CHAPTER 3

CHOOSING AN APPROPRIATE STATISTICAL TEST

When alternative statistical tests are available for a given research design, as is very often the case, it is necessary to employ some rationale for choosing among them. In Chap. 2 we presented one criterion to use in choosing among alternative statistical tests: the criterion of power. In this chapter other criteria will be presented.

The reader will remember that the *power* of a statistical analysis is partly a function of the statistical test employed in the analysis. A statistical test is a good one if it has a small probability of rejecting H_0 when H_0 is true, but a large probability of rejecting H_0 when H_0 is false. Suppose we find two statistical tests, A and B, which have the same probability of rejecting H_0 when it is true. It might seem that we should simply select the one that has the larger probability of rejecting H_0 when it is false.

However, there are considerations other than power which enter into the choice of a statistical test. In this choice we must consider the manner in which the sample of scores was drawn, the nature of the population from which the sample was drawn, and the kind of measurement or scaling which was employed in the operational definitions of the variables involved, i.e., in the scores. All these matters enter into determining which statistical test is optimum or most appropriate for analyzing a particular set of research data.

THE STATISTICAL MODEL

When we have asserted the nature of the population and the manner of sampling, we have established a statistical model. Associated with every statistical test is a model and a measurement requirement; the test is valid under certain conditions, and the model and the measurement requirement specify those conditions. Sometimes we are able to test whether the conditions of a particular statistical model are met, but more often we have to *assume* that they are met. Thus the conditions of the statistical model of a test are often called the "assumptions" of the test. All decisions arrived at by the use of any statistical test must

arry with them this qualification: "If the model used was correct, and the measurement requirement was satisfied, then . . ."

It is obvious that the fewer or weaker are the assumptions that define particular model, the less qualifying we need to do about our decision rrived at by the statistical test associated with that model. That is, he fewer or weaker are the assumptions, the more general are the onclusions.

However, the most powerful tests are those which have the strongest r most extensive assumptions. The parametric tests, for example, the or F tests, have a variety of strong assumptions underlying their use. When those assumptions are valid, these tests are the most likely of all ests to reject H_0 when H_0 is false. That is, when research data may ppropriately be analyzed by a parametric test, that test will be more owerful than any other in rejecting H_0 when it is false. Notice, however, the requirement that the research data must be appropriate for the est. What constitutes such appropriateness? What are the conditions that are associated with the statistical model and the measurement equirement underlying, say, the t test? The conditions which must be atisfied to make the t test the most powerful one, and in fact before any onfidence can be placed in any probability statement obtained by the se of the t test, are at least these:

1. The observations must be independent. That is, the selection of any one case from the population for inclusion in the sample must not bias the chances of any other case for inclusion, and the score which is assigned to any case must not bias the score which is assigned to any other case.

2. The observations must be drawn from normally distributed populations.

3. These populations must have the same variance (or, in special cases, they must have a known ratio of variances).

4. The variables involved must have been measured in *at least* an interval scale, so that it is possible to use the operations of arithmetic (adding, dividing, finding means, etc.) on the scores.

In the case of the analysis of variance (the F test), another condition is added to those already given:

5. The means of these normal and homoscedastic populations must be linear combinations of effects due to columns and/or rows. That is, the effects must be additive.

All the above conditions [except (4), which states the measurement requirement] are elements of the parametric statistical model. With the possible exception of the assumption of homoscedasticity (equal variances) these conditions are ordinarily not tested in the course of the performance of a statistical analysis. Rather, they are presumptions

which are accepted, and their truth or falsity determines the meaningfulness of the probability statement arrived at by the parametric test.

When we have reason to believe that these conditions are met in the data under analysis, then we should certainly choose a parametric statistical test, such as t or F, for analyzing those data. Such a choice is optimum because the parametric test will be most powerful for rejecting H_0 when it should be rejected.

But what if these conditions are not met? What happens when the population is *not* normally distributed? What happens when the measurement is *not* so strong as an interval scale? What happens when the populations are *not* equal in variance?

