, the nonparametric tests suitable for testing hypotheses analogous to those tested by the *t* and *F* tests vary in power-efficiency from 63 percent to 100 per cent. The weaker ones are the tests which use classificatory data—for example, if scores suitable for treatment by the *t* test were split at their

median and tested for differences in location by the median test, the power-efficiency of that test would be about 63 per cent. Many tests which use ranked data have power-efficiency around 95 per cent. The randomization tests, which are used when the scores have numerical meaning, have 100 per cent power-efficiency.

In considering these values, the reader is cautioned to remember that they compare the power of parametric and nonparametric tests *when used with data for which the parametric tests are appropriate,* i.e., when used with data which meet the assumptions and requirements of the statistical model for parametric tests. That is, when we say, for example, that the Mann-Whitney test has power-efficiency of about 95 per cent, we mean that when the Mann-Whitney test is used on two samples of scores which were independently drawn, from normally distributed populations of equal variance, then with $N = 40$ it will reject H_0 at the same level of significance that the t test will with $N = 38$. But if the two tests were both used with data from non-normal populations or with data from populations differing in variance, the Mann-Whitney test might very well reject H_0 at a more stringent significance level than would a t test. Whitney (14) has shown that for some population distributions a nonparametric statistical test is clearly superior in power to a parametric one. With such data, we must rely on the inference based on the nonparametric test, for with such data a t test is inappropriate and therefore yields less meaningful results.

Advantages of Nonparametric Tests

At present, for a great many research designs, both parametric and nonparametric statistical tests are available. We have suggested that the choice between the alternative tests suitable for a given research design should be based on three criteria: (a) the applicability of the statistical models on which the tests are based to the data of the research, (b) the level of measurement achieved in the research, and (c) the power-efficiency of each alternative test. In terms of these criteria, nonparametric tests have certain merits. The enumeration of these may serve as a summary of the arguments presented above, and will introduce certain additional considerations as well.

1. Probability statements obtained from most nonparametric statistical tests are *exact* probabilities (except in the case of large samples, for which excellent approximations are available), regardless of the shape of the population distribution from which the random sample was drawn. That is, a conclusion based on a nonparametric test does not carry stringent qualifiers, as does a conclusion based on a parametric test.

2. If samples as small as 6 are used, there is no alternative to using a nonparametric statistical test unless the nature of the population distribution is known exactly. This is an advantage in pilot testing and in research with populations whose nature precludes the use of large samples (e.g., populations of persons having a rare form of illness).

3. There are suitable nonparametric statistical tests for treating samples made up of observations from several different populations. None of the parametric tests can handle such data unless seemingly unrealistic assumptions are made.

4. Nonparametric statistical tests are available to treat data which are inherently in ranks as well as scores which have merely the strength of ranks. In many fields of investigation, ordinal measurement is the strongest that usually can be achieved. This is the case, for example, in the behavioral sciences. Such data, as well as those for which only gross ordering (plus or minus, for example) can be achieved, can be treated by nonparametric methods, whereas they cannot be treated by parametric methods unless precarious, untestable, and perhaps unrealistic assumptions are made about the underlying distributions.

5. Nonparametric methods are available to treat classificatory data. No parametric technique applies to such data.

6. Nonparametric statistical tests are typically much easier to learn and to apply than are parametric tests.

The advantage which parametric tests hold over their nonparametric counterpart is, of course, that if all the assumptions of the parametric statistical model are in fact met in the research and if the measurement is of the required strength, then nonparametric statistical tests are wasteful of data. The degree of wastefulness in such cases is expressed by the power-efficiency of the nonparametric tests.

Table 2 indicates the variety of nonparametric tests which are now available, and shows the research design and the level of measurement for which each is useful. This list is by no means exhaustive, but an attempt has been made to include a diversity of tests and measures of association and to include those which are most commonly used.

To save space, citations for all tests are not given in this article. In most cases, Table 2 gives the names of the authors of the tests, and the reader may turn to (9) or (10) for references for these tests. The randomization tests were originated by Fisher; early work on their development was presented by Pitman and Welch, and more recently Kempthorne (4, 5) has made important contributions to them. Included in Table 2 are citations for the two most recent tests.

