1972

A Simulation Study of the Power Efficiency of Certain Nonparametric Statistical Tests for Normal Alternatives.

Travis Hillman Willis
Louisiana State University and Agricultural & Mechanical College

A Simulation Study of the Power Efficiency of Certain Nonparametric Statistical Tests for Normal Alternatives

A Dissertation

Submitted to the Graduate Faculty of the
Louisiana State University and
Agricultural and Mechanical College
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

in

The Department of Quantitative Methods

by

Travis Hillman Willis
B.S., Louisiana State University, 1962
M.B.A., Memphis State University, 1968
August, 1972

ACKNOWLEDGEMENTS

My sincere thanks go to the members of my committee: Drs. Vincent E. Cangelosi (chairman), Roger L. Burford, Carolyn N. Hooper, Eugene C. McCann, and G. Randolph Rice for their kind assistance.

I am also indebted to the Louisiana State University Computer Research Center for their understanding in granting permission to utilize considerably more than the usual amount of computer time.

Appreciation must also be extended to the Louisiana State University graduate school for awarding me a dissertation-year fellowship which helped substantially to relieve some of the financial burden.

Of course, my appreciation also goes to my wife, Ilona, whose understanding and moral support made it all possible; not to mention the typing of numerous rough drafts. I am also grateful to my two daughters, Ann and Kris, who contributed more than they will ever know.

TABLE OF CONTENTS

LIST OF TABLES

ABSTRACT

A Monte Carlo simulation technique was used to investigate the power efficiency of three nonparametric two-sample tests. The power of the sign test, the Kolmogorov-Smirnov test, and the Mann-Whitney U test was compared with the power of their t-test equivalent--the paired t-test in the case of the sign test, and the t-test for independent samples for the Kolmogorov-Smirnov test and the Mann-Whitney test.

The simulation process permitted the investigation of a wide range of parameters. Each test was investigated for one-tailed significance levels of .05 and .01; equal samples of size $m = n = 6(2)20, 30, 40, 50$; and location-shift alternatives $\Theta = 0.0(0.2)1.0, 2.0, 3.0$, where $\Theta = \frac{\mu_2 - \mu_1}{\sigma}$. Restrictions on computer time prevented the analysis from encompassing a wider range of parameters.

The analysis was performed on an IBM 360/65 computer with a simulation process based on a Monte Carlo procedure of generating random normal deviates. Random samples of equal size were generated from normal distributions with equal variances of one; the first sample being drawn from a distribution with $\mu = 0$ and the second sample from a distribution with $\mu = \Theta$. Two thousand separate samples were tested for each set of parameters for samples 6 to 20 and 1,000 repetitions for samples 30 to 50. Power was obtained by establishing a decision rule and determining the number of rejections in the total number of test samples.

The findings were divided into two categories--probability of a Type I error ($\Theta = 0.0$) and power efficiency.

The results obtained from simulating the probability of a Type I error indicate that, in general, each nonparametric and parametric test was operating under similar test conditions, and, therefore, valid findings were produced in the study. However, for the Kolmogorov-Smirnov test, which is based upon the establishment of cumulative frequency distributions, it was necessary to increase the number of class intervals in the cumulative distributions to $2(n + m)$ before valid results were obtained.

The power efficiency of the sign test decreased from approximately 80 percent for the smaller parameter values of n and Θ to approximately 60 percent as the parameters increased. Over the same range of parameter values, the relative efficiency of the K-S test increased from approximately 50 to 70-75 percent, and all of the power efficiency values for the U-test fluctuated, primarily, between 90 and 100 percent. A slight increase in power efficiency was noted for both the sign test and the Kolmogorov-Smirnov test as the significance level decreased. Sampling error prevented any patterns from emerging as parameters changed for the U-test.

It was anticipated that the K-S test would outperform the sign test for all parameter values. This proved not to be true for the smaller parameters. The power of the K-S test relies upon the assumption of continuous distributions and if this assumption is violated by creating too few classes then performance suffers. Therefore, the researcher is advised to use at least $2(n + m)$ class intervals in the test procedure.

The power of the U-test was found to be very close to that of the t-test. The U-test is recommended over the t-test in all cases for testing the hypothesis of equal means, except those in which the underlying

distributions can be safely assumed to be normal. The Kolmogorov-Smirnov test is preferred to the sign test when large samples or large location-shift alternatives are encountered. However, when small samples or alternatives are involved the evidence of this study favors the sign test, especially when the ease of computation is considered.

CHAPTER I

INTRODUCTION

Numerous occasions occur within the business complex in which a two-sample statistical test is appropriate for analyzing data. Consider the research and development division of a firm which must determine which of two types of sun tan oil is most effective, or consider a production problem in which two different machine settings are compared to determine if they result in a significant difference in tolerances. The traditional method for analyzing such data has been with the use of the t-test--a test based on underlying normal distributions.

Within the past thirty years a number of two-sample statistical tests have been developed that do not depend upon any stringent assumptions concerning the underlying distributions. These nonparametric, or distribution-free tests as they are sometimes called, seldom assume more than continuously distributed data and independent sampling. Although the terms nonparametric and distribution-free are often used interchangeably, they are not synonymous. As Bradley (1968:15) pointed out, nonparametric tests have no hypothesis about the value of any parameter, whereas distribution-free tests make no assumptions concerning the type of population being sampled. Since it is common to assume an underlying continuous distribution, the term distribution-free is not completely accurate.

When a nonparametric test is being considered for analyzing data, the question arises as to how the nonparametric test compares with the parametric test, assuming that the assumptions of both tests are met. This comparison is usually made on the basis of the relative efficiency of the nonparametric test. Relative efficiency is defined as the ratio of sample sizes that is necessary to equate the powers of the two statistical tests. Since the comparison relies upon respective powers, the more descriptive term, power efficiency, is often used.

THE PROBLEM

There exists a need to provide the researcher with an a priori power efficiency value for the particular test that is being used, given the parameters that apply. The problem can be approached through any of three methods: an asymptotic approach, a deterministic study of finite samples, or an empirical investigation.* The last method is used in this study because it was felt that this procedure provided the greatest flexibility.

Statement of the Problem

It is the purpose of this study to empirically determine the power efficiency of three selected nonparametric statistical tests for various parameter values, using a simulation technique. The three tests, which are discussed below, were selected because of their popularity and wide applicability in business and economic analysis. Since power is a

---

*The terms simulation and empirical are used synonymously in this paper which follows common usage in the literature. Although these terms have different meanings in a strict sense, simulation is empirical but uses artificial rather than actual data.

function of three parameters, these three parameters were assigned various values to establish a spectrum of power efficiencies. The following parameter values were investigated: (1) significance levels of .05 and .01 for one-sided tests; (2) a range of various equal sample sizes from 6 to 50; and (3) a range of mean differences in normal populations of $\Theta = 0.0(0.2)1.0, 2.0, 3.0$,* where $\Theta = \frac{\mu_2 - \mu_1}{\sigma}$ for one-sided tests.

It should be pointed out that the samples were generated from normal distributions with equal variances of one. The reason for choosing an underlying normal distribution, rather than some other distribution, is that power efficiency values are customarily given on the basis of normality. When any parametric test is used the normal distribution is assumed. Thus, a comparison of equivalent tests is more meaningful when the assumptions of both tests are valid.

Nonparametric Tests Considered in the Study

The nonparametric tests that were examined are the sign test, the Kolmogorov-Smirnov two-sample test, and the Mann-Whitney U test. These three tests are among the most popular of the nonparametric tests used in the social sciences, and one fact in support of this popularity is the voluminous literature that exists on these tests.

The sign test is one of the simplest and easiest two-sample tests to apply. When two parent populations are symmetrical and continuous, the sign test can be used to test for a zero difference between population

---

* $\Theta = 0.0(0.2)1.0, 2.0, 3.0$ is read as follows: Θ ranges from 0.0 to 1.0 in increments of 0.2 and then takes values of 2.0 and 3.0.

medians, or population means, since the mean and the median are identical in a symmetrical distribution. If one of the samples receives a particular treatment, then the sign test is appropriate for determining whether the two conditions are significantly different.

In cases in which each pair of samples is related in some manner and is independent of any other pair of sample observations, the sign test is especially appropriate. The example mentioned earlier concerning the testing of two sun tan oils fits this situation. In this example, each subject supposedly coats one arm with one oil and the other arm with the second oil. After a certain amount of exposure to the sun, the oils on each person are rated for tanning effectiveness. Each sample pair is related in that both oils are applied to every person.

The sign test is often used a quick preliminary check to determine if the application of a more sophisticated test is justified.

Although the Kolmogorov-Smirnov two-sample test is more difficult to compute than the sign test, it remains very popular, partly because it is so well tabled. Developed by two Russian mathematicians, this test is sensitive to any kind of difference in the distributions from which the samples are drawn. Significant differences in location (central tendency), dispersion, skewness, etc., influence the Kolmogorov-Smirnov test statistic. The Kolmogorov-Smirnov test is one of a large class of maximum-deviation tests which is based on differences in cumulative distribution functions.

Consider a situation in which a business firm wishes to know if male and female responses to television advertising differ in a particular fashion. More specifically, do men and women differ in the time that they wait to buy a certain product after their initial exposure to the

advertisement? The most appropriate test for this experiment is the Kolmogorov-Smirnov one-tailed test.

Another test that is germane to this type of problem is the final test investigated in this study--the Mann-Whitney U test. The Mann-Whitney U test has the distinction of being one of the more powerful of the nonparametric tests. The U-test is sensitive to differences in populations, but it is different from the Kolmogorov-Smirnov test in that it is especially sensitive to unequal locations. If the experimenter randomly draws two independent samples from the same population and subjects one set of samples to a particular treatment and the other set of samples to another treatment, the Mann-Whitney test could be used to determine if the two treatments are the same. It is common for one sample to receive a treatment and the other sample to serve as a control, i.e., to receive no treatment.

The Mann-Whitney test is also appropriate for testing the hypothesis that two populations differ. For example, assume that a cereal company has produced two dietary cereal products and wishes to know which cereal results in the greatest amount of weight loss in individuals. If the cereals are assigned to individuals in a random manner, then the U-test is almost as effective as the t-test for testing the null hypothesis.

The parametric test that is equivalent to these three nonparametric tests, and thus will provide the comparative base for the power efficiencies, is Student's t-test. The exact configuration of the t-test is discussed in Chapter III.

Relevance and Limitations of the Study

The concept of power efficiency is basic in nonparametric statistics. This is the primary criterion upon which various tests are

compared. If a researcher can determine fairly accurately the power efficiency of his test, even before computing the test statistic, this is of interest from an applied as well as a theoretical standpoint. Such information tells the researcher what sacrifices in power are being made when the sample size and significance level are set and, if these two parameters are flexible, how the relative efficiency can be affected by a change in these parameters.

The efficiencies that were computed in this research were based upon normal shift alternatives. There is certainly no technical reason for not investigating non-normal alternatives. In fact, as will be pointed out in the next chapter, a large number of studies have dealt with this situation, in which such underlying distributions as the uniform, exponential, logistic, and Cauchy have been investigated. Such research is certainly not superfluous; but when one goes beyond normality, comparisons become less meaningful because of the numerous possibilities that exist. Thus, the scope of this study was limited to normal alternatives.

Any study of this type must suffer certain limitations to keep the subject matter manageable. As will be pointed out in the next chapter, previous studies have limited their approach, usually by one of two methods. Many have taken an asymptotic approach, computing the asymptotic relative efficiency (a limiting efficiency function as $n \to \infty$) of various tests. The disadvantage of this method is that these efficiencies provide limited insight for the researcher who works with finite samples.

The other common approach has been to view the problem from a deterministic standpoint and compute the exact powers and power

efficiencies for a few selected finite sample sizes. The inherent difficulty with this is the complex and sometimes intractable power functions that must be dealt with. As a result of having to deal with these intricate functions, the research has often covered only a limited number of alternatives (sample sizes, significance levels, or shifts in location).

This study overcame some of these limitations by including a large combination of alternatives--those that are likely to exist in field experiments. To broaden the spectrum of alternatives, a simulation technique, based upon a Monte Carlo normal deviate generation process, was used. Simulation proved to have an inherent flexibility that could not be approached by deterministic methods.

Perhaps a justified objection to simulation is that it is merely an approximation of the true case. But in order to cover a large number of alternatives, simulation was the most practical approach. The simulation, itself, is set in a stochastic framework, as are the tests being simulated. Therefore, it did not seem inappropriate to use an artificial method of data generation when the analysis itself is a synthetic situation.

ORGANIZATION OF REMAINDER OF THE THESIS

A review of the literature is presented in Chapter II. Because the two-sample statistical test is frequently encountered in all areas of applied research, the tests have come under considerable review and analysis. There exists a fairly extensive collection of research material that is devoted to the study of nonparametric power.

An attempt has been made to cover in depth the literature that discusses relative efficiency and to concentrate particularly on the empirical studies. The literature related to Student's t-test is reviewed first, followed by writings pertaining to the sign test, the Kolmogorov-Smirnov test, and the Mann-Whitney U test. Finally, the purely empirical investigations are summarized.

In the third chapter the structure of the problem and the methodology are discussed. The first part of the chapter is devoted to a brief review of power efficiency, followed by an explanation of the formulation of the three distribution-free tests and their parametric equivalents. Next, the rudiments of the simulation procedure are analyzed. Included in this section is primarily an outline of the method used to generate the necessary data, and secondarily, a discussion of how certain problems were handled.

The results of the study are presented in Chapter IV. The power efficiency data are presented in tabular form and the important outcomes are discussed.

The final chapter, Chapter V, is devoted to a summary of the developments of the previous material and the conclusions drawn from the results.

CHAPTER II

REVIEW OF THE LITERATURE

The considerable literature on two-sample statistical tests reflects the prominence of this test in research. The two-sample test is appropriate for determining the difference between two populations or two population means. The parametric test that is usually applied in this situation is reviewed first--the Student t-test. Following the t-test there is a brief review of asymptotic relative efficiency. An investigation of this important concept is necessary prior to reviewing the literature concerning the three nonparametric tests and their power efficiencies. Finally, the findings and limitations of previous simulation studies are covered.

STUDENT T-TEST

If certain assumptions can be met, the parametric t-test (Student, 1908) is the most powerful test that can be applied in certain practical situations. These specific assumptions and the assumptions of all of the tests that are investigated in this manuscript are enumerated in the following chapter. Since the relative efficiency of a statistical test is based on a comparison of powers, it is the power of the respective tests that is of interest to researchers.

Owen (1965) is just one of many authors that have investigated the power of Student's t-test. As is the procedure in many articles discussing power, Owen evaluated both normal and non-normal conditions.

Although the present study is concerned with normal alternatives, a significant amount of literature deals with the problem of non-normality and other parametric assumption violations. If a test has the ability to withstand violations to its underlying assumptions, it is referred to as being robust. Robustness oftentimes enters the picture of power analysis because research in power efficiency has often been conducted in terms of normal alternatives vis-à-vis non-normal alternatives. A number of studies have shown that the t-test is quite robust to various violations (for example, see Boneau, 1960).