When the assumptions constituting the statistical model for a test are in fact not met, or when the measurement is not of the required strength, then it is difficult if not impossible to say what is really the power of the test. It is even difficult to estimate the extent to which a probability statement about the hypothesis in question is meaningful when that probability statement results from the unacceptable application of a 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.

POWER-EFFICIENCY

We have already noticed that the fewer or weaker are the assumptions that constitute a particular model, the more general are the conclusions derived from the application of the statistical test associated with that model but the less powerful is the test of H_0. This assertion is generally true for any given sample size. But it may not be true in the comparison of two statistical tests which are applied to two samples of unequal size. That is, if $N = 30$ in both instances, test A may be more powerful than test B. But the same test B may be more powerful with $N = 30$ than is test A with $N = 20$. In other words, we can avoid the dilemma of having to choose between power and generality by selecting a statistical test which has broad generality and then increasing its power to that of the most powerful test available by enlarging the size of the sample.

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 test A. If test A is the most powerful known test of its type (when used with data which meet its conditions), and if test B is another test for the same research design which is just as powerful with N_b cases as is test A 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 $N = 25$ cases to have the same power as test A has with $N = 20$ cases, then test B has power-efficiency of $(100)\frac{20}{25}$ per cent, i.e., its power-efficiency is 80 per cent. A power-efficiency of 80 per cent means that in order to equate the power of test A and test B (when all the conditions of both tests are met, and when test A is the more powerful) we need to draw 10 cases for test B for every 8 cases drawn for test A.

Thus we can avoid having to meet some of the assumptions of the most powerful tests, the parametric tests, without losing power by simply choosing a different test and drawing a larger N. In other words, by choosing another statistical test with fewer assumptions in its model and thus with greater generality than the t and F tests, and by enlarging our N, we can avoid having to make assumptions 2, 3, and 5 above, and still retain equivalent power to reject H_0.

Two other conditions, 1 and 4 above, underlie parametric statistical tests. Assumption 1, that the scores are independently drawn from the population, is an assumption which underlies all statistical tests, parametric or nonparametric. But assumption 4, which concerns the strength of measurement required for parametric tests—measurement must be at least in an interval scale—is not shared by all statistical tests. Different tests require measurement of different strengths. In order to understand the measurement requirements of the various statistical tests, the reader should be conversant with some of the basic notions in the theory of measurement. The discussion of measurement which occupies the next few pages gives the required information.

MEASUREMENT

When a physical scientist talks about measurement, he usually means the assigning of numbers to observations in such a way that the numbers are amenable to analysis by manipulation or operation according to certain rules. This analysis by manipulation will reveal new information about the objects being measured. In other words, the relation between the things being observed and the numbers assigned to the observations is so direct that by manipulating the numbers the physical scientist obtains new information about the things. For example, he may determine how much a homogeneous mass of material would weigh if cut in half by simply dividing its weight by 2.

The social scientist, taking physics as his model, usually attempts to

do likewise in his scoring or measurement of social variables. But in his scaling the social scientist very often overlooks a fundamental fact in measurement theory. He overlooks the fact that in order for him to be able to make certain operations with numbers that have been assigned to observations, the structure of his method of mapping numbers (assigning scores) to observations 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.

For example, if a researcher collects data made up of numerical scores and then manipulates these scores by, say, adding and dividing (which are necessary operations in finding means and standard deviations), he is assuming that the structure of his measurement is isomorphic to that numerical structure known as arithmetic. That is, he is assuming that he has attained a high level of measurement.

The theory of measurement consists of a set of separate or distinct theories, each concerning a distinct *level* of measurement. The operations allowable on a given set of scores are dependent on the level of measurement achieved. Here we will discuss four levels of measurement—nominal, ordinal, interval, and ratio—and will discuss the operations and thus the statistics and statistical tests that are permitted with each level.

The Nominal or Classificatory Scale

Definition. Measurement at its weakest level exists when numbers or other symbols are used simply to classify an object, person, or characteristic. When numbers or other symbols are used to identify the groups to which various objects belong, these numbers or symbols constitute a nominal or classificatory scale.