The inclusion of several tests in the same category in Table 2 does not imply that the several tests are equivalent or interchangeable. For example, five tests are listed for use with k independent samples when ordinal

Table 2—Nonparametric Statistical Tests and Measures of Correlation for Various Designs and Various Levels of Measurement

LEVEL OF MEASUREMENT	NONPARAMETRIC STATISTICAL TEST					NONPARAMETRIC MEASURE OF CORRELATION
	One-Sample Case	Two-Sample Case		k-Sample Case		
		Related Samples	Independent Samples	Related Samples or Randomized Blocks	Independent Samples	
Nominal	Binomial test χ^2 test	McNemar test	Fisher exact probability test χ^2 test	Bowker test Cochran Q test	χ^2 test	Contingency coefficient
Ordinal	Kolmogorov-Smirnov test Runs test	Sign test Wilcoxon test	Festinger test Kolmogorov-Smirnov test Mann-Whitney U test Median test Moses test of extreme reactions Wald-Wolfowitz runs test White test Wilcoxon test	Friedman test Wilson test (15)	Extension of the median test Jonckheere test Kruskal-Wallis test Mood runs test Mosteller slippage test Whitney extension of the U test	Cureton biserial rank correlation (1) Kendall rank correlation coefficient Kendall partial rank correlation coefficient Kendall coefficient of concordance Moran multiple rank correlation Spearman rank correlation coefficient
Interval	Walsh test Randomization test	Walsh test Randomization test	Randomization test	Randomization test	Randomization test	Olmstead-Tukey corner test Randomization test

measurement has been achieved. Each of these has a different application. The extension of the median test is useful when only incomplete ordering has been achieved, so that any observation may be classed either above or below the common median. The Kruskal-Wallis test is a more general test for data in which complete ordering has been achieved, and thus it is more powerful than the extension of the median test. Whitney's extension of the U test is not an analogue of the analysis of variance, as is the Kruskal-Wallis test, being a significance test for only three samples which tests the prediction that the three averages will occur in a specific order. The Jonckheere test is a test against ordered alternatives; the Mosteller technique tests whether one group has slipped significantly to the right of the others; and the k-sample runs test is sensitive to any sorts of differences among groups, not just differences in location.

REFERENCES

1. Cureton, E. E., "Rank-Biserial Correlation," *Psychometrika*, 1956, 21, 287-290.
2. Dixon, W. J., and Massey, F. J., *Introduction to Statistical Analysis.* (2nd Ed.) New York: McGraw-Hill, 1957.
3. Fraser, D. A. S., *Nonparametric Methods in Statistics.* New York: Wiley, 1957.
4. Kempthorne, O., *The Design and Analysis of Experiments.* New York: Wiley, 1952.
5. Kempthorne, O., "The Randomization Theory of Experimental Inference," *J. Amer. Statist. Assn.*, 1955, 50, 946-967.
6. Mood, A. M., *Introduction to the Theory of Statistics.* New York: McGraw-Hill, 1950.
7. Moses, L. W., "Non-Parametric Statistics for Psychological Research," *Psychol. Bull.*, 1952, 49, 122-143.
8. Mosteller, F., and Bush, R. R. "Selected Quantitative Techniques." In G. Lindzey (Ed.) *Handbook of Social Psychology.* Vol. 1. *Theory and Method.* Cambridge, Mass.: Addison-Wesley, 1954. Pp. 289-334.
9. Savage, I. R., "Bibliography of Nonparametric Statistics and Related Topics," *J. Amer. Statist. Assn.*, 1953, 48, 844-906.
10. Siegel, S., *Nonparametric Statistics: For the Behavioral Sciences.* New York: McGraw-Hill, 1956.
11. Smith, K., "Distribution-Free Statistical Methods and the Concept of Power Efficiency." In L. Festinger and D. Katz (Eds.), *Research Methods in the Behavioral Sciences.* New York: Dryden, 1953. Pp. 536-577.
12. Stevens, S. S., "On the Theory of Scales of Measurement," *Science*, 1946, 103, 677-680.
13. Tukey, J. W. *The Simplest Signed-Rank Tests.* Mimeographed Report No. 17, Statistical Research Group, Princeton University, 1949.
14. Whitney, D. R. "A Comparison of the Power of Non-Parametric Tests and Tests Based on the Normal Distribution under Non-Normal Alternatives.' Unpublished doctor's dissertation, Ohio State University, 1948.
15. Wilson, K. V., "A Distribution-Free Test of Analysis of Variance Hypotheses," *Psychol. Bull.*, 1956, 53, 96-101.