When the assumptions of normality hold, the power of the t-test may be calculated exactly. Two publications have appeared recently which contain extensive power tables of the t-test (see Cohen, 1969, and Milton, 1970). However, as is shown later, the power of most nonparametric tests is not so easily calculated.

A measure of relative efficiency is usually determined with the power values of a nonparametric test and its parametric equivalent. The traditional approach to defining relative efficiency has been in an asymptotic context.

ASYMPTOTIC RELATIVE EFFICIENCY

Asymptotic relative efficiency (A.R.E.) provides an analytical solution to the problem of power efficiency. An asymptotic approach is the only feasible approach that will give a single summary measure of the efficiency of a test. The A.R.E. is the limit of the reciprocal of the ratio of sample sizes required to achieve the same power. As the sample sizes tend to infinity, the alternative hypothesis approaches the null hypothesis to keep the powers of the tests bound away from one. Asymptotic

relative efficiency is credited primarily to Pitman (1948) and was extended by Noether (1955), Hoeffding and Rosenblatt (1955), and Witting (1960).

Pitman's theorem of asymptotic efficiency was succinctly presented in an article by Noether (1958). Let $T_n = T(x_1, \ldots, x_n)$ be a consistent test statistic for testing the hypothesis $\Theta = \Theta_0$ against the alternative $\Theta = \Theta_1$. If $E(T_n) = \Psi_n(\Theta)$ and var $(T_n) = \sigma_n^2(\Theta)$ then the quantity

$$R_n^2(\Theta_0) = \frac{[\Psi_n'(\Theta_0)]^2}{\sigma_n^2(\Theta_0)} \tag{2.1}$$

is called the "efficacy" of T_n. When the alternative hypothesis is stated $\Theta = \Theta_1 = \Theta_0 + \frac{k}{\sqrt{n}}$ when k is an arbitrary, but fixed, positive constant, it is clear that $\Theta_1 \to \Theta_0$ as the sample size n increases. Suppose there are two tests of the same hypothesis with efficacies $R_{1,n}^2(\Theta_0)$ and $R_{2,n}^2(\Theta_0)$. The ratio of these two efficacies in a limiting form gives Pitman's theorem,

$$\frac{n_1}{n_2} = e = \lim_{n \to \infty} \frac{R_{2,n}^2(\Theta_0)}{R_{1,n}^2(\Theta_0)}. \tag{2.2}$$

This is the asymptotic efficiency of the second test relative to the first test. Stuart (1954a) has shown that Pitman's theorem is equivalent to measuring test efficiency by the estimating efficiency of the test statistic. This was supported by Sundrum (1954) and, thus, Pitman's efficiency can be reduced to

$$e = \lim_{n \to \infty} \frac{\sigma_{2n}^2}{\sigma_{1n}^2}, \tag{2.3}$$

which is the ratio of the variances of the two test statistics. Therefore, the A.R.E. of two consistent tests is equal to the ratio of the asymptotic variances of two consistent estimators of Θ on which these tests are based.

Since only very large samples are considered, the A.R.E. represents a theoretical lower limit to the power efficiency function. The limiting conditions under which the A.R.E. is computed do not change from test to test, so the A.R.E. can be considered standardized, and thus, provides a useful index for comparing various tests. The assumptions of the A.R.E. concept make it manageable from a mathematical standpoint.

Pitman was actually preceded by Cochran (1937) when Cochran computed the asymptotic efficiency of the binomial series, or sign test. Cochran's asymptotic value of $2/\pi$ for the sign test was verified by Pitman. Cochran restricted his analysis to the sign test and did not develop limiting functions as Pitman did eleven years later. Following these initial developments, a number of variations to computing asymptotic efficiency have been set forth.

Bahadur (1960a), (1960b), and (1967) presented variations to Pitman's basic concept. In one approach, instead of allowing $\Theta_1 \rightarrow \Theta_0$ as Pitman did, Bahadur held the alternative hypothesis constant and permitted power, β, to converge stochastically to zero while the significance level, α, remained a stochastic uniform variable. Another variation by Bahadur allowed α to converge to zero while β was fixed at 1-p and the alternative hypothesis was fixed. According to Gleser (1964), the Bahadur efficiencies are only approximate measures of asymptotic relative efficiency. Bahadur's measures of asymptotic relative efficiency were summarized and contrasted with Pitman efficiency by Savage (1969).

Another variation to computing A.R.E. was introduced by Blomquist (1950). Blomquist computed what he referred to as an asymptotic

local efficiency by taking the ratio of the respective sample sizes under the assumption that the power functions of the two tests have equal slopes at $\Theta = \Theta_0$. However, for the larger samples this is essentially equivalent to A.R.E.

Blyth (1958) also defined A.R.E. in an unusual manner by abandoning the usual method of establishing a ratio of sample sizes. Blyth's method consisted of incorporating three loss functions into the computational scheme.

Another author, Witting (1960), extended Pitman's efficiency concept to encompass finite sample sizes. The zero-order approximation to Witting's formulation was equal to Pitman's efficiency.

The attempt to generalize from the asymptotic level to the finite level illustrates the shortcomings of A.R.E. The conditions which are responsible for the tractability of A.R.E. (infinite sample sizes and converging alternative hypothesis) also limit its praticality. As Bradley (1968:58-59) put it, ". . . while relative efficiency is realistic but not sufficiently general, A.R.E. is general (at least in the sense of being 'standardized') but not sufficiently realistic."

In an attempt to fill this gap, Hodges and Lehmann (1956) proposed a definition of efficiency for small sample theory that may be used in rough comparison with A.R.E. Let N_a and N_b represent the sample size for test a and test b, respectively. For alternative hypothesis, Θ, and Type I error, α, the Hodges-Lehmann efficiency is expressed as

$$e_{a,b}(\Theta,\alpha) = \frac{N_b^*}{N_a} \quad , \tag{2.4}$$

where N_b^* is the randomized sample size for test b needed to match the power of test a. Rarely will N^* be an integer, so linear interpolation between consecutive integer samples is required to equate powers. Hodges

and Lehmann prefer to hold α and Θ fixed and permit $N \to \infty$, which results in an approximate A.R.E. figure. If the location-shift is allowed to approach zero, the Hodges-Lehmann efficiency approaches Pitman efficiency. It is important to note that the asymptotic relative efficiencies proposed by Pitman (1948), Bahadur (1960b), and Hodges and Lehmann (1956) do not always agree, even locally for $\Theta \to 0$. Both Hodges and Lehmann and Bahadur efficiencies approach Pitman efficiency as $\Theta \to 0$ (see Tsutakawa, 1968).

In addition to those references mentioned previously, A.R.E. is summarized in Basu (1956), Stuart (1954b; 1957), and Mood (1954). As a measure of power efficiency, A.R.E. has its shortcomings, but it does provide boundary values that demonstrate the range of the power efficiency of most nonparametric tests. As each nonparametric test is discussed in the next section, the asymptotic values that have been calculated are mentioned.

NONPARAMETRIC STATISTICAL TESTS

Since the initial development of nonparametric tests in the 1930's, there has been a proliferation of literature in which these statistical methods are discussed and developed. There are very few textbooks on statistics that do not give at least a cursory mention of nonparametric methods. Typical of some of the texts that give an above average treatment to nonparametric techniques are Dixon and Massey (1969), Harshbarger (1971), Hoel (1962), Noether (1971), Roscoe (1969), and Walker and Lev (1953). Two of the most popular textbooks that cover nonparametric tests exclusively are Bradley (1968) and Siegel (1956). Recent publications of this type include Conover (1971) and Gibbons

(1971). More mathematical approaches to the subject of nonparametrics are taken in Hájek (1969), Fraser (1957), and Noether (1967). Two extensive bibliographies have been published by Savage (1953; 1962) on subject matter pertaining to nonparametric statistics. The latter reference contains approximately 2,500 entries, which gives some indication of the mass of literature in existence.

Numerous survey articles and monographs discuss general topics and techniques in nonparametric statistics. A few of the better-known articles that discuss the general concept of the nonparametric tests included in this study are Blum and Fattu (1954), Bradley (1967), Gaito (1959), Moses (1952), Siegel (1957), and Smith (1953).

Because this study concentrates on three specific distribution-free tests, the literature dealing with the power efficiency of these tests is reviewed in detail. As each test is discussed, the historical development of that test is covered first. This is followed by the findings of asymptotic relative efficiency. The remaining literature on power efficiency is then presented in, basically, a chronological format.

Sign Test

The sign test, which is based on the binomial distribution, was developed by Cochran (1937). Cochran computed the relative efficiency of the sign test for $\frac{\mu}{\sigma} = 0$ at which the function had a value of $2/\pi \simeq .637$. The asymptotic relative efficiency of $2/\pi$ for the sign test for normal alternatives has been confirmed by many statisticians since Cochran (for example, see Dixon and Mood, 1946; Pitman, 1948; Mood, 1954; and Hodges and Lehmann, 1956).

Pitman (1948) found the asymptotic efficiency of the two-sample sign test relative to the t-test to be

$$e = 8\sigma^2[\int f^2(x)dx]^2 , \qquad (2.5)$$

where σ^2 is the variance of f(x). For the normal distribution, this function is equal to $2/\pi$. Pitman felt that efficiency would be greater for small samples. Hodges and Lehmann (1956) pointed out that in the case where $f(x) = 0$, $e = 0$, thus the function has no positive lower bound. On the other hand, the function has no upper bound either, since the sign test is applicable to distributions having an infinite variance. Hodges and Lehmann found that if f(x) is unimodal, $e \geq 0.333$, this minimum value being attained for the rectangular distribution. When $n \to \infty$ the asymptotic efficiency function appeared to be independent of α but dependent upon Θ. As was found previously, as $\Theta \to 0$ or as $\mu_2 \to \mu_1$, $e \to 2/\pi$. Hodges and Lehmann also pointed out that as μ_2 departs from μ_1, or $\Theta \to \infty$, e decreases steadily from .637 to .500 for one-tailed tests.

An A.R.E. of $2/\pi$ was also calculated by Mood (1954) for the median test for location; a test whose efficiency values are equivalent to the sign test. Mood calculated asymptotic efficiencies for five two-sample tests for normal alternatives. Other authors have investigated the asymptotic efficiency of the sign test using power functions to make the computations (see Bahadur, 1960c; Blyth, 1958; and David and Perez, 1960). A generalized Pitman efficiency was established for the sign test by Witting (1960).

One of the first studies that provided greater insight into the power of the sign test for finite samples was undertaken by Mac Stewart (1941). Mac Stewart constructed a table for determining the size of the

sample that is required for the sign test to attain a certain power for a given alternative hypothesis for $\alpha \leq .05$.

Five years later, Dixon and Mood (1946) wrote an excellent survey article on the sign test. The authors derived the asymptotic power of the sign test and presented power efficiencies for three selected sample sizes. At a significance level of .10, it was found the power of the sign test for a sample of 18 equaled the power of the t-test with a sample of 12, resulting in a power efficiency of .667. For a sample size of 30, power efficiency ranged from .667 to .700, and for samples of 44, efficiency ranged from .636 to .659.

Walsh wrote two articles concerning the sign test, one in 1946 and the other in 1949. Although both articles reviewed general concepts of the median test (sign test), in the first article Walsh (1946) investigated the power of the sign test relative to the t-test for slippage for the case of normal populations. For one-tailed tests with samples of 4, 5, and 6, relative efficiency was found to be approximately 95 percent. As the sample size was increased, the relative efficiency dropped, but only to approximately 75 percent for samples of size 13. Thus, for small samples, the sign test exhibited fairly high efficiency.

Walsh defined power efficiency in an unique manner. For a given sample size for the sign test, the degrees of freedom for the t-test were varied as was necessary to make the algebraic sum of the areas between the two power functions equal to zero. Subsequent research (see Jeeves and Richards, 1950; and Dixon, 1953) revealed that Walsh's calculating procedure would cause the power efficiency to have an upward bias.

To avoid the upward bias experienced by Walsh, Jeeves and Richards (1950) computed a randomized relative efficiency value for $\alpha = .05$ and .01. Three techniques were utilized to measure efficiency: (1) Walsh's procedure of balancing the area between power curves, (2) minimizing the maximum difference, and (3) equalizing the power functions at certain fixed points. The initial findings disclosed that the three methods did not result in significantly different power efficiencies. For $\alpha = .05$, power efficiency was about .70 for sample sizes of 6 to 20, and slightly higher for $\alpha = .01$. The relative efficiency slowly approached the asymptotic value of .6366 as n was increased.

Dixon (1953) confirmed the biasedness of Walsh's values and determined that the efficiency of .70 for a sample of 6($\alpha = .05$), which was stipulated by Jeeves and Richards (1950), was too low. Dixon believed that it was necessary to determine the power efficiency for every parameter and alternative to get a truly accurate picture of relative power. With this goal in mind, Dixon explored the power efficiency of the sign test on a larger scale than heretofore taken. The power function for the sign test was tabulated for various sample sizes (5, 10, and 20), for normal alternatives, at levels of significance chosen on the basis of the discreteness of the binomial distribution. A linear interpolation method was used to determine fractional degrees of freedom for the t-test--a method which proved satisfactory except for shift alternatives near zero. In general, the results indicated a decreasing power efficiency for an increasing sample size, an increasing significance level, and an increasing shift alternative.

One year later, Dixon (1954) compared the power of the rank sum test, the maximum absolute deviation test, the median test, and the

total number of runs test with each other and the t-test. This comprehensive research is one of the most incisive investigations of power efficiency that has been performed. For a randomized significance level of .025 and samples equal to five, the power efficiency of the median test ranged from .70 to .73 as Θ increased from 0.5 to 4.5. This increase in power efficiency as the alternative, Θ, increased is interesting in that it did not support one of the conclusions drawn in Dixon's previous article. Since three of the tests that Dixon studied are covered in this research, it is interesting to note that the rank sum test, the maximum absolute deviation test, and the median test ranked in that order in efficiency, relative to the t-test.

Milton (1970:39) published extensive tables of power and power comparisons for four nonparametric tests. The tests included the Wilcoxon test (Mann-Whitney U test), the normal scores test, the median test (sign test), and the Kolmogorov-Smirnov test. Power efficiencies were computed in the Hodges-Lehmann form that was outlined previously. The Hodges-Lehmann efficiency of the median test was tabulated for one-sided tests with $n = m = 5$, 6, and 7 for significance levels of .05 and .01 and shift alternatives, $\Theta = 0.2(0.2)1.0$, 1.5, 2.0, 3.0. For a sample of six, $\alpha = .01$, power efficiency of the median test decreased steadily from .6905 to .6322 as Θ increased from 0.2 to 2.0. The power efficiencies for the corresponding alternatives with $\alpha = .05$ were slightly higher; power efficiencies of the other two samples were lower for $\alpha = .01$ than for $\alpha = .05$. In some instances efficiency increased ($n = m = 5$, $\alpha = .01$; $n = m = 6$, $\alpha = .05$) as the shift increased. There were not enough samples to determine any trend in the power efficiency for a given alternative as the sample size increased (for samples 5, 6, and 7, the power efficiency

decreased and then rose for $\alpha = .01$, and just the opposite for $\alpha = .05$). The power efficiency values obtained by Milton were limited to too few sample sizes to draw any definite conclusions. In addition to obtaining a few isolated values, an important point to be attained from the study was the need to broaden the scope of parameter values.