Examples. The psychiatric system of diagnostic groups constitutes a nominal scale. When a diagnostician identifies a person as "schizophrenic," "paranoid," "manic-depressive," or "psychoneurotic," he is using a symbol to represent the class of persons to which this person belongs, and thus he is using nominal scaling.

The numbers on automobile license plates constitute a nominal scale. If the assignment of plate numbers is purely arbitrary, then each plated car is a member of a unique subclass. But if, as is common in the United States, a certain number or letter on the license plate indicates the county in which the car owner resides, then each subclass in the nominal scale consists of a group of entities: all owners residing in the same county. Here the assignment of numbers must be such that the same number (or letter) is given to all persons residing in the same county and that different numbers (or letters) are given to people residing in different counties. That is, the number or letter on the license plate must clearly indicate to which of a set of mutually exclusive subclasses the owner belongs.

Numbers on football jerseys and social-security numbers are other examples of the use of numbers in nominal scaling.

Formal properties. All scales have certain formal properties. These properties provide fairly exact definitions of the scale's characteristics, more exact definitions than we can give in verbal terms. These properties may be formulated more abstractly than we have done here by a set of axioms which specify the operations of scaling and the relations among the objects that have been scaled.

In a nominal scale, the scaling operation is partitioning a given class into a set of mutually exclusive subclasses. The only relation involved is that of *equivalence*. That is, the members of any one subclass must be equivalent in the property being scaled. This relation is symbolized by the familiar sign: $=$. The equivalence relation is reflexive, symmetrical, and transitive.[1]

Admissible operations. Since in any nominal scale the classification may be equally well represented by any set of symbols, the nominal scale is said to be "unique up to a one-to-one transformation." The symbols designating the various subclasses in the scale may be interchanged, if this is done consistently and completely. For example, when new license plates are issued, the license number which formerly stood for one county can be interchanged with that which had stood for another county. Nominal scaling would be preserved if this change-over were performed consistently and thoroughly in the issuing of all license plates. Such one-to-one transformations are sometimes called "the symmetric group of transformations."

Since the symbols which designate the various groups on a nominal scale may be interchanged without altering the essential information in the scale, the only kinds of admissible descriptive statistics are those which would be unchanged by such a transformation: the mode, frequency counts, etc. Under certain conditions, we can test hypotheses regarding the distribution of cases among categories by using the nonparametric statistical test, χ^2, or by using a test based on the binomial expansion. These tests are appropriate for nominal data because they focus on frequencies in categories, i.e., on enumerative data. The most common measure of association for nominal data is the contingency coefficient, C, a nonparametric statistic.

The Ordinal or Ranking Scale

Definition. It may happen that the objects in one category of a scale are not just different from the objects in other categories of that scale,

[1] *Reflexive:* $x = x$ for all values of x. *Symmetrical:* if $x = y$, then $y = x$. *Transitive:* if $x = y$ and $y = z$, then $x = z$.

but that they stand in some kind of *relation* to them. Typical relations among classes are: higher, more preferred, more difficult, more disturbed, more mature, etc. Such relations may be designated by the carat (>) which, in general, means "greater than." In reference to particular scales, > may be used to designate *is preferred to, is higher than, is more difficult than*, etc. Its specific meaning depends on the nature of the relation that defines the scale.

Given a group of equivalence classes (i.e., given a nominal scale), 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*.

Examples. Socioeconomic status, as conceived by Warner and his associates,[1] constitutes an ordinal scale. In prestige or social acceptability, all members of the upper middle class are higher than (>) all members of the lower middle class. The lower middles, in turn, are higher than the upper lowers. The = relation holds among members of the same class, and the > relation holds between any pair of classes.

The system of grades in the military services is another example of an ordinal scale. Sergeant > corporal > private.

Many personality inventories and tests of ability or aptitude result in scores which have the strength of ranks. Although the scores may appear to be more precise than ranks, generally these scales do not meet the requirements of any higher level of measurement and may properly be viewed as ordinal.

Formal properties. Axiomatically, 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" (>). The latter relation is irreflexive, asymmetrical, and transitive.[2]

Admissible operations. Since any order-preserving transformation does not change the information contained in an ordinal scale, the scale is said to be "unique up to a monotonic transformation." That is, it does not matter what numbers we give to a pair of classes or to members of those classes, just as long as we give a higher number to the members of the class which is "greater" or "more preferred." (Of course, one may use the lower numbers for the "more preferred" grades. Thus we usually refer to excellent performance as "first-class," and to progressively inferior performances as "second-class" and "third-class." So long as we are consistent, it does not matter whether higher or lower numbers are used to denote "greater" or "more preferred.")

[1] Warner, W. L., Meeker, M., and Eells, K. 1949. *Social class in America.* New York: Science Research Associates.

[2] *Irreflexive:* it is not true for any x that $x > x$. *Asymmetrical:* if $x > y$, then $y \not> x$. *Transitive:* if $x > y$ and $y > z$, then $x > z$.

For example, a corporal in the army wears two stripes on his sleeve and a sergeant wears three. These insignia denote that sergeant > corporal. This relation would be equally well expressed if the corporal wore four stripes and the sergeant wore seven. That is, a transformation which does not change the order of the classes is completely admissible because *it does not involve any loss of information.* Any or all the numbers applied to classes in an ordinal scale may be changed in any fashion which does not alter the ordering (ranking) of the objects.

The statistic most appropriate for describing the central tendency of scores in an ordinal scale is the median, since 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 scaling, hypotheses can be tested by using that large group of nonparametric statistical tests which are sometimes called "order statistics" or "ranking statistics." Correlation coefficients based on rankings (e.g., the Spearman r_S or the Kendall τ) are appropriate.

The only assumption made by some ranking tests is that the scores we observe are drawn from an underlying continuous distribution. Parametric tests also make this assumption. An underlying continuous variate is one that is not restricted to having only isolated values. It may have any value in a certain interval. A discrete variate, on the other hand, is one which can take on only a finite number of values; a continuous variate is one which can (but may not) take on a continuous infinity of values.

For some nonparametric techniques which require ordinal measurement, the requirement is that there be a continuum *underlying* the observed scores. The actual scores we observe may fall into discrete categories. For example, the actual scores may be either "pass" or "fail" on a particular item. We may well assume that underlying such a dichotomy there is a continuum of possible results. That is, some individuals who were categorized as failing may have been closer to passing than were others who were categorized as failing. Similarly, some passed only minimally, whereas others passed with ease and dispatch. The assumption is that "pass" and "fail" represent a continuum dichotomized into two intervals.

Similarly, in matters of opinion those who are classified as "agree" and "disagree" may be thought to fall on a continuum. Some who score as "agree" are actually not very concerned with the issue, whereas others are strongly convinced of their position. Those who "disagree" include those who are only mildly in disagreement as well as die-hard opponents.

Frequently the grossness of our measuring devices obscures the underlying continuity that may exist. If a variate is truly continuously distributed, then the probability of a tie is zero. However, tied scores fre-

quently occur. Tied scores are almost invariably a reflection of the lack of sensitivity of our measuring instruments, which fail to distinguish the small differences which really exist between the tied observations. Therefore even when ties are observed it may not be unreasonable to assume that a continuous distribution underlies our gross measures.

At the risk of being excessively repetitious, the writer wishes to emphasize here that parametric statistical tests, which use means and standard deviations (i.e., which require the operations of arithmetic on the original scores), ought not to be used with data in an ordinal scale. The properties of an ordinal scale are *not* isomorphic to the numerical system known as arithmetic. When only the rank order of scores is known, means and standard deviations found on the scores themselves are *in error* to the extent that the successive intervals (distances between classes) on the scale are not equal. When parametric techniques of statistical inference are used with such data, any decisions about hypotheses are doubtful. Probability statements derived from the application of parametric statistical tests to ordinal data are in error to the extent that the structure of the method of collecting the data is not isomorphic to arithmetic. Inasmuch as most of the measurements made by behavioral scientists culminate in ordinal scales (this seems to be the case except in the field of psychophysics, and possibly in the use of a few carefully standardized tests), this point deserves strong emphasis.