Research with non-normal alternatives has not been ignored. In a doctoral dissertation, one of the authors of the Mann-Whitney U test (Whitney, 1948) investigated the sign test for normal, rectangular, double rectangular, triple rectangular, and Cauchy alternatives. It was found that for many non-normal alternatives the sign test performed well, especially for alternatives for which the variables were concentrated at the mean or median. Gibbons (1964) investigated the performance of the sign test under several combinations of skewness and kurtosis in the underlying distribution.

The evidence in the literature supports the contention that the power of the sign test does not compare very favorably with the t-test. One piece of evidence already presented suggested that the Kolmogorov-Smirnov two-sample test was more powerful than the sign test.

Kolmogorov-Smirnov Test

The historical development of the Kolmogorov-Smirnov test is presented in a thorough and lucid manner by Darling (1957). It appears that the initial development of the Kolmogorov-Smirnov test took place when Kolmogorov (1933) developed a test based on the maximum deviation of two empirical distributions. In 1939, Smirnov made a distribution-free test of Kolmogorov's test, determined the limiting ditribution for the test, and presented a table of critical values. Kolmogorov (1941)

authored a brief article two years later which summarized the work that had been done on his original test up to that time. A similar survey article was authored by Smirnov (1944). Smirnov (1948) republished, in English, the tables that he had originally presented in Russian in 1939. The test that resulted from the combined efforts of Kolmogorov and Smirnov has proven to be a very useful test in the social sciences.

The particular configuration of critical values for the Kolmogorov-Smirnov test (or simply K-S test) can be tabled a number of ways, and it is merely a matter of personal requirements as to which table is most suitable (see Massey, 1950a; Massey, 1951a; Massey, 1951b; Birnbaum, 1952; Massey, 1952a; Goodman, 1954; Miller, 1956; Birnbaum and Hall, 1960; Owen, 1962; and Lilliefors, 1967).

Other articles of an expository nature in addition to Darling's excellent summary of the K-S test include Massey (1951b), Birnbaum (1953), and Goodman (1954). The article by Massey (1951b) is limited to the Kolmogorov-Smirnov goodness-of-fit test which is the one-sample version of the two-sample test that is covered in this manuscript. However, his calculation of a lower bound for the K-S test can be applied to the two-sample case.

In two other articles, as well, Massey (1950b; 1952b) computed the lower bound for the K-S test and demonstrated that the test was consistent against all alternatives $F(x) \neq G(y)$, assuming the smaller of the two sample sizes approaches infinity while the ratio of sample sizes remains away from zero and infinity. It was also shown that the K-S test is biased for finite sample sizes. Massey presented the lower bound for the K-S test as

$$1 - \frac{1}{\sqrt{2}} \int_{2[\Delta\sqrt{N} - d_\alpha(N)]}^{2[\Delta\sqrt{N} + d_\alpha(N)]} e^{-t^2/2} dt. \qquad (2.6)$$

The rationale for computing a lower bound for the K-S test instead of the A.R.E. is that the K-S statistic does not have the characteristics necessary for computing a conventional A.R.E.

In an excellent article, Capon (1965) pointed out that, in general, the power of the K-S test cannot be computed because the limiting distribution under the alternative hypothesis is not known; and, because the usual assumptions concerning asymptotic normality are not satisfied, asymptotic relative efficiency cannot be computed. However, a lower bound for the power of the test can be calculated and, following Massey (1950b), Capon derived a lower bound for the asymptotic efficiency of the K-S test relative to the optimum likelihood ratio test. Capon made essentially the same assumptions as Massey--as m = n approaches infinity, the ratio $\frac{n}{m}$ is bound away from zero and infinity. Applications were made for the Cauchy, exponential, and normal distributions. When sampling took place from two normal populations that differed only in location, the lower bound of the K-S test relative to the optimum likelihood ratio test was $\frac{2}{\pi} \approx .637$, and the upper bound was 1.0. Bradley (1968) felt that the true A.R.E. was somewhere between these two values. The lower and upper bound were also computed for the K-S test relative to Student's t-test. It was found that when two unspecified populations of the same type differ only in location, the lower bound was greater than or equal to $\frac{1}{3}$ and the upper bound was capable of being large for certain populations.

Further study of the asymptotic efficiency of the K-S test was carried out by Klotz (1967). Asymptotic efficiency was derived and evaluated for normal location and normal scale alternatives. Using equal sample sizes, the limiting efficiency was obtained by letting $\alpha \to 0$ and

fixing β, $0 < \beta < 1$. The limit of the relative efficiency was $\frac{2}{\pi}$ for normal location shift alternatives that approached the null hypothesis.

Studies of the power of the K-S test have been somewhat limited as compared to the other nonparametric tests covered in this paper. Darling (1957) found that information concerning the power efficiency of the K-S test was quite fragmentary. This is probably due mainly to the difficulties encountered with A.R.E. and the complexity of the power function. The first significant power comparisons were made by van der Waerden.

Van der Waerden (1953b) investigated the power of the K-S test and the Mann-Whitney U test for a number of sample sizes under the assumption of normality and equal variances--situations in which the t-test is the most powerful test. For all the cases investigated, the K-S test proved less powerful than the Mann-Whitney U test. The relative efficiency of the K-S two-sample test with one sample being large and the other equal to five was 65 percent for both one-tailed and two-tailed tests. When the smaller sample exceeded five, van der Waerden expected efficiency to fall. In a continuation of the same article, van der Waerden (1953b) investigated non-normal distributions and unequal variances. In another article, van der Waerden (1953a) suggested that the K-S test demonstrated inferiority to the classical test in detecting mean differences because of the universal nature of the K-S test as compared to the single purpose of the classical test to detect a difference in means.

Dixon (1954) investigated the power of the maximum absolute deviation test (K-S test) in the same study that was mentioned in the sign test review. Power comparisons were made by numerically integrating

power functions in a deterministic framework. The power efficiency of equal samples of size 3, 4, and 5 drawn from normal distributions with equal variances for various $\Theta = \frac{|\mu_1 - \mu_2|}{\sigma}$ were studied. Unfortunately, computational complexity restricted the number of different samples and levels of significance that were included in the study. In order to equate powers, fractional sample sizes of the t-test were found by polynomial interpolation. Hodges and Lehmann (1956) attacked this procedure as lacking "functional meaning."

The level of significance was randomized to a value of .025 for equal samples of five to make comparisons among the nonparametric tests. For alternatives from 0.5 to 4.5, the power efficiency of the K-S test relative to the t-test decreased from .81 to .74; each value was lower than the Mann-Whitney U test, but higher than the sign test. However, the advantage over the sign test was very small for large alternatives ($\Theta > 3.0$). In general, the power efficiency decreased slightly as the shift alternative increased, and as the level of significance increased.

The evidence that Dixon presented does not support the contention of Siegel (1956:136) that ". . . whereas for very small samples the Kolmogorov-Smirnov test is slightly more efficient than the Mann-Whitney U test, for large samples the converse holds." Dixon's study supports the conclusion that the Mann-Whitney U test is more powerful than the K-S test for every parameter.

Lee (1966) compared the exact power of the K-S test with a standard parametric test--the normal test. The evaluation included samples of size five considered drawn from normal distributions differing in means. For $\alpha = .05$ and .01, the relative efficiency increased from .84 to .98 and from .76 to .92, respectively, as the shift

alternative increased from 0.5 to 2.0. It should be noted that the efficiency increased as the location-shift increased which is atypical in light of a majority of the findings.

Recently, Knott (1970) and Milton (1970) have investigated the power and relative efficiency of the K-S test. Knott computed efficiencies of the K-S test relative to the optimum normal test and found that performance did not deteriorate substantially as the sample size increased. General efficiencies of 75 percent for $\alpha = .05$ and 72 percent for $\alpha = .01$ were found. In addition, Knott obtained the lower bound for the K-S test, $2/\pi$.

Milton (1970) presented tables of the exact power of four nonparametric tests for both one-sided and two-sided tests for all sample sizes $2 \leq n \leq m \leq 7$. Various levels of significance were investigated for $\theta = 0.2(0.2)1.0, 1.5, 2.0, 3.0$. As mentioned in the review of the sign test, Hodges-Lehmann efficiencies were computed for the one-sided K-S test relative to the t-test. One result taken from Milton (1970:40) had power efficiency falling steadily from .8632 to .8583 for increasing location-shifts with $n = m = 6$ and $\alpha = .01$. The corresponding power efficiencies were generally lower for $\alpha = .05$ although noted exceptions existed for the larger location-shifts. Power efficienty decreased fairly consistently as the location-shift alternative increased. As with the sign test, not enough sample sizes were included in the report to determine any definite trend in power efficiency as sample size increased.

Although the evidence is not complete, it appears that the K-S test is more efficient than the sign test. The literature shows that the Mann-Whitney U test is the most powerful of the three tests.

Mann-Whitney U Test

The Mann-Whitney U test is a linear transformation of the Wilcoxon rank-sum two-sample test. Therefore, all of the information that is pertinent to the power of the Wilcoxon test also applies to the Mann-Whitney U test.

Wilcoxon (1945) developed a test that is based on the sum of the rankings of the observations. The Wilcoxon test was generalized and extended by Mann and Whitney (1947), who considered both unequal and equal samples. A table of critical values was established for samples up to $m = n = 8$; for larger samples, Mann and Whitney felt that the normal approximation was appropriate.

The first individual to investigate the asymptotic relative efficiency of the Wilcoxon or the Mann-Whitney U test was Pitman (1948). Pitman's efficiency of the U-test is given as

$$e = 12\,\sigma^2[\int f^2(x)dx]^2. \tag{2.7}$$

For normal populations this is equal to $\frac{3}{\pi} \simeq .955$. Several writers have verified this result (for example, see van der Vaart, 1950; van der Waerden, 1952 and 1953b; and Mood, 1954). Hodges and Lehmann (1956) found that the A.R.E. of the Mann-Whitney U test never falls below .864 for any underlying continuous distribution. They also discovered that for certain non-normal distributions the relative efficiency of the U-test could be arbitrarily large. Thus, Hodges and Lehmann correctly concluded that using the U-test instead of the t-test could never entail a serious loss of efficiency.

Witting (1960) developed a generalized Pitman efficiency for the Mann-Whitney U test which was equal to Pitman's efficiency of $\frac{3}{\pi}$ for the

zero-order approximation in the case of normal alternatives and equal to 1.0 for the uniform distribution. A comparison of Bahadur efficiency and Pitman efficiency for the Mann-Whitney test was made by Hollander (1967).

Tables of critical values for the Mann-Whitney U test have appeared in Auble (1953), Jacobson (1963), Milton (1964), Mc Cornack (1965), and Claypool (1970); Jacobson (1963) also includes a thorough bibliography.

One of the first analytical investigations of the power of the Mann-Whitney U test was undertaken by Whitney (1948). The U-test was compared with the normal test and the t-test under three separate conditions: $\sigma_x^2 = \sigma_y^2$, $\sigma_x^2 = \frac{\sigma_y^2}{4}$, and $\sigma_x^2 = 4\sigma_y^2$. It was found that under certain non-normal conditions, the Mann-Whitney test was superior to both parametric tests and very close in power under normal conditions.

Perhaps the first person to study the small sample power of the U-test against normal alternatives was van der Vaart (1950). Comparisons of power against the t-test were made by evaluating the ratio of the derivatives of the power function at the null hypothesis for one-tailed tests with $m + n \leq 5$ and for two-tailed tests with $m + n \leq 6$. The power of the U-test compared very favorably with the power of the t-test for small samples at selected significance levels. Indications were that, even for large samples, the difference in power was not too great. The ratio of the second derivatives of the power functions yielded the asymptotic efficiency of $\frac{3}{\pi}$. In a later article, van der Vaart (1953) investigated the power function of the Wilcoxon two-sample test when the underlying distributions were not normal.

A slightly different approach was used by van der Waerden (1952) who computed the actual power of the Wilcoxon and the t-test for particular alternatives. For $m = n = 2$, a mean difference of two, and a standard deviation of one, the power of the Wilcoxon test was approximately 62 percent while the power of the t-test was a little higher, 65 percent. In another article by van der Waerden (1953b) the asymptotic efficiency of the Wilcoxon test was verified, the power of the U-test was compared to the K-S test, and non-normality was investigated.

All of the studies that have been mentioned up to this point suffer a common malady--comparisons of power have been made on the basis of an extremely limited number of alternatives. Dixon (1954) emphasized that a comprehensive efficiency comparison must be based on an evaluation of all possible values of n, α, and Θ. Obviously, this is not possible, but Dixon did extend his analysis to cover more parameter values than previous studies. As mentioned in reviewing the two previous tests, Dixon used a numerical procedure to evaluate the power of the nonparametric tests. Power efficiency was computed for equal samples of five for $\alpha = .025$ for the rank-sum test (Mann-Whitney test, Wilcoxon test) as with the median and the maximum absolute deviation tests. The power of the rank-sum test proved to be superior to all of the other nonparametric tests evaluated. Power efficiency fell steadily from .964 to .88 as the mean difference increased from 0 to 4.5. It was found that, as the level of significance increased, the power efficiency of the U-test increased slightly which is just the opposite to what happened with the sign test and the K-S test. The local power efficiencies for the U-test were very high; for all cases, they were greater than the asymptotic efficiency of $\frac{3}{\pi}$.

Hodges and Lehmann (1956) compared some of their efficiency values with those obtained by Dixon and found that, while Dixon's power efficiency values decreased steadily as the alternative increased, their values increased as Θ increased beyond 3.0. Hodges and Lehmann attributed this to the different methods used in interpolating the parametric sample size. Another result obtained by Dixon that has not been substantiated by other research had to do with an increasing power efficiency associated with an increasing significance level. In this situation, Bradley (1968) among others, felt that efficiency should decrease, not increase.

As with the median test and the Kolmogorov-Smirnov test, Milton (1970:37) investigated the small sample power of the Wilcoxon two-sample rank-sum test for the same alternatives. Extensive power tables were computed for all possible combinations of m,n from 2,1 to 7,7 for various shift alternatives. Power efficiency values for the one-tailed test were given for a range of location-shifts for samples $m = n = 5$, 6, and 7, and for $\alpha = .01$ and .05. For samples $m = n = 6$ and $\alpha = .01$, the power efficiency of the Wilcoxon test decreased steadily from .9667 to .9443 as the alternative ranged from 0.2 to 3.0. The power efficiency values were generally lower for corresponding alternatives at $\alpha = .05$ (exceptions were noted for the higher location-shifts, smaller samples). All of the power efficiencies tended to decrease steadily as Θ increased (a few exceptions were noted for the larger values of Θ). Unequal sample sizes of $m = 7$ and $n = 6$ were also tabulated and the results of the parameters were not significantly different from those for equal samples. Again, as with the other two tests, it should be recognized that the samples that were presented were too limited to draw any definite conclusions.