Since this book is addressed to the behavioral scientist, and since the scales used by behavioral scientists typically are at best no stronger than ordinal, the major portion of this book is devoted to those methods which are appropriate for testing hypotheses with data measured in an ordinal scale. These methods, which also have much less circumscribing or restrictive assumptions in their statistical models than have parametric tests, make up the bulk of the nonparametric tests.

The Interval Scale

Definition. When a scale has all the characteristics of an ordinal scale, and when in addition the *distances* between any two numbers on the scale are of known size, then measurement considerably stronger than ordinality has been achieved. In such a case measurement has been achieved in the sense of an interval scale. That is, if our mapping of several classes of objects is so precise that we know just how large are the intervals (distances) between all objects on the scale, then we have achieved interval measurement. An interval scale is characterized by a common and constant unit of measurement which assigns a real number to all pairs of objects in the ordered set. In this sort of measurement, the ratio of any two intervals is independent of the unit of measurement

and of the zero point. In an interval scale, the zero point and the unit of measurement are arbitrary.

Examples. We measure temperature on an interval scale. In fact, two different scales—centigrade and Fahrenheit—are commonly used. The unit of measurement and the zero point in measuring temperature are arbitrary; they are different for the two scales. However, both scales contain the same amount and the same kind of information. This is the case because they are linearly related. That is, a reading on one scale can be transformed to the equivalent reading on the other by the linear transformation

$$F = \tfrac{9}{5}C + 32$$

where F = number of degrees on Fahrenheit scale
C = number of degrees on centrigrade scale

It can be shown that the ratios of temperature differences (intervals) are independent of the unit of measurement and of the zero point. For instance, "freezing" occurs at 0 degrees on the centigrade scale, and "boiling" occurs at 100 degrees. On the Fahrenheit scale, "freezing" occurs at 32 degrees and "boiling" at 212 degrees. Some other readings of the same temperature 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 $\frac{30 - 10}{10 - 0} = 2$. For the comparable readings on the Fahrenheit scale, the ratio is $\frac{86 - 50}{50 - 32} = 2$. The ratio is the same in both cases: 2. In an interval scale, in other words, the ratio of any two intervals is independent of the unit used and of the zero point, both of which are arbitrary.

Most behavioral scientists aspire to create interval scales, and on infrequent occasions they succeed. Usually, however, what is taken for success comes because of the untested assumptions the scale maker is willing to make. One frequent assumption is that the variable being scaled is normally distributed in the individuals being tested. Having made this assumption, the scale maker manipulates the units of the scale until the *assumed* normal distribution is recovered from the individuals' scores. This procedure, of course, is only as good as the intuition of the investigator when he hits upon the distribution to assume.

Another assumption which is often made in order to create an apparent interval scale is the assumption that the person's answer of "yes" on any one item is exactly equivalent to his answering affirmatively on any other item. This assumption is made in order to satisfy the requirement that an interval scale have a common and constant unit of measurement. In ability or aptitude scales, the equivalent assumption is that giving the correct answer to any one item is exactly equivalent (in amount of ability shown) to giving the correct answer to any other item.

Formal properties. Axiomatically, it can be shown that the operations and relations which give rise to the structure of an interval scale are such that the *differences* in the scale are isomorphic to the structure of arithmetic. Numbers may be associated with the positions of the objects on an interval scale so that the operations of arithmetic may be meaningfully performed on the *differences* between these numbers.

In constructing an interval scale, one must not only be able to specify equivalences, as in a nominal scale, and greater-than relations, as in an ordinal scale, but one must also be able to specify the ratio of any two intervals.

Admissible operations. Any change in the numbers associated with the positions of the objects measured in an interval scale must preserve not only the ordering of the objects but also the relative differences between the objects. That is, the interval scale is "unique up to a linear transformation." Thus the information yielded by the scale is not affected if each number is multiplied by a positive constant and then a constant is added to this product, that is, $f(x) = ax + b$. (In the temperature example, $a = \frac{9}{5}$ and $b = 32$.)