Another study that used a numerical approach was by Tsao (1957) who computed power values for the Mann-Whitney U test for m = n = 2 and 3 and Θ = .25(.25)1.5. These small samples were evaluated by means of polynomial interpolation and asymptotic efficiency was investigated by letting $\Theta \rightarrow 0$. The Wilcoxon test was compared with the normal scores test for normal alternatives in Hodges and Lehmann (1961); and Witting (1960) investigated the efficiency of the Wilcoxon test for finite sample sizes in the case of normal and rectangular alternatives. For m = n = 5, efficiency equaled .9563. Other numerical investigations were undertaken by Lehmann (1953), Barton (1957), and Gibbons (1963).

One of the many studies that have examined non-normal alternatives for the Wilcoxon test was undertaken by Wetherill (1960), who considered the situation in which the two underlying distributions differed slightly in shape so the assumptions of neither test were met. The study concentrated on normal populations that had unequal variances. Wetherill concluded that Wilcoxon's test was a little more robust to differences in population variance than the t-test, but the Wilcoxon test was much more sensitive to skewness and kurtosis. In cases in which the underlying populations were identical, but non-normal, the Wilcoxon test was preferred over the t-test.

This evidence supports the theory that the Mann-Whitney U test is not just a test of location. It is sensitive to the rapidity of build-up from a specified direction. Thus, an extremely skewed population may result in a significant U even though the two populations may have equal locations.

The literature stresses two points. For normal alternatives, the power of the Mann-Whitney U test is very close to the power of the

t-test. For most non-normal alternatives, the power of the U-test exceeds the power of the t-test. Considering all alternatives, the Mann-Whitney U test is one of the more powerful nonparametric tests.

All of the literature discussed thus far has approached the problem of power efficiency from a deterministic standpoint. An alternative approach (the approach taken in this paper) utilizes the simulation technique.

SIMULATION STUDIES

One of the first simulation or empirical studies of nonparametric power was conducted by Dixon and Teichroew (1954). Only small samples were involved but the sampling was extensive enough to be able to rank the nonparametric tests according to power in the following order, starting with the most powerful test: (1) rank-sum test (U-test), (2) maximum deviation test (K-S test), (3) median test (sign test), and (4) run test. Although the complete results were not available, samples of size $m = n = 5$, 10, 20; $m = 5$, $n = 10$; $m = 10$, $n = 20$ for significance levels .01, .05, and .10 were examined for normal shift alternatives. Power estimates of the rank-sum test which were based on either 100 or 150 pairs of samples, were very close to the t-test. This study closely paralleled Dixon's other paper (Dixon, 1954).

Teichroew (1955) used a similar technique a year later to obtain power values for another particular ranking test. Even though the empirical process was based on 1,000 to 7,000 random samples, the sample sizes never exceeded four.

Another comparison of nonparametric tests against normal shift alternatives was conducted for the purpose of applying the information to life testing (see Epstein, 1955). Two of the nonparametric tests that are of interest were the rank-sum test (U-test) and the maximum deviation test (K-S test). Equal samples of ten were drawn from normal populations which had a common variance of one and differed in location. With $\alpha = .05$, 200 pairs of samples were generated to apply the tests. The results indicated that the Kolmogorov-Smirnov test was not as powerful as the Mann-Whitney U test.

Hemelrijk (1961) compared the power of Wilcoxon's two-sample test with Student's t-test for normal alternatives. The power of one-tailed tests was estimated with $m = n = 10$ and $\alpha = .025$. Because of discreteness, the true level of significance for the Wilcoxon test was .022, but the results indicated that the difference in significance levels had essentially no effect. Hemelrijk generated 250 pairs of samples for various normal alternatives and found that the t-test was superior to the Wilcoxon test for all mean differences. Results from non-normal alternatives indicated the opposite superiority relationship.

A study similar to Hemelrijk's was conducted in the following year by Boneau (1962). Normal, rectangular, and exponential alternatives were simulated for various values of α(.05 and .01), sample sizes (5 and 15), and variances (1 and 4) to compare the Mann-Whitney U test with the t-test. One thousand U's and t's were generated for each condition. The findings, which were presented graphically, revealed that the U-test might be biased and that it was certainly not distribution-free. The U-test was affected by skewness and heterogeneous variances but appeared relatively robust to these non-normal conditions.

Two other statisticians (van der Laan and Oosterhoff, 1955) used a Monte Carlo technique to determine the power functions of the Wilcoxon, van der Waerden, and Terry tests and to compare these tests with each other and the t-test. Although sample sizes $m = n = 6, 8, 10$; $m = 8$, $n = 12$; and $m = 5$, $n = 15$ were studied for various significance levels, only the results for $m = n = 6$ were given. The power of all three tests increased as the significance level increased, and as expected, the power of the Wilcoxon test was very close to the power of the t-test.

Neave and Granger (1968) conducted a simulation study involving eight tests for differences in mean. Three of the tests included the t-test, the Mann-Whitney U test, and the Kolmogorov-Smirnov text. Various combinations of sample size (20 and 40), significance level, variance, and parent distribution were simulated, each involving 500 pairs of samples. As expected, the t-test was inimitable for normal alternatives, followed by the U-test and then the K-S test. Neave and Granger noted that the K-S test was designed to detect more general differences between distributions than the t-test or the U-test and therefore did not perform as well as these tests for detecting shifts in location. The U-test was superior for non-normal distributions.

An empirical comparison of the permutation t-test, the Student t-test, and the Mann-Whitney U test was the subject of a doctoral thesis (see Toothaker, 1969). Location-shift alternatives were studied for normal, uniform, and skewed distributions. The shift or effect size (Θ) was chosen so that the power of the t-test would be .30, .60, and .90 for normal alternatives. One thousand samples were generated from the three types of populations for all sample combinations from 2,3 to 5,5. The experiment was limited to these small samples to avoid using an inordinate

amount of computer time. The t-test demonstrated consistent superiority for the location-shift alternatives, and only for the skewed distribution did the U-test exhibit superior power. Toothaker indicated that 1,000 samples were not sufficient to eliminate the sampling error that occurred in his results.

It is significant to note that even the empirical studies have been somewhat limited in their coverage of power analysis. The size of the samples that were included in the studies were often very small, and even when the samples were larger, only one or two different sample sizes were usually investigated. Certainly, the investigations have not been extensive enough to draw any specific conclusions in the realm of power efficiency. Seemingly, the potential of simulation to expand the analysis to a larger number of parameter values has not been fully explored. The rudiments of this simulation study are disclosed in the following chapter.

In general, the evidence concerning nonparametric tests suggests relatively high power efficiencies associated with small samples, which fall ultimately to the asymptotic relative efficiency value as n increases. There also appears to be some support for the general contention that power efficiency decreases as either the significance level, the mean difference, or the sample size increases. However, the numerous findings of conflicting evidence, even from the deterministic studies, certainly accentuates the need for further research to clear the issue.

CHAPTER III

METHODOLOGY AND STRUCTURE OF THE PROBLEM

Comparisons of two-sample statistical tests in applied research are usually made on the basis of power, in the guise of a power efficiency value. This chapter begins with a brief look at the basic concept of power efficiency. Next, the assumptions and the particular formulation of the statistical tests that were covered in this research are presented. The final portion of the chapter is devoted to an explanation of the specific simulation technique that was used to develop the power efficiencies.

POWER EFFICIENCY CONCEPT

The efficiency of a statistical test is determined by its power; i.e., its ability to avoid accepting a false hypothesis. In other words, the power of a test is the probability that the test will reject a false hypothesis. This ability to reject a false hypothesis is related to a Type II error, β, (the probability of accepting a false hypothesis) in the following manner,

$$\text{Power} = 1 - \beta.$$

When the null hypothesis is, in fact, true, the probability of a Type II error is zero. In this case, the probability of rejecting a true hypothesis is given by the significance level, α. So the power concept has meaning only when the null hypothesis is false.

A very useful method for comparing tests is on the basis of power, and one of the most useful measures for comparing power is relative efficiency. Suppose a researcher has a particular experiment that has assumptions that are met by two different statistical tests. One test, a nonparametric test, requires a sample size of n_2 to have the same power as the other test, its parametric equivalent, which has a sample size of n_1. Then

$$\text{power efficiency} = \frac{n_1}{n_2} .$$

It is customary to put the sample size of the parametric test in the numerator. This sample is usually the smaller sample because the parametric test usually has greater power. If powers are equated when $n_2 = 20$ and $n_1 = 15$, the power efficiency of the nonparametric test is 75 percent. The nonparametric test requires a sample 33.3 percent larger than the sample of the parametric test for the two tests to have equal power. For normal shift alternatives the power efficiency of a nonparametric test should lie between 0.0 and 1.0 where a value of 1.0 signifies equal efficiency or power for a given set of parameters.

FORMULATION OF TESTS

Student's t-test is used to test the hypothesis of equality between two population means when the populations are normal and have equal variances. The t-statistic is calculated for two independent samples as

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\Sigma(X_i - \bar{X})^2 + \Sigma(Y_i - \bar{Y})^2}{m + n - 2} \cdot \left(\frac{1}{m} + \frac{1}{n}\right)}} \qquad (3.1)$$

where $\bar{X}$ is the mean of the values, X_i, from a sample of size m, drawn from the X population and $\bar{Y}$ is the mean of values, Y_i, from a sample of size n, drawn from the Y population. A t-value computed by (3.1) that is greater than or equal to the tabled critical value with m + n - 2 degrees of freedom is significant at the stated significance level. A significant result means that the null hypothesis can be rejected with the probability of α that an error has been made.

For independent samples, the formula,

$$t = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sum_{i=1}^{m} X_i^2 + \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{m} X_i\right)^2}{m} - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n}}{m + n - 2} \cdot \left(\frac{m + n}{mn}\right)}} \qquad (3.2)$$

proved to be more efficient, computationally, than formula (3.1) for computing the two-sample t-test equivalent of the Kolmogorov-Smirnov test and the Mann-Whitney U test.

A different computational formula was used to compute the parametric equivalent of the sign test. The sign test is used primarily in situations in which each sample pair is related in some manner. When there is some sort of relationship between sample pairs, the paired t-test is most appropriate. The paired t-test with n - 1 degrees of freedom is calculated as

$$t = \frac{\Sigma d}{\sqrt{\frac{n\Sigma d^2 - (\Sigma d)^2}{n - 1}}}, \qquad (3.3)$$

where d is equal to the difference between each sample pair $(Y_i - X_i)$ and n is the number of sample pairs.

When the assumptions of normality and equal variances hold, the t-test is the most powerful test. The assumptions for the t-test are:

(1) the observations are independent, (2) the samples are drawn from normal populations, (3) the variances of the populations are equal, and (4) the data are measurable on at least an interval scale.* Only the first assumption is shared by nonparametric tests.

One of the simplest nonparametric tests to apply, the sign test, is based on the binomial distribution. The null hypothesis can be stated as

$$P(X_i > Y_i) = P(X_i < Y_i) = 1/2,$$

or equivalently, that the median difference between two populations is zero, or that the number of pluses and minuses resulting from population differences are the same. The binomial distribution which is stated as

$$\binom{n}{x} p^x q^{n-x}, \tag{3.4}$$

requires values for two parameters, n and p, to determine the probability of x successes in n trials. For the sign test, p = 1/2, n is the total number of pairs of samples showing a directional difference, and x is the smaller number of plus or minus signs taken from each difference $Y_i - X_i$. Given these values, the sign test can be calculated with

$$\left(\frac{1}{2}\right)^n \sum_{j=0}^{x} \frac{n!}{j!(n-j)!} . \tag{3.5}$$

This equation is the cumulative binomial distribution and the probability for the one-tailed test can be read directly from the cumulative binomial table for p = 1/2. For one-tailed tests the direction of the alternative hypothesis is declared in advance, which means that the alternative

---

*There is a difference of opinion among statisticians as to the validity of the last assumption. The present concensus seems to be that an interval scale measurement is not required to satisfy the applicability of a parametric test (see Anderson, 1961; Gaito, 1959 and 1960; Savage, 1957; and Stevens, 1946 and 1968).

hypothesis indicates whether the number of fewer signs will be pluses or minuses. If the probability in the cumulative binomial table is less than or equal to the chosen significance level, the null hypothesis may be rejected.

For large samples ($n > 25$), the normal approximation,

$$Z = \frac{(X + .5) - \frac{n}{2}}{\frac{1}{2}\sqrt{n}} \tag{3.6}$$

may be used. The probability of a value as small or smaller than Z is found in the normal probability table.

The sign test assumes that each pair of observations is independent and that the variable under consideration has a continuous distribution.

The Kolmogorov-Smirnov test is sensitive to differences in location, dispersion, skewness, and kurtosis. The one-sided K-S test is given by

$$D = \text{maximum } [S_1(X) - S_2(Y)], \tag{3.7}$$

where $S_1(X)$ and $S_2(Y)$ are observed cumulative step functions. Only equal samples were considered in this study, therefore $m = n$. Let $S_1(X) = k_1/m$ where k_1 = number of scores less than or equal to X and $S_2(Y) = k_2/n$ where k_2 = number of scores less than or equal to Y. To compute these values, it is first necessary to rank into two separate groups the values sampled from each distribution. Then class intervals must be constructed to make a cumulative frequency distribution for each sample of observations, using the same intervals for both distributions. The best use of information is made if there is a large number of intervals, so $2(n + m)$ intervals were established for every case in the study. After each value is placed into its proper class interval, the differences in the frequency counts for each class are noted. The maximum difference is designated D.

Probabilities for D can be found in most of the tables mentioned in the previous chapter. However, for the purposes of this study a table of critical values was adapted from Birnbaum and Hall (1960) and Massey (1951a). It turned out that all of the values were not used because the number of samples was reduced to conserve computer time. However, this table is presented in Appendix A because it appears to be the only source of D values for the two-sample test that includes every sample size from 5 to 40. For samples larger than 40, the approximate values suggested in Conover (1971:399) were used.

The K-S test has the same basic assumptions as most nonparametric tests, i.e., independent observations and continuous distributions. When the assumption of continuity is violated, the K-S test loses much of its power. The result is a test that is much more conservative than it would be, otherwise. The occurrence of numerous ties in the data is an indication of lack of continuity.

The Mann-Whitney U test is a ranking test that is used to test the hypothesis that two populations are identical, particularly in terms of respective locations. Specifically, the one-tailed hypothesis can be stated:

H_0: $P(X > Y) = P(X < Y) = 1/2$

H_1: $P(X > Y) < 1/2$ or $P(X > Y) > 1/2$.

The Mann-Whitney test is calculated with the test statistic, U, which for given samples m and n, is based upon the number of times a Y value exceeds an X. Thus

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} d_{ij} \tag{3.9}$$

where $d_{ij} = 1$ if $X_i < Y_j$

$= 0$ otherwise.

Calculating U using the procedure required by (3.9) can be very tedious when the samples are large, so Mann and Whitney (1947) developed a formula for calculating U that avoids this cumbersome counting.

$$U = mn + \frac{m(m + 1)}{2} - \Sigma R_m \tag{3.10}$$

or equivalently,

$$U' = mn + \frac{n(n + 1)}{2} - \Sigma R_n \tag{3.11}$$

where U and U' are related in the manner

$$U = mn - U'. \tag{3.12}$$

The values ΣR_m and ΣR_n are the sums of the ranks of m observations and n observations in the X and Y sample, respectively. These rank-sums, which represent Wilcoxon's statistic, are obtained after the scores from both groups are ranked together in ascending order. The smaller of U or U' is the value of interest because this is the value that is tabled. The null hypothesis is rejected if the U (or U') computed from (3.10) or (3.11) is less than or equal to the tabled value.

For the one-sided test in which the direction of the alternative hypothesis (H_1: $\mu_1 < \mu_2$) is predicted, Harshbarger (1971) presented a formula for computing the U-statistic. In this case, the computation of U' to determine whether U or U' is smaller is superfluous because U must be smaller to reject the null hypothesis. This formula,

$$U = \Sigma R_m - \frac{m(m + 1)}{2}, \tag{3.13}$$

was used in this investigation to compute the U-statistic. In this equation ΣR_m and m are values related to the X population for which $\mu = 0$. It ties occurred in the Monte Carlo simulation process (an unlikely occurrence as is explained later) each tied value was assigned the value of the average rank.

Mann and Whitney felt that for samples larger than m = n = 8 the normal approximation could be used with safety. This statement probably reflects the laborious task of constructing critical values for samples larger than eight rather than theoretical accuracy. In any case, the normal approximation was used for sample sizes m = n = 14 and larger because of the unavailability of the necessary critical values. Gibbons (1971:145) reported that the normal approximation has been found reasonably accurate for equal samples of size six. Mann and Whitney determined that

$$E(U) = \frac{mn}{2} \tag{3.14}$$

$$\text{and} \quad VAR(U) = \frac{mn\,(m + n + 1)}{12} \quad . \tag{3.15}$$

Using these equations, the normal approximation for the one-tailed U-test is given as

$$Z = \frac{U - \frac{mn}{2}}{\sqrt{\frac{(m)(n)(m + n + 1)}{12}}} \tag{3.16}$$

Substituting $U = \Sigma R_m - \frac{m(m + 1)}{2}$ in (3.16) and simplifying gives

$$Z = \frac{2\Sigma R_m - m(N + 1)}{\sqrt{\frac{mn(N + 1)}{3}}} \quad , \tag{3.17}$$

where N = m + n. Equation (3.17) proved to be more efficient than (3.16).

The validity of the Mann-Whitney U test relies upon the same assumptions as the Kolmogorov-Smirnov test--independence of observations and an underlying continuous distribution.

The formulas that were stated as being used in this research were programmed in FORTRAN IV language to carry out their particular statistical analysis on each set of data.

SIMULATION PROCEDURE

As indicated in the previous chapter, much of the analysis of power efficiency has been severely limited to a very few parameter values, due mainly to the complexities involved with manipulating power functions. The technique of simulation has yet to be fully utilized as an effective tool in revealing the comparative powers of statistical tests for a broad range of parameters. Therefore, the simulation method was adopted for this study and an outline of the exact procedure follows.

The writer has written three separate computer programs to simulate the performance of each nonparametric test and its corresponding t-test. It was impractical to include all three tests in one program because of the length of each program and the compiling time involved.

A Monte Carlo process was used to generate the normal variables. Equal samples of size $m = n$ were generated from two unit normal distributions with a mean of $\mu = 0$ for one sampling distribution and $\mu = \Theta$ for the second sampling distribution with equal variances of one. These underlying distributions satisfy the assumptions of the t-test, making the t-test the uniformly most powerful test. An underlying normal distribution also satisfies the assumption of continuity which is required by the nonparametric tests.

The computer programs were run on an IBM System 360/65 and IBM library subroutines were used to generate the normal deviates. Subroutine GAUSS (see IBM Scientific Subroutine Package) was used to compute a normally distributed random variable with a given mean and a standard deviation of one. The subroutine uses a sequence of uniform random numbers to approximate a normally distributed deviate, Y, using

$$Y = \frac{\sum_{i=1}^{K} X_i - \frac{K}{2}}{\sqrt{K/12}}, \qquad (3.18)$$

where X_i is a uniform random number, $0 < X_i < 1$. K is the number of values of X_i to be used. As K approaches infinity, Y approaches a normal distribution. For simplicity, K was given a value of 12, thus reducing (3.18) to

$$Y = \sum_{i=1}^{12} X_i - 6.0 \; . \qquad (3.19)$$

Finally, Y was adjusted for the desired mean and standard deviation with

$$Y' = Y\sigma + \mu \; ,$$

where Y' is the normal deviate with mean, μ, and standard deviation, σ.

Another subroutine in the IBM Scientific Subroutine Package was used to generate the uniform random numbers required in GAUSS. The random number generator, RANDU, generates a maximum of 2^{29} or 536,870,912 random numbers, each in the interval, zero to one, before repeating, which was deemed adequate for this study.

If ties occurred in the data, this was an indication that the assumption of continuous distributions was being violated, which diminished the validity of all of the tests that were under investigation. However, because the random number generator produced a normal variable that had a substantial number of significant digits, the chance of a tied observation was extremely remote. Remedial procedures have been developed for most nonparametric tests to offset the effect of tied scores. If a tie happens to occur in the data generated for the sign test, the sample size is simply reduced accordingly and that observed pair is ignored. As indicated before, tied observations that occurred in the data of the Mann-Whitney U test were assigned the value of the average rank. However,

if there is a significant number of ties in the data of the U-test, then a correction factor should be applied. As far as the Kolmogorov-Smirnov test is concerned, ties simply reduce the power of the test. Since the probability of a tied observation was so small, the possibility of this happening in the simulation was ignored and no corrective procedures were installed. Certainly, tied observations would be so infrequent, if they occurred at all, that the effect on the power estimates would be negligible.

Power is defined as the probability of rejecting a hypothesis which is known to be false. Therefore, for each positive mean difference, $\theta > 0$, power was determined by the percentage of rejections over the total number of tests performed. Statistical theory demonstrates that the power of the tests under study increases as the location-shift increases, as the sample size increases, or as the significance level increases. Thus, for each given significance level and location-shift alternative, the sample size of a test can be changed in order to increase or decrease the power of the test.

The process of increasing or decreasing the sample size to manipulate statistical power was used in this empirical study. For a given significance level and shift alternative for the one-tailed test, an estimate of the power of the nonparametric test was obtained for a given sample by calculating the proportion of rejections for the stipulated number of samples. As each sample was drawn and tested by the nonparametric test, the same data were also tested by the parametric test equivalent--the t-test. Thus, after the initial sampling was completed an estimate of the power of both tests was available. At this point, power was compared to determine if the sample size for the t-test had to be increased or decreased to make the power of the t-test equal the power of

the nonparametric test. If the powers were equal on the initial sampling, then a power efficiency of 100 percent would be recorded since the sample sizes were the same. The case most often encountered was that the power of the t-test exceeded the power of the nonparametric test and the sample size of the t-test had to be decreased for its power to equal or envelop the power of the nonparametric test. This follows necessarily from the fact that the parametric assumptions were satisfied, giving the t-test superior power. The only feasible explanation for getting a power efficiency that exceeded 100 percent, in which case the samples for the t-test were increased, was the existence of sampling error in the random sampling process.

After the power of the nonparametric test had been enclosed, a linear interpolation method, similar to Hodges and Lehmann (1956), was used to equate powers. Linear interpolation was applied to the enclosing consecutive sample sizes of the t-test to determine a fractional sample size that equated power with the integer sample size of the nonparametric test. If, as the sample size for the t-test was being reduced, the power for a given sample equated exactly with that of the nonparametric test, then interpolation was not necessary and power efficiency was calculated by the ratio of integer values.

When the location-shift is zero, an empirical estimate of the probability of a Type I error (α) is given. When $\Theta = 0$ the null hypothesis is true, therefore the probability of a Type II error has no meaning. Only a Type I error can be made in this case, so for all mean differences of zero the proportion of rejections of the null hypothesis is an empirical estimate of the significance level. The accuracy of this empirical

α is an indication of the randomness of the simulation process and the validity of the tests.

Tables of critical values were read into the computer for use in testing the significance of the null hypothesis. Various sources provided these critical values. The extensive tables of Owen (1962) provided the critical t-values. As mentioned previously, the critical values for the Kolmogorov-Smirnov two-sample test for $m = n \leq 40$ were adapted from the tables in Massey (1951a) and Birnbaum and Hall (1960), and Conover (1971) provided the critical values for $m = n = 50$. The tables contained in Noether (1971) and in the appendix of Dixon and Massey (1969) were helpful in furnishing the critical values for the Mann-Whitney U test. Finally, as previously indicated, the probabilities for the sign test are given in any cumulative binomial table (for example, Walker and Lev, 1953). After each test was calculated, the value obtained was compared with the table value to determine significance.

The parameters that were ultimately evaluated comprised part of a much more comprehensive array of parameters that were originally intended for investigation, but available computer time restricted the number of alternatives that were evaluated. The study was originally begun by testing 1,000 pairs of samples. After running a substantial number of various parameter combinations, it was found that the results fluctuated too much to be of much value. Thus, despite the increased computer time involved, it was decided to decrease some of the parameters evaluated in order to increase the simulation to 2,000 test repetitions.

The choice between evaluating one-sided or two-sided tests was made in favor of one-sided alternatives because a directional alternatives hypothesis is the more powerful and the more meaningful test. The

analysis of one-tailed tests at α can be considered equivalent to the two-tailed versions at 2α. So even though only one-tailed tests were investigated for $\alpha = .05$ and .01, this could be considered the same as two-tailed tests with $\alpha = .10$ and .02. For the Kolmogorov-Smirnov two-sample test this symmetrical relationship is not exact but is close enough for most practical applications (see Bradley, 1968:292).

It was decided to investigate the power efficiency for samples $m = n = 6(2)20$* for the three tests. Each pair of samples of this size were evaluated for 2,000 repetitions. The results that were run initially for 1,000 samplings were retained and presented for samples $m = n = 30$, 40, and 50. It was felt that for these larger samples, samplings greater than 1,000 would be prohibitive in terms of computer time. For certain larger sample sizes the normal approximation was used instead of the exact nonparametric test.

This last point concerning the normal approximation discloses a basic problem in nonparametric statistical analysis. As mentioned previously, the results that were run with 1,000 repetitions displayed some significant fluctuations. One of the reasons for this is the discreteness of the underlying distributions of the nonparametric tests. Because these tests are based on discrete, and not continuous distributions, the significance levels are merely approximations, not exact. For example, if one was to apply the one-tailed sign test with a sample of 10 and $\alpha = .05$, the test must have one or fewer signs of the same kind to reject the null hypothesis. But the exact probability of obtaining one or fewer signs of the same kind in a sample of 10 is .011, not .05. The

---

*$m = n = 6(2)20$ is read as follows: Samples m and n range in size from 6 to 20, simultaneously, in increments of 2.

next higher critical value, two or fewer signs, has a probability of .055 of occurring. As another example, suppose the Kolmogorov-Smirnov two-sample one-tailed test was being applied to a sample of $m = n = 6$ with $\alpha = .01$. The tabled critical value given for these parameters represents an exact probability of rejecting a true hypothesis of .0011, not .01. The succeeding critical value has a probability of .0130 of occurring. Thus, one can see that the discreteness of the distributions can distort the power values that are obtained.

To rectify this, the power of the nonparametric test for the smaller samples was adjusted by interpolating the level of significance. As each set of data was tested, significance was checked for critical value bordering above and below the chosen significance level. Taking the two resultant empirical power values, a linear interpoliation was made to adjust the theoretical significance level to .05 or .01, whichever parameter was being considered. This was done by determining beforehand the factor that was necessary to correct α, reading this value into the computer, and simply calling for this value and multiplying as necessary to make the interpolation. Dixon (1954) and others have used a similar randomization technique to help eliminate the effect of discreteness.

A randomization procedure might be criticized on the grounds that the practitioner does not randomize the level of significance in field experiments and empirically it is not a true representation. However, from a theoretical standpoint, randomization is necessary because power efficiency, by definition, is based upon the assumption that all parameters, except sample size, are equal. The results varied a great deal between those that were randomized and those that were not. The interpoliation

procedure tended to "smooth" the power efficiency values and remove some of the variation manifested in the nonrandomized results.

Another criticism might stem from the fact that linear interpolation was applied to a nonlinear relationship. However, it was felt that the effect of this approximation would not distort any of the results to an appreciable degree.

Interpolation was performed on the significance level for the sign test and the Kolmogorov-Smirnov test for all samples from 6 to 20. The significance level of the Mann-Whitney U test was randomized for samples up to 12 only because the table values for exact probabilities did not exist for samples 14 and above. For samples $m = n \geq 14$, the normal approximation for the U-test was used, which was considered a relatively safe approximation (Mann and Whitney recommended the normal approximation for samples larger than eight). As the sample size approaches infinity, the discrete distribution of the nonparametric test approaches a continuous normal distribution by the central limit theorem. Thus, interpolation of the significance level is not as important for the larger samples as with smaller samples. Randomization was not performed on the larger samples (30, 40, and 50) for this reason.

The normal approximation was used for the sign test on samples greater than 20 and as mentioned above, the normal approximation was used for the Mann-Whitney U test on sample sizes 14 and above. When the normal approximation was applied to these larger samples, a continuous test was being used to estimate a discrete nonparametric test. This reduces the necessity of randomization; especially in view of the fact that increasing sample sizes approach continuity.

In summary, results are given for sample sizes $m = n = 6(2)20$, 30, 40, 50. Randomization was performed on some of the smaller samples and the normal approximation was used to calculate the sign test and the U-test for some of the larger sized samples. Statistical tests were performed 2,000 times each on samples $m = n = 6(2)20$ and 1,000 times on samples $m = n = 30$, 40, and 50. The one-tailed test for $\alpha = .05$ and .01 was investigated for normal location-shift alternatives $\Theta = 0.0(0.2)1.0$, 2.0, 3.0. The decision to restrict the analysis to equal samples and the selection of all parameters, in general, was guided primarily by computer time considerations. The choice of significance levels and sample sizes was also made in consideration of the common usage of these parameters in applied research.

A brief summary of each computer program should clarify the procedure used to calculate the power efficiency of each of the three nonparametric tests. All of the programs followed the same basic format. The sign test program began by generating two samples of size n. First, an observation was generated from a normal distribution with $\mu = 0$ and then from a normal distribution with $\mu = \Theta$. This was repeated n times to generate n pairs of samples. As each pair of scores were generated, the necessary values for computing the t-test were also compiled.

The sign test statistic was computed differently depending upon whether the sample size was larger than 20 or not. For $n \leq 20$, the critical value for the sign test was determined by counting the smaller number of signs. Each critical value, thus obtained, was compared with the two table values that enclosed the true significance level to determine if the computed value could be significant in either case. Then the t-test

(3.3) was performed on the same data. This process was repeated 2,000 times. Power values were then obtained by dividing the number of rejections in each case by the total number of trials, 2,000. At this point the power of the sign test was interpolated for the exact significance level. Depending upon whether the sample size for the t-test had to be increased or decreased to enclose the power of the sign test, the sample pairs were increased or decreased by one, new data generated, and the t-statistic computed for the new sample. The t-statistic calculated from each set of data was compared with the tabled value for n - 1 degrees of freedom to test the null hypothesis for significance. This was also repeated 2,000 times. After the power of the sign test was enclosed by the power resulting from two parametric samples, linear interpolation was used to determine a fractional sample size of the t-test that equated powers. Finally, the ratio of the two samples that resulted in equal power was printed as the power efficiency for that set of parameters.

The main difference that existed when the sample size exceeded 20 was that the sign test was calculated using the normal approximation (3.6). The resulting statistic was checked for significance with the IBM subroutine, NDTR (see IBM reference manual). This subroutine computes $Pr(X \leq x)$ where X is a random variable distributed normally with $\mu = 0$ and $\sigma^2 = 1$. Randomization of the significance level was only performed for the smaller sample sizes and not for samples of 30, 40, and 50. Moreover, only 1,000 tests were performed for each given sample of these larger sizes.

The steps in simulating the Kolmogorov-Smirnov test followed essentially the same order as the sign test. The entire sample of size m was generated from a normal distribution with $\mu = 0$ and then sample n

was generated from the same distribution with $\mu = \Theta$. The scores generated from each distribution were sorted into two separate groups in ascending order. After the classes were established and the frequency counts determined, the value D (3.7) was located. Then the test statistic D was compared with each of the two table values of the K-S statistic that enclosed the chosen significance level to determine if the test was significant at either level of significance. The t-statistic was also computed (3.2) and compared with the tabled t-value with $m + n - 2$ degrees of freedom to check for significance. The remainder of the procedure for the K-S test was the same as for the sign test.

Samples for the U-test were generated in the same manner as with the K-S test. The samples that were generated from separate distributions were ranked together in ascending order and then the ranks of the scores that were taken from the X distribution ($\mu = 0$) were summed. This value, ΣR_m, was necessary to calculate the value of U for all samples. For samples smaller than 14, the one-tailed test statistic for the Mann-Whitney U test was computed with (3.13). For samples sizes $m = n \geq 14$, the normal approximation (3.17) and the NDTR subroutine were utilized. Other facets of the program were similar to the Kolmogorov-Smirnov program.

The power efficiency values, which constituted the primary objective of this study, were computed in a similar fashion for all of the tests.

CHAPTER IV

RESULTS AND DISCUSSION

The results of the study are presented in two segments in the form of tables and discussion. The first segment contains the empirical probabilities of a Type I error. These simulated probabilities are given for the sign test, the Kolmogorov-Smirnov test, the Mann-Whitney U test, and their parametric equivalent--the t-test. The second segment contains the power efficiencies which are given for the same nonparametric tests for normal shift alternatives for significance levels of .05 and .01. The investigation included tests of the one-tailed variety. Accompanying each table is an analytical discussion concerning the important findings. In instances in which previous research provided data that was comparable with the results of this study, comparisons and general comments are made as to how and why these results support or dispute the previous findings.

EMPIRICAL PROBABILITY OF A TYPE I ERROR

When the means of the two sampling distributions are identical, the location-shift alternative is zero ($\Theta = 0.0$), in which case the only error that a statistical test can make is of the first type. The proportion of rejections of the null hypothesis gives an empirical estimate of the significance level. If the sampling is random, the empirical probability of a Type I error should approach the chosen significance level. However, this approximation is affected by, not only sampling error, but

the discreteness of the underlying distributions of the nonparametric tests as outlined in the previous chapter.

In most sampling processes, an element of error is expected. Error estimates can be made for the empirical probabilities of a Type I error with the standard error of proportions,

$$\sigma_p = \sqrt{\frac{\pi(1 - \pi)}{n}} \tag{4.1}$$

since the significance level is a proportion. A confidence interval, using the standard error of proportions, was established for each empirical value depending upon the significance level and the number of samples. A 95 percent confidence interval based on normal populations is given by

$$\pi \pm 1.96 \sqrt{\frac{\pi(1 - \pi)}{n}} , \tag{4.2}$$

where π is equal to the chosen significance level, .05 or .01. Although the confidence interval is nonsymmetrical for $\pi \neq .5$, equation (4.2) was used which implies a symmetrical interval. This approximation was made because $n\pi$ and $n(1 - \pi)$ are both greater than five which indicates that the normal approximation to the binomial is appropriate. Even for $n = 1{,}000$ and $\pi = .01$, $n\pi = 10.0$. The following confidence intervals were established depending upon the level of significance and the number of samples.

$\alpha = .05$, $n = 2{,}000$; .0404 to .0596

$\alpha = .05$, $n = 1{,}000$; .0365 to .0635

$\alpha = .01$, $n = 2{,}000$; .0056 to .0144

$\alpha = .01$, $n = 1{,}000$; .0038 to .0162

Values that lie outside their respective intervals were considered to have been influenced by an unusual amount of sampling error.

Table 1 contains the empirical probabilities of a Type I error for the one-tailed sign test and the one-tailed paired t-test under simulated test conditions. The note at the bottom of the table (and at the bottom of every table) explains the various conditions under which the results were generated, as was detailed in the previous chapter. The results in all of the tables for sample sizes 6 to 20 were based on 2,000 samples, whereas samples 30 to 50 were predicated on 1,000 samples. In addition, the significance level for certain small samples of the nonparametric tests were interpolated to correct for the discreteness of the test, and approximations to the sign test and the U-test were applied for certain large samples. Type I error results are given for the same sample sizes ($n = 6(2)20, 30, 40, 50$) and levels of significance ($\alpha = .05$ and .01) for which power efficiency results are given later in the chapter.

It should be noted in Table 1 that no values are presented for $\alpha = .01$, $n = 6$. The reason for this is that a sample size of six is too small to have a one-tailed significant difference in means at the .01 level of significance. The hypothesis can be rejected at the .0156 level of significance, but not at .01. Only two values in the table fall outside the 95 percent confidence interval. The sign test values for $\alpha = .05$ or .0250, associated with $n = 30$; and .0290, associated with $n = 50$, lie outside the confidence interval. Such values are, of course, to be expected as a result of sampling error. Undoubtedly, a non-interpolated significance level, a smaller number of samples taken, and the utilization of the normal approximation contributed to a significant portion of the error.

Table 1

Empirical Probability of a Type I Error for the Sign Test and the t-test for Various Sample Sizes

Sample Size n	$\alpha = .05$		$\alpha = .01$	
	Sign test	t-test	Sign test	t-test
6	.0478	.0480	-	-
8	.0460	.0475	.0115	.0130
10	.0494	.0550	.0095	.0115
12	.0531	.0505	.0082	.0075
14	.0491	.0490	.0099	.0110
16	.0528	.0485	.0092	.0110
18	.0478	.0485	.0092	.0065
20	.0468	.0475	.0111	.0110
30	.0250	.0509	.0070	.0120
40	.0460	.0529	.0090	.0060
50	.0290	.0609	.0050	.0100
Mean	.0448	.0508	.0090	.0088

Note: Probabilities for samples 6 through 20 were based on 2,000 test samples.

Probabilities for samples 30 through 50 were based on 1,000 test samples.

The significance level of the nonparametric test was randomized for samples 6 through 20.

The normal approximation for the sign test was used for samples 30, 40, and 50.

For the samples subjected to 2,000 repetitions, neither the sign test nor the t-test demonstrated any superiority in accepting the true hypothesis. In other words, the probability of making a Type I error was not consistently higher for either the sign test or the t-test for a given sample size. Apparently, only sampling variation caused the empirical probabilities to deviate around the given significance levels of .05 and .01.

Evidently, the situation was different for the larger samples that were subjected to only 1,000 test repetitions. For both $\alpha = .05$ and .01, the empirical probability of rejecting a true hypothesis for the t-test exceeds the corresponding values for the sign test five out of six times. And as pointed out previously, the confidence interval fails to enclose two of the values for the nonparametric test in this range. It appears that the lack of randomization and the fewer samples decreased the probabilities for the sign test. Five out of the six values, for the sign test, for samples 30, 40, and 50 are lower than any of the other sign test values for samples 6 through 20. At the same time, the empirical probabilities for the t-test reflect no significant differences between the larger samples and smaller samples. It should be emphasized that the data on which the nonparametric tests were performed were, in each particular test situation, the exact data used in the t-test for testing the null hypothesis of mean differences.

In general, the findings for the sign test indicate that for $n \leq 20$ both tests were performing in a random manner, but for $n > 20$, the sign test was influenced by sampling error more than the t-test. The arithmetic mean of each column is presented for an overall comparison. These means reflect a similar performance on the part of both tests.

The simulated probabilities of a Type I error for the Kolmogorov-Smirnov two-sample test and the t-test for independent samples are presented in Table 2. The results shown in Table 2 represent the fruit of two modifications made to the Kolmogorov-Smirnov test. The first modification, which was made on all of the tests, was to interpolate the significance level to correct for discreteness. This interpolation process helped to increase the previous empirical probabilities to more realistic values, as compared to the t-test figures. This indicated that the two tests were working in a more similar fashion than before.

After this first modification however, the empirical probabilities of a Type I error for the K-S test were still lower than one would expect solely on the basis of variation caused by sampling error. It was apparent that there was an additional element contributing to the distortion of the values.

This distorting factor resulted from the characteristics of the test procedure that was used. The Kolmogorov-Smirnov two-sample test as described in Siegel (1956:127-136) and Roscoe (1969:214-218) is calculated by determining the maximum difference between two cumulative frequency distributions. These distributions are established by setting up a given number of classes and determining the frequency count associated with each class. The K-S test statistic, D, is then given by the maximum frequency difference between respective classes. The question arises as to how many classes to establish. The general requirement, "as many as feasible," is rather nebulous in many practical situations. The decision to use n class intervals was a rather unfortunate one, because the results demonstrated fluctuations that simply could not be explained by sampling error alone. For example, every one of the empirical probabilities of a

Table 2

Empirical Probability of a Type I Error for the Kolmogorov-Smirnov Test and the t-test for Various Sample Sizes

Sample Size	$\alpha = .05$		$\alpha = .01$	
m = n	K-S test	t-test	K-S test	t-test
6	.0445	.0615	.0103	.0130
8	.0409	.0565	.0090	.0085
10	.0407	.0455	.0105	.0115
12	.0427	.0550	.0069	.0090
14	.0360	.0445	.0070	.0110
16	.0477	.0540	.0072	.0075
18	.0428	.0550	.0068	.0090
20	.0476	.0515	.0093	.0100
30	.0280	.0400	.0020	.0060
40	.0539	.0490	.0050	.0090
50	.0330	.0519	.0050	.0070
Mean	.0416	.0513	.0072	.0092

Note: Probabilities for samples 6 through 20 were based on 2,000 test samples.

Probabilities for samples 30 through 50 were based on 1,000 test samples.

The significance level of the nonparametric test was randomized for samples 6 through 20.

Type I error fell below the chosen significance level and less than 20 percent of the values were enclosed by a 95 percent confidence interval. These results are presented in Appendix B so that a comparison can be made with the final results.

The first revision consisted of increasing the number of classes to n + m. The results were improved, but were still not completely satisfactory. The final results, which are presented in this chapter, were based upon 2(n + m) classes. Comparing the findings in Table 2 with those in Appendix B reveals a significant improvement in the performance of the K-S test. Although the results were not as good as expected, it was felt that the K-S test and the t-test performed closely enough to substantiate the validity of the power efficiencies. Further improvements could have been realized by increasing the number of classes even more, but such computational detail was unrealistic from an applied standpoint. Perhaps a better solution would have been to treat the individual observations as discrete variables and thus, avoid the establishment of classes. This procedure was recently suggested by Conover (1971:309-314), but the traditional approach used in this study follows Siegel's technique (1956). Siegel's method was used because of the popularity and wide use of his book on nonparametric statistics. It was felt that since a majority of the analysts would probably follow Siegel's "bible" that the study would be most meaningful using his technique.

For $\alpha = .05$, only one of the 11 empirical probabilities exceeds .05, which reflects a downward bias as a result of too few classes. Three of the probability figures for the K-S test failed to be enclosed by the 95 percent confidence interval. These were for samples 14, 30, and 50. In contrast, only one of the probabilities for the t-test ($\alpha = .05$, $n = 6$)

failed to lie within the confidence limits. When all of the samples are considered, the mean probability for the K-S test is .0416 as compared to .0513 for the t-test.

The results for $\alpha = .01$ reflect a slight improvement over those for $\alpha = .05$. Only one value failed to be enclosed by the confidence interval. This was the probability for the K-S test associated with $m = n = 30$. It appears that the fewer samples and lack of interpolation was detrimental to the results obtained for the larger samples because three of five values for $m = n = 30$, 40, and 50 were outside the confidence limits. The average probability for the K-S test is .0072 and .0092 for the t-test, considering all samples when $\alpha = .01$. If the larger samples are ignored, these averages are .0084 and .0099 for the K-S test and t-test, respectively; not far apart.

In summary, the performance of the K-S test relies a great deal upon the method that is used to construct the cumulative frequency distributions. If an insufficient number of classes are established, power will suffer. The empirical probabilities reveal that the K-S test and the t-test performed in a fairly similar manner, although more of the K-S test values fell outside of the confidence limits than did the t-test values. The results for the samples for which the significance level was randomized and subjected to 2,000 test samples showed an improvement over the results for those samples that were not.

Table 3 contains the probabilities of a Type I error for the Mann-Whitney U test and the t-test for two independent samples. The only value that lies outside the confidence limits is the t-test value of .0615 for $n = 6$, $\alpha = .05$, which again can only be explained by sampling error. All of the U-test values were enclosed by the 95 percent confidence

Table 3

Empirical Probability of a Type I Error for the Mann-Whitney U Test and the t-test for Various Sample Sizes

Sample Size	α = .05		α = .01	
m = n	U-test	t-test	U-test	t-test
6	.0566	.0615	.0122	.0130
8	.0502	.0490	.0122	.0105
10	.0507	.0505	.0075	.0075
12	.0499	.0520	.0117	.0100
14	.0515	.0540	.0095	.0115
16	.0510	.0465	.0110	.0135
18	.0510	.0470	.0110	.0105
20	.0485	.0485	.0125	.0100
30	.0450	.0480	.0090	.0070
40	.0509	.0490	.0040	.0060
50	.0500	.0500	.0070	.0080
Mean	.0505	.0505	.0098	.0098

Note: Probabilities for samples 6 through 20 were based on 2,000 test samples.

Probabilities for samples 30 through 50 were based on 1,000 test samples.

The significance level of the nonparametric test was randomized for samples 6 through 12.

The normal approximation for the U-test was used for samples 14 through 50.

interval. Of the three nonparametric tests, the U-test performed closest to the t-test when $\Theta = 0.0$.

No trends in the probabilities are evident as the sample size increases, and the probability values for the U-test show no apparent advantage over the t-test or vice versa. About one half of the values that are different are greater for the t-test than for the U-test, so no advantage for either test is evident. The probability values for the U-test, which were interpolated ($m = n = 6$ to 12), do not demonstrate any discernible advantage over those sample values that were not interpolated ($m = n = 14$-50). Also, the normal approximation that was used to determine the critical values of the U-test for samples of sizes 14 to 50 had no apparent effect on the results.

It should be noted that the probability values for the U-test and the t-test for $\alpha = .01$ and $m = n = 30$ to 50 are all less than .01. Although these values are not significantly different from .01, the minor difference is probably attributable to the fewer samples that were taken (1,000) for these larger sized samples.

The Mann-Whitney U test was the only nonparametric test that had all of the probability values within the 95 percent confidence interval. The results of these empirical significance levels indicate that the U-test was closest to the t-test in performance. This is reflected in the equal arithmetic means that are presented at the bottom of Table 3. The U-test was followed by the sign test and finally the Kolmogorov-Smirnov test. Despite an occasional outlier, the empirical probabilities of a Type I error for each nonparametric and parametric test demonstrate a fairly equal chance of rejecting a true hypothesis. Thus, the power

efficiencies can be assumed to have been developed under comparable significance levels and test conditions.

POWER EFFICIENCY RESULTS

The primary goal of the present research is to provide power efficiencies for one-tailed versions of the two-sample sign test, Kolmogorov-Smirnov test, and Mann-Whitney U test for various sample sizes, significance levels, and normal location-shift alternatives. Power efficiencies are presented in tables according to the type of nonparametric test and significance level (.05 and .01). Each table covers sample sizes $m = n = 6(2)20, 30, 40, 50$, and normal shift alternatives $\Theta = 0.2(0.2)1.0, 2.0, 3.0$. As with the previous tables, footnotes to each table explain the various conditions under which the results were generated.

The tables are particularly useful to researchers for determining, prior to performing their tests, how much efficiency or power is being sacrificed by using these particular nonparametric tests in lieu of the t-test. Of course, this assumes that the particular nonparametric test, sample size, and chosen significance level are among those included in this study. Therefore, an attempt was made to cover as wide an array of parameter values as possible while avoiding the use of an excessive amount of computer time.

The power of a test, and thus power efficiency, is dependent upon three parameters--the significance level, the sample size, and the true difference between the result obtained by sampling and an established, or assumed, standard. For the purposes of this study, differences between population means are considered. This difference, which has been

defined as $\Theta = \frac{\mu_2 - \mu_1}{\sigma}$, is usually difficult to determine because the population mean is unknown in most practical situations. However, if the researcher is able to make an informative estimate as to the degree of the mean difference, then a fairly accurate power efficiency can be determined. To assist in this decision, Cohen (1969:22-25) has suggested values of $\Theta = 0.2$, 0.5, and 0.8 to represent "small," "medium," and "large" mean differences. Although these values are basically arbitrary, Cohen justified his choice in a logical manner that makes them conducive to practical application. Therefore, if the researcher has no better basis for estimating the extent to which the phenomenon exists in his data, then he can merely choose one of three relative measures that he feels best fits the situation. Since the tables do not contain a $\Theta = .05$ column, it will be necessary to determine the value midway between 0.4 and 0.6 whenever a "medium" difference in locations is predicted. The necessity for interpolating to obtain the power efficiency for $\Theta = .05$ will certainly not discredit the resultant value.

As will be evident later, all of the power efficiency values were subjected to a certain amount of sampling error. Therefore, it is not advisable to look at one value for a given set of parameters and say that that value is the exact power efficiency. A recommended procedure is to investigate the power efficiencies that immediately surround the value of interest. A cursory investigation of this sort should reveal if the particular value has been affected by a disproportionate amount of sampling error. If it appears that it has, then it would be appropriate to use a mean computed from the surrounding values and the value of interest. For example, should the power efficiency lie within the body of the table, then the eight surrounding efficiencies plus the

efficiency under investigation would comprise the mean. If the particular value is positioned in a corner of the table, that value plus the three adjacent values should be averaged. This procedure should help to evenly distribute the sampling error that is reflected in some power efficiencies to a greater degree than in others. In most cases, however, this adjustment will not be necessary.

Certain trends that are evident in the results will be explored, as well as deviations from these trends. In situations in which the results of this study can be compared with the findings of previous studies, such comparisons will be made. It should be emphasized that comparisons of this type are legitimate only in cases in which the methodology and parameters investigated are identical. As Bradley (1968:57) points out, "It [power efficiency] is an index which is both highly peculiar to experimental test conditions and highly realistic to them. (It is also highly peculiar to the mathematical procedures used to obtain it; other perfectly realistic definitions of relative efficiency, based on slightly different procedures, may lead to quite contrasting results.)" Therefore, a certain degree of tolerance should be allowed when comparing the results of this simulation study with those of a deterministic approach. Even though the parameters studied may be similar, differences in methodology are likely to cause disparate results.

Sign Test

The power efficiencies for the sign test with $\alpha = .05$ are presented in Table 4. Certain trends in the results are clearly evident. First is the presence of a fairly smooth transition from a power efficiency of about 80 percent for the very small samples to around 55 percent

Table 4

Empirical Power Efficiency of the Sign Test for Various Normal Shift Alternatives for Various Sample Sizes for $\alpha = .05$

Sample Size	Θ							Mean
n	0.2	0.4	0.6	0.8	1.0	2.0	3.0	
6	0.842	0.872	0.784	0.745	0.713	0.689	0.671	0.759
8	0.717	0.716	0.763	0.766	0.768	0.709	0.720	0.737
10	0.837	0.713	0.747	0.753	0.712	0.698	0.592	0.722
12	0.785	0.638	0.691	0.713	0.638	0.659	0.613	0.677
14	0.602	0.659	0.664	0.677	0.640	0.654	1.000	0.649*
16	0.902	0.613	0.740	0.688	0.674	0.641	1.000	0.710*
18	0.954	0.676	0.690	0.694	0.723	0.650	1.000	0.731*
20	0.960	0.684	0.689	0.634	0.637	0.592	1.000	0.699*
30	0.100	0.321	0.374	0.489	0.514	1.000	1.000	0.360*
40	0.365	0.561	0.597	0.601	0.681	1.000	1.000	0.561*
50	0.403	0.520	0.563	0.540	0.580	1.000	1.000	0.521*
Mean	0.679	0.634	0.664	0.664	0.662	0.662*	0.649*	

Note: Power efficiencies for samples 6 through 20 were based on 2,000 test samples.
Power efficiencies for samples 30 through 50 were based on 1,000 test samples.
The significance level of the nonparametric test was randomized for samples 6-20.
The normal approximation for the sign test was used for samples 30, 40, and 50.

* Modified mean excluding unbounded power efficiencies of 1.0.

for the larger samples. As might be expected, sampling error prevented a consistent transition from the higher to lower efficiencies.

A second noteworthy trend consists of a fairly steady decrease in power efficiency as the sample size and shift alternative increases. This is evidenced somewhat by the means that were calculated for every shift alternative and sample size in the table, but is more clearly evident in the individual rows and columns. For small samples and small alternatives, power efficiency is approximately 80 percent, decreasing to approximately 70 percent as Θ increases to 3.0. As n increases, the power efficiency tends to fall from these values for each location-shift alternative.

The literature review indicated that a significant amount of research has been conducted in reference to the sign test. The main reason for this is that, in contrast to most nonparametric tests, the power function of the sign test is simple to determine and fairly easy to manipulate. Although there are obvious differences in methodology, there was an opportunity for comparing the results of this study with previous investigations.

The trends in Table 4 support, in part, the findings of Dixon. By integrating the power function of the sign test, Dixon (1953) found that a decreasing power efficiency was generally associated with an increasing sample size and shift alternative. This is the same conclusion drawn from Table 4. However, it should be pointed out that Dixon (1954) later obtained results for the sign test (or median test) which conflicted with his previous conclusions. These results (with $n = 5$, $\alpha = .025$) indicated an increasing power efficiency associated with an increase in mean differences. The findings of most of the other researchers (Walsh,

Jeeves and Richards, and Milton) were not extensive enough to disclose any trends for similar situations.

Both of Dixon's studies were based on a deterministic analysis of the power functions. In addition to the basic differences in methodology between Dixon's work and the present study (see p. 67), comparisons must be made in light of any other differences that exist--particularly differences in parameters such as significance level and sample size. These seemingly insignificant differences theoretically invalidate legitimate comparisons. However, because so few efficiencies have been computed that are directly comparable with the results of this study, the theoretical framework will be stretched to include certain artificial comparisons to demonstrate the validity of the results of this investigation. This will be done with the Kolmogorov-Smirnov test and the Mann-Whitney U test as well as with the sign test.

Various degrees of sampling error are in evidence in Table 4. The efficiencies for $\Theta = 0.2$ are particularly susceptible to fluctuation. This applies not only to Table 4, but to all of the power efficiency tables. The values in Table 4 of 0.902, 0.954, and 0.960 for samples 16, 18, and 20, respectively, exemplify this variation. These three values appear to be higher than normal, as indicated by general trends. There are two main reasons for the excessive fluctuations associated with the small shift alternatives. In a previous study, Dixon (1953) found that applying linear interpolation to the sample sizes of the t-test to equate powers was inaccurate only for small location shifts. A similar linear interpolation process was used in the present study. The second reason stems from the ratio that was used to compute power efficiency. For small alternatives, statistical power is usually small; and a given

absolute change in sample size to equate powers will have a greater effect on the power efficiency ratio than when the same absolute adjustment is made to powers closer to unity, which is usually the case for large shift alternatives. Simply stated, a given change to values in a ratio that are close to zero changes that ratio relatively more than a given change to values in a ratio that are close to unity. Thus, fluctuations for $\Theta = 0.2$ were to be expected.

The power efficiencies for samples of 30 are unusually low. This is the smallest sample size that consisted of only 1,000 repetitions and an uninterpolated significance level. This might explain, at least partially, the unexpectedly small values.

It should be noted that, for $n = 14$, $\Theta = 3.0$ and certain other parameter combinations primarily in the lower right corner of the table, the power efficiencies are 1.000. In these cases the power of both tests have equaled 1.0 or have attained the same high power with equal sized samples. This situation points to one of the advantages of asymptotic relative efficiency. The A.R.E. theoretical construct avoids this possibility by assuming that $\Theta \to 0.0$ as $n \to \infty$, which keeps the powers for large samples bound from 1.0. Increasing the number of samples would have refined the simulation to a point that would have ultimately prevented the efficiencies from attaining values of one, but this was precluded by computing time considerations. Certain combinations of large samples and large location-shift alternatives produced power efficiencies of 1.0 in every table.

The relative efficiency value for $n = 50$ and $\Theta = 0.2$ (0.403) is smaller than the A.R.E. theory indicates that it should have been.

However, the asymptotic value of .637 is approximated by two values in the n = 40 row.

In general, the power efficiency of the sign test for $\alpha = .05$ decreases from 80 percent to about 70 percent as n and Θ begin to increase, and finally to just below 60 percent for large n. This is in general agreement with the deterministic findings of Dixon (1953) who found an efficiency above 70 percent for n = 6, and Jeeves and Richards (1950) who found a relative efficiency of approximately 70 percent for samples between 6 and 20. The results in Table 4 for a sample size of six are also fairly consistent with the analytical findings of Milton (1970:39). However, it should be noted that Milton's efficiencies, which were computed in the Hodges-Lehmann form, increased rather than decreased as the shift alternative increased.

A few inconsistencies among the findings of the analytical studies were discovered. Thus, it is clear that if power efficiency is obtained under different procedures, different values are likely to occur. However, despite obvious differences in methodology and parameters, the results for the sign test for $\alpha = .05$ appear to be valid and should be of benefit to researchers.

The power efficiencies for the sign test for $\alpha = .01$ are presented in Table 5. The results in Table 5 appear to fluctuate a little less than the values in the previous table. The power efficiencies do not demonstrate any large deviations, especially for $\Theta = 0.2$. The power efficiency values that are also noticeably more uniform are those for n = 30. The values for the larger samples are, as a whole, much more consistent with the established trends than the values in the previous table for the sign test. Power efficiency decreases fairly steadily as the sample size

Table 5

Empirical Power Efficiency of the Sign Test for Various Normal Shift Alternatives for Various Sample Sizes for $\alpha = .01$

Sample Size	Θ							Mean
n	0.2	0.4	0.6	0.8	1.0	2.0	3.0	
6	———	———	———	———	———	———	———	———
8	0.940	0.890	0.754	0.765	0.784	0.744	0.743	0.803
10	0.837	0.766	0.763	0.766	0.746	0.764	0.715	0.765
12	0.862	0.799	0.764	0.733	0.745	0.697	0.650	0.750
14	0.809	0.808	0.755	0.774	0.739	0.694	0.691	0.753
16	0.796	0.719	0.814	0.759	0.716	0.676	0.615	0.728
18	0.850	0.805	0.704	0.695	0.679	0.657	0.641	0.719
20	0.696	0.735	0.698	0.679	0.702	0.635	1.000	0.691*
30	0.386	0.688	0.667	0.673	0.656	0.592	1.000	0.610*
40	0.533	0.628	0.657	0.596	0.629	1.000	1.000	0.609*
50	0.682	0.631	0.600	0.613	0.632	1.000	1.000	0.632*
Mean	0.739	0.747	0.718	0.705	0.703	0.682*	0.676*	

Note: Power efficiencies for samples 6 through 20 were based on 2,000 test samples.
Power efficiencies for samples 30 through 50 were based on 1,000 test samples.
The significance level of the nonparametric test was randomized for samples 6-20.
The normal approximation for the sign test was used for samples 30, 40, and 50.

* Modified mean excluding unbounded power efficiencies of 1.0.

increases and as the shift alternative increases, as was true for $\alpha = .05$. This is evidenced from the individual values in the table, as well as the means in the table margins. It should be pointed out that these means do not include the power efficiency values of 1.0 that are in the lower right corner of all of the tables. It was felt that these atypical values distort the true power efficiency of the tests.

For small samples and small shifts the power efficiency was a little higher (about 90 percent) than in Table 4. For $n = 8$, power efficiency decreases from approximately .94 to .74 as Θ increases from 0.2 to 3.0. For medium sized samples (14 to 20) and shift alternatives (0.4 to 1.0), relative efficiency is roughly 75 percent. The power efficiencies of 1.0 are, again, present for large n and large Θ. As the sample size increases beyond 20, the efficiency of the sign test drops to less than 70 percent. For $n = 50$ and $\Theta = 0.2$, the power efficiency is 0.682, which is very close to the asymptotic relative efficiency of 0.637. In fact, the first five values in the last row average 0.631.

A noteworthy point is that generally the power efficiency values for $\alpha = .01$ are larger than the power efficiency values for $\alpha = .05$. Fifty of the total 63 power efficiencies that are different, are greater for $\alpha = .01$ than for $\alpha = .05$. Also, 15 of 17 means in Table 5 are larger than in Table 4. This evidence supports the hypothesis of Jeeves and Richards (1950) and Dixon (1953) that efficiency should increase as the significance level decreases. The results of Milton (1970:39) were too inconsistent to draw any definite conclusions concerning the effect of the significance level on power efficiency.

The same general conclusions and comparisons that were made in reference to the sign test with $\alpha = .05$ also apply when $\alpha = .01$. The

relative efficiencies in Table 5 support and broaden previous findings. In addition to those previously mentioned, the postulates of Siegel and Walsh are also supported. Siegel (1956) felt that the power efficiency of the sign test would be about 95 percent for n = 6, and decline steadily to 63 percent as n increased. Walsh (1946) believed that the power efficiency for very small samples would be approximately 95 percent and decrease as n increased, obtaining a value of around 75 percent for n = 13. The power efficiencies of Table 5 appear to be valid, even considering isolated fluctuations due to sampling error.

Kolmogorov-Smirnov Test

The power efficiencies of the one-tailed Kolmogorov-Smirnov test for $\alpha = .05$ are presented in Table 6. The efficiencies in Tables 6 and 7 are based upon an empirical cumulative frequency distribution that consisted of $2(n + m)$ class intervals. As mentioned previously, the initial simulation run of the K-S test consisted of n classes. However, these results proved unsatisfactory and the number of classes was increased to $2(n + m)$ before reasonable results were obtained.

The initial results (which are not presented) had power efficiencies of zero for the smaller samples and mean differences. These have been eliminated in Table 6, but the efficiencies in this range are still lower than indicated by previous research. Undoubtedly, increasing the number of classes would have increased the power efficiencies beyond the present values because an increase in efficiency accompanied each incremental increase in the number of classes. This situation was evaluated in the discussion pertaining to Table 2.

Table 6

Empirical Power Efficiency of the Kolmogorov-Smirnov Two-Sample Test for Various Normal Shift Alternatives for Various Sample Sizes for $\alpha = .05$

Sample Size	Θ							Mean
m = n	0.2	0.4	0.6	0.8	1.0	2.0	3.0	
6	0.399	0.675	0.655	0.648	0.681	0.667	0.658	0.626
8	0.425	0.534	0.636	0.661	0.716	0.728	0.731	0.633
10	0.713	0.499	0.590	0.668	0.654	0.695	1.000	0.636*
12	0.439	0.602	0.621	0.712	0.724	0.746	1.000	0.641*
14	0.534	0.606	0.702	0.670	0.675	0.663	1.000	0.642*
16	0.702	0.682	0.687	0.669	0.699	0.753	1.000	0.699*
18	0.536	0.624	0.630	0.648	0.727	0.706	1.000	0.645*
20	0.490	0.632	0.655	0.702	0.685	0.660	1.000	0.637*
30	0.405	0.634	0.615	0.617	0.680	1.000	1.000	0.590*
40	0.612	0.770	0.766	0.694	0.781	1.000	1.000	0.725*
50	0.392	0.582	0.606	0.640	0.740	1.000	1.000	0.592*
Mean	0.513	0.622	0.651	0.666	0.706	0.702*	0.694*	

Note: Power efficiencies for samples 6 through 20 were based on 2,000 test samples.
Power efficiencies for samples 30 through 50 were based on 1,000 test samples.
The significance level of the nonparametric test was randomized for samples 6-20.

* Modified mean excluding unbounded power efficiencies of 1.0.

The power efficiencies in Table 6 are around 50 percent for small samples and shift alternatives and gradually increase as the shift alternative increases. However, both the individual values in the table and the averages reveal that the transition from lower to higher power efficiencies is not smooth. For the larger mean differences and sample sizes, power efficiency is roughly 70 percent.

Research on the power of the K-S test has been somewhat restricted, not only by the difficulties encountered with the power function of the K-S test, but also by the fact that the K-S test does not have the necessary characteristics for computing A.R.E. But enough evidence is available to make a number of comparisons, which again must be made in light of differences in methodology and parameters, making the comparisons approximate at best.

The relatively low efficiency of the K-S test for small samples and shift alternatives, as indicated primarily by the means, conflicts with the findings of most of the earlier studies. Dixon (1954) determined a power efficiency of about 80 percent for $\Theta = 0.5$, $m = n = 5$, and $\alpha = .025$. Milton's (1970:40) power efficiencies for $m = n = 6$ and $\alpha = .05$ fell from 0.785 to 0.717 as Θ increased from 0.2 to 3.0. His value for $\Theta = 3.0$ (0.717) is not too far from that in Table 6 (0.658), but this is not true for the results for $\Theta = 0.2$.

Another general conclusion of previous investigations indicated that the K-S test had a power that was superior to the sign test for equal parameters. Dixon (1954), in particular, concluded that the relative efficiency of the K-S test would exceed that of the sign test for small alternatives, but that the advantage would fall as Θ increased. A comparison of Table 6 with Table 4 reveals that, for the smaller

parameters, the power efficiency of the sign test was superior, but the converse held for the larger parameter values.

A major portion of the explanation for the relatively poor performance of the K-S test for the smaller parameters derives from the test procedure, i.e., the power of the K-S test is directly dependent upon the number of empirical classes that are established in the computation of the K-S test statistic. The reason for this is that the K-S test is based upon the assumption of continuous data. Therefore, if the data are grouped, or if the continuous data are divided into too few classes, the test loses power. As Roscoe (1969:214-218) warned, a violation of the assumption of continuity of distribution could result in a great loss in power.

Perhaps another reason for the low power has to do with the test itself. As one of the early investigators of the Kolmogorov-Smirnov two-sample test pointed out (van der Waerden, 1953a), the K-S test is designed to detect differences of any type between populations. Thus, he suggested that the K-S test demonstrated inferiority to the classical tests in detecting, solely, mean differences because of the general nature of the K-S test. As with the classical tests, Gibbons (1971:173) noted that the median (sign) and the Mann-Whitney U tests were particularly sensitive to differences in location when the populations were identical otherwise. Therefore, a comparison of the sign test and the K-S test must be made under consideration of the types of differences that exist in the underlying populations. Only differences in location are considered when normal alternatives are under investigation, as in the present case. Roscoe (1969:217) summarizes it best; "Generally, a statistical test that may be rejected because of any one of several different kinds of departures from

the sampling distribution will be less powerful than a statistical test that concentrates on a single alternative to the null hypothesis." As the shift became more distant, the K-S test was able to utilize the information in the data more efficiently than the sign test, which resulted in generally higher power efficiencies.

The values in Table 6 are fairly close to the findings of Knott. Knott (1970) computed the efficiencies of the K-S test relative to a parametric test and obtained a general efficiency of 75 percent for $\alpha = .05$. This is comparable to many of the table figures for the larger shifts and samples.

The fairly consistent increase in power efficiency that was associated with an increase in mean difference supports certain previous findings and disputes others. Dixon (1954) concluded that power efficiency would fall as the normal alternative increases. Milton (1970:40), who used a numerical integration technique similar to Dixon's, finalized a similar conclusion by obtaining only one of many values that was contrary to Dixon's results. On the other hand, Lee (1966), who compared the exact power of the K-S test with that of the normal test for $m = n = 5$, and $\alpha = .05$ and .01, found an increasing relative efficiency for increasing normal alternatives. The reason the conclusions of Lee, rather than Dixon and Milton, are manifested in Tables 6 and 7 stems primarily from the arguments previously put forth concerning the relative performance of the K-S test. Differences in methodology and computational schemes for power efficiency must also be considered.

As with the sign test, the power efficiency values fluctuate more in the first column ($\Theta = 0.2$) for increasing samples than for any other trend segment in the table. Also, as with the sign test, Tables 6 and 7

show that the simulation process was not refined enough to eliminate power efficiencies of 1.0 for certain large parameter combinations.

The inconsistent means for the last three rows ($m = n = 30$, 40, and 50) indicate that only 1,000 samples and the lack of interpolating the significance level increased sampling error and prevented truly representative results from being generated for these three sample sizes. Despite this distortion, the estimated A.R.E. of .637 is approximated by some of the values in the last two rows of Table 6.

The simulated power efficiencies for the Kolmogorov-Smirnov two-sample test for $\alpha = .01$ are presented in Table 7. Because of the numerous similarities between the two tables, most of the comments made in reference to the previous table (Table 6) also apply here. Therefore, certain trends will be mentioned, but when they are similar to those in Table 6 an explanation will not be repeated.

One of the more obvious differences between the two tables is that the power efficiencies for $\alpha = .01$ are generally higher than those for $\alpha = .05$. Only two of the mean values in Table 7 are less than the respective means in Table 6. Dixon (1953; 1954), among others, predicted a decrease in efficiency as the significance level increases. The findings of Milton (1970:40) were inconsistent with respect to the performance of the test relative to the significance level. Thus, the results for the K-S test substantiate Dixon's hypothesis.

The means located at the bottom of Table 7 indicate an increase in power efficiency for more distant alternatives, as in Table 6. Another similarity between the values in the two tables is the power efficiencies of 1.0 for the larger parameter combinations. No trend in power efficiency is evident as the sample size increases.

Table 7

Empirical Power Efficiency of the Kolmogorov-Smirnov Two-Sample Test for Various Normal Shift Alternatives for Various Sample Sizes for $\alpha = .01$

Sample Size	Θ							Mean
m = n	0.2	0.4	0.6	0.8	1.0	2.0	3.0	
6	0.455	0.616	0.647	0.693	0.764	0.759	0.752	0.669
8	0.406	0.604	0.780	0.761	0.725	0.746	0.782	0.686
10	0.914	0.620	0.665	0.716	0.757	0.734	0.835	0.749
12	0.533	0.678	0.755	0.734	0.685	0.725	0.813	0.703
14	0.591	0.878	0.706	0.676	0.762	0.727	1.000	0.723*
16	0.703	0.647	0.725	0.775	0.710	0.727	1.000	0.714*
18	0.453	0.693	0.677	0.711	0.706	0.721	1.000	0.660*
20	0.799	0.660	0.618	0.648	0.738	0.743	1.000	0.701*
30	0.668	0.567	0.664	0.672	0.707	1.000	1.000	0.656*
40	0.711	0.447	0.646	0.702	0.725	1.000	1.000	0.646*
50	0.527	0.523	0.588	0.662	0.647	1.000	1.000	0.589*
Mean	0.614	0.630	0.679	0.704	0.720	0.735*	0.796*	

Note: Power efficiencies for samples 6 through 20 were based on 2,000 test samples.
Power efficiencies for samples 30 through 50 were based on 1,000 test samples.
The significance level of the nonparametric test was randomized for samples 6-20.

* Modified mean excluding unbounded power efficiencies of 1.0.

Although researchers would be interested in individual values, the overall efficiency appears to be in the 70-75 percent range. This compares very favorably with the general ratio of 72 percent of the K-S test that was obtained by Knott (1970) for $\alpha = .01$.

Siegel (1956:136) stated that, "The evidence seems to indicate that whereas for very small samples the Kolmogorov-Smirnov test is slightly more efficient than the Mann-Whitney test, for large samples the converse holds." Unfortunately, supporting data were not furnished. The conclusions gathered from this study and from that by Dixon (1954) agree that the Mann-Whitney U test is everywhere more powerful than the K-S test.

Mann-Whitney U Test

The power efficiency values for the Mann-Whitney U test for $\alpha = .05$ are presented in Table 8. The most striking feature of the U-test is that its power is obviously very close to the power of the t-test. For example, the lowest value in the table is 0.816 for $m = n = 6$ and $\Theta = 0.2$, which means that if the U-test is used instead of the t-test for a given sample, there is a sacrifice in power of less than 20 percent. A majority of the values in the table exceed 96 percent, which indicates a very high power for the U-test.

The power efficiency values fluctuate randomly throughout the table. The U-test does not exhibit any of the patterns that are evident in the tables for the sign test and the K-S test. The main reason no patterns are evident is that the power of the U-test is very close to the power of the t-test and the sampling procedure that was used in the simulation was not sufficiently refined to amplify the minute differences in

Table 8

Empirical Power Efficiency of the Mann-Whitney U Test for Various Normal Shift Alternatives for Various Sample Sizes for $\alpha = .05$

Sample Size	Θ							Mean
m = n	0.2	0.4	0.6	0.8	1.0	2.0	3.0	
6	0.816	0.965	0.931	0.931	0.935	0.945	0.974	0.928
8	0.845	0.963	0.998	0.957	0.929	0.896	1.000	0.931*
10	0.974	0.982	0.866	0.931	0.899	0.946	1.000	0.933*
12	0.962	0.973	0.866	0.932	0.902	0.957	1.000	0.932*
14	1.033	0.915	1.006	0.905	0.972	1.000	1.000	0.966*
16	0.896	0.888	0.863	0.943	0.949	1.000	1.000	0.908*
18	0.994	0.972	0.963	0.983	0.938	1.000	1.000	0.970*
20	0.912	0.993	0.974	0.926	0.932	1.000	1.000	0.947*
30	0.888	0.983	0.973	0.932	0.993	1.000	1.000	0.954*
40	0.990	0.960	0.962	0.950	0.988	1.000	1.000	0.970*
50	1.035	0.943	0.942	0.886	1.000	1.000	1.000	0.952*
Mean	0.940	0.958	0.940	0.934	0.944*	0.936*	0.974*	

Note: Power efficiencies for samples 6 through 20 were based on 2,000 test samples.
Power efficiencies for samples 30 through 50 were based on 1,000 test samples.
The significance level of the nonparametric test was randomized for samples 6-12.
The normal approximation for the U-test was used for samples 14 through 50.

* Modified mean excluding unbounded power efficiencies of 1.0.

power. There were simply not enough samples to eliminate the sampling error in the results--an error which confused the slight difference in power often enough to eliminate the possibility of any trends. This problem could have been eliminated by increasing the number of samples to get a more precise experiment, but computer time restrictions prevented any increase in the number of samples. This situation also applies to Table 9.

No trends were apparent for an increasing sample size or an increasing location-shift alternative. There are no noticeable differences between the results for which the exact U-test was used and the results for which the normal approximation was used. The exact U-test was applied to samples 6 to 12 and the normal approximation was used for samples 14 to 50. This point also represents the division between the samples that had an interpolated significance level and those that were not interpolated. Also, no differences could be observed between the efficiencies resulting from 2,000 samples ($m = n = 6$-20) and those resulting from 1,000 samples ($m = n = 30$-50).

Despite the fact that the power function of the Mann-Whitney test is extremely tedious to evaluate, much research has been conducted on its power. Probably the main reason for this is that the U-test is an extremely powerful and useful nonparametric test. The high values in Tables 8 and 9 are in general agreement with the deterministic efficiencies of Dixon (1954) and Milton (1970:37), and with the results of asymptotic studies. While Dixon (1954) obtained a power efficiency that decreased steadily as the alternative increased, Milton (1970:37) and Hodges and Lehmann (1956) found efficiencies that decreased and then increased as the shift alternative increased beyond a certain point