We have already noticed that the zero point in an interval scale is arbitrary. This is inherent in the fact that the scale is subject to transformations which consist of adding a constant to the numbers making up the scale.

The interval scale is the first truly *quantitative* scale that we have encountered. All the common parametric statistics (means, standard deviations, Pearson correlations, etc.) are applicable to data in an interval scale, as are the common parametric statistical tests (t test, F test, etc.). If measurement in the sense of an interval scale has in fact been achieved, and if all of the assumptions in the statistical model (given on page 19) are adequately met, then the researcher *should* utilize parametric statistical tests. In such a case, nonparametric methods usually would not take advantage of all the information contained in the research data.

The Ratio Scale

Definition. 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. In a ratio scale, the ratio of any two scale points is independent of the unit of measurement.

Example. We measure mass or weight in a ratio scale. The scale of ounces and pounds has a true zero point. So does the scale of grams. The ratio between any two weights is independent of the unit of measurement. For example, if we 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 is identical to the ratio of the two gram weights.

Formal properties. The operations and relations which give rise to the numerical values in a ratio scale are such that the scale is isomorphic to the structure of arithmetic. Therefore the operations of arithmetic are permissible on the numerical values assigned to the objects themselves, as well as on the intervals between numbers as is the case in the interval scale.

Ratio scales, most commonly encountered in the physical sciences, are achieved only when all four of these relations are operationally possible to attain: (*a*) equivalence, (*b*) greater than, (*c*) known ratio of any two intervals, and (*d*) known ratio of any two scale values.

Admissible operations. The numbers associated with the ratio scale values are "true" numbers with a true zero; only the unit of measurement is arbitrary. Thus the ratio scale is "unique up to multiplication by a positive constant." That is, the ratios between any two numbers are preserved when the scale values are all multiplied by a positive constant, and thus such a transformation does not alter the information contained in the scale.

Any statistical test is usable when ratio measurement has been achieved. In addition to using those previously mentioned as being appropriate for use with data in interval scales, with ratio scales one may use such statistics as the geometric mean and the coefficient of variation—statistics which require knowledge of the true zero point.

Summary

Measurement is the process of mapping or assigning numbers to objects or observations. The kind of measurement which is achieved is a function of the rules under which the numbers were assigned. The operations and relations employed in obtaining the scores define and limit the manipulations and operations which are permissible in handling the scores; the manipulations and operations must be those of the numerical structure to which the measurement is isomorphic.

Four of the most general scales were discussed: the nominal, ordinal, interval, and ratio scales. Nominal and ordinal measurement are the most common types achieved in the behavioral sciences. Data measured by either nominal or ordinal scales should be analyzed by the non-

parametric methods. Data measured in interval or ratio scales may be analyzed by parametric methods, if the assumptions of the parametric statistical model are tenable.

Table 3.1 summarizes the information in our discussion of various levels of measurement and of the kinds of statistics and statistical tests which are appropriate to each level when the assumptions of the tests' statistical models are satisfied.

TABLE 3.1. 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) Greater than	Median Percentile Spearman r_S Kendall τ Kendall W	
Interval	(1) Equivalence (2) Greater than (3) Known ratio of any two intervals	Mean Standard deviation Pearson product-moment correlation Multiple product-moment correlation	Nonparametric and parametric statistical tests
Ratio	(1) Equivalence (2) Greater than (3) Known ratio of any two intervals (4) Known ratio of any two scale values	Geometric mean Coefficient of variation	

The reader may find other discussions of measurement in Bergman and Spence (1944), Coombs (1950; 1952), Davidson, Siegel, and Suppes (1955), Hempel (1952), Siegel (1956), and Stevens (1946; 1951).

PARAMETRIC AND NONPARAMETRIC STATISTICAL TESTS

A parametric statistical test is a test whose model specifies certain conditions (given on page 19) about the parameters of the population

from which the research sample was drawn. Since these conditions are not ordinarily tested, they are assumed to hold. The meaningfulness of the results of a parametric test depends on the validity of these assumptions. Parametric tests also require that the scores under analysis result from measurement in the strength of at least an interval scale.

A nonparametric statistical test is a test whose model does not specify conditions about the parameters of the population from which the sample was drawn. Certain assumptions are associated with most nonparametric statistical tests, i.e., that the observations are independent and that the variable under study has underlying continuity, but these assumptions are fewer and much weaker than those associated with parametric tests. Moreover, nonparametric tests do not require measurement so strong as that required for the parametric tests; most nonparametric tests apply to data in an ordinal scale, and some apply also to data in a nominal scale.

In this chapter we have discussed the various criteria which should be considered in the choice of a statistical test for use in making a decision about a research hypothesis. These criteria are (*a*) the power of the test, (*b*) the applicability of the statistical model on which the test is based to the data of the research, (*c*) power-efficiency, and (*d*) the level of measurement achieved in the research. It has been stated that a parametric statistical test is most powerful when all the assumptions of its statistical model are met and when the variables under analysis are measured in at least an interval scale. However, even when all the parametric test's assumptions about the population and requirements about strength of measurement are satisfied, we know from the concept of power-efficiency that by increasing the sample size by an appropriate amount we can use a nonparametric test rather than the parametric one and yet retain the same power to reject H_0.

Because the power of any nonparametric test may be increased by simply increasing the size of N, and because behavioral scientists rarely achieve the sort of measurement which permits the meaningful use of parametric tests, nonparametric statistical tests deserve an increasingly prominent role in research in the behavioral sciences. This book presents a variety of nonparametric tests for the use of behavioral scientists. The use of parametric tests in research has been presented well in a variety of sources[1] and therefore we will not review those tests here.

In many of the nonparametric statistical tests to be presented, the data are changed from scores to ranks or even to signs. Such methods

[1] Among the many sources on parametric statistical tests, these are especially useful: Anderson and Bancroft (1952), Dixon and Massey (1951), Edwards (1954), Fisher (1934; 1935), McNemar (1955), Mood (1950), Snedecor (1946), Walker and Lev (1953).

may arouse the criticism that they "do not use all of the information in the sample" or that they "throw away information." The answer to this objection is contained in the answers to these questions: (*a*) Of the methods available, parametric and nonparametric, which uses the information in the sample most appropriately? (*b*) How important is it that the conclusions from the research apply generally rather than only to populations with normal distributions?

The answer to the first question depends on the level of measurement achieved in the research and on the researcher's knowledge of the population. If the measurement is weaker than that of an interval scale, by using parametric tests the researcher would "add information" and thereby create distortions which may be as great and as damaging as those introduced by the "throwing away of information" which occurs when scores are converted to ranks. Moreover, the assumptions which must be made to justify the use of parametric tests usually rest on conjecture and hope, for knowledge about the population parameters is almost invariably lacking. Finally, for some population distributions a nonparametric statistical test is clearly superior in power to a parametric one (Whitney, 1948).

The answer to the second question can be given only by the investigator as he considers the substantive aspects of the research problem.

The relevance of the discussion of this chapter to the choice between parametric and nonparametric statistical tests may be sharpened by the summary below, which lists the advantages and disadvantages of nonparametric statistical tests.

Advantages of Nonparametric Statistical Tests

1. Probability statements obtained from most nonparametric statistical tests are *exact* probabilities (except in the case of large samples, where excellent approximations are available), regardless of the shape of the population distribution from which the random sample was drawn. The accuracy of the probability statement does not depend on the shape of the population, although some nonparametric tests may assume identity of shape of two or more population distributions, and some others assume symmetrical population distributions. In certain cases, the nonparametric tests do assume that the underlying distribution is continuous, an assumption which they share with parametric tests.

2. If sample sizes as small as $N = 6$ are used, there is no alternative to using a nonparametric statistical test unless the nature of the population distribution is *known exactly*.

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 without requiring us to make seemingly unrealistic assumptions.

4. Nonparametric statistical tests are available to treat data which are inherently in ranks as well as data whose seemingly numerical scores have the strength of ranks. That is, the researcher may only be able to say of his subjects that one has more or less of the characteristic than another, without being able to say *how much* more or less. For example, in studying such a variable as anxiety, we may be able to state that subject *A* is more anxious than subject *B* without knowing at all exactly how much more anxious *A* is. If data are inherently in ranks, or even if they can only be categorized as plus or minus (more or less, better or worse), they can be treated by nonparametric methods, whereas they cannot be treated by parametric methods unless precarious and perhaps unrealistic assumptions are made about the underlying distributions.

5. Nonparametric methods are available to treat data which are simply classificatory, i.e., are measured in a nominal scale. No parametric technique applies to such data.

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

Disadvantages of Nonparametric Statistical Tests

1. If all the assumptions of the parametric statistical model are in fact met in the data, and if the measurement is of the required strength, then nonparametric statistical tests are wasteful of data. The degree of wastefulness is expressed by the power-efficiency of the nonparametric test. (It will be remembered that if a nonparametric statistical test has power-efficiency of, say, 90 per cent, this means that *where all the conditions of the parametric test are satisfied* the appropriate parametric test would be just as effective with a sample which is 10 per cent smaller than that used in the nonparametric analysis.)

2. There are as yet no nonparametric methods for testing interactions in the analysis of variance model, unless special assumptions are made about additivity. (Perhaps we should disregard this distinction because parametric statistical tests are also forced to make the assumption of additivity. However, the problem of higher-ordered interactions has yet to be dealt with in the literature of nonparametric methods.)[1]

Another objection that has been entered against nonparametric methods is that the tests and their accompanying tables of significant values have been widely scattered in various publications, many highly

[1] After this book had been set in type, a nonparametric test was presented which contributes to the solution of this problem. See Wilson, K. V. 1956. A distribution-free test of analysis of variance hypotheses. *Psychol. Bull.*, **53**, 96-101.

specialized, and they have therefore been comparatively inaccessible to the behavioral scientist. In preparing this book, the writer's intention has been to rob that objection of its force. This book attempts to present all the nonparametric techniques of statistical inference and measures of association that the behavioral scientist is likely to need, and it gives all of the tables necessary for the use of these techniques. Although this text is not exhaustive in its coverage of nonparametric tests—it could not be without being excessively redundant—enough tests are included in the chapters which follow to give the behavioral scientist wide latitude in choosing a nonparametric technique appropriate to his research design and useful for testing his research hypothesis.

CHAPTER 4

THE ONE-SAMPLE CASE

In this chapter we present those nonparametric statistical tests which may be used to test a hypothesis which calls for drawing just one sample. The tests tell us whether the particular sample could have come from some specified population. These tests are in contrast to the two-sample tests, which may be more familiar, which compare two samples and test whether it is likely that the two came from the same population.

The one-sample test is usually of the goodness-of-fit type. In the typical case, we draw a random sample and then test the hypothesis that this sample was drawn from a population with a specified distribution. Thus the one-sample test can answer questions like these: Is there a significant difference in location (central tendency) between the sample and the population? Is there a significant difference between the observed frequencies and the frequencies we would expect on the basis of some principle? Is there a significant difference between observed and expected proportions? Is it reasonable to believe that this sample has been drawn from a population of a specified shape or form (e.g., normal, rectangular)? Is it reasonable to believe that this sample is a random sample from some known population?

In the one-sample case a common parametric technique is to apply a t test to the difference between the observed (sample) mean and the expected (population) mean. The t test, strictly speaking, assumes that the observations or scores in the sample have come from a normally distributed population. The t test also requires that the observations be measured at least in an interval scale.

There are many sorts of data to which the t test may be inapplicable. The experimenter may find that (*a*) the assumptions and requirements of the t test are unrealistic for his data, (*b*) it is preferable to avoid making the assumptions of the t test and thus to gain greater generality for the conclusions, (*c*) the data of his research are inherently in ranks and thus not amenable to analysis by the t test, (*d*) the data may be simply classificatory or enumerative and thus not amenable to analysis by the t test, or (*e*) he is not interested only in differences in location but rather wishes to expose *any* kind of difference whatsoever. In such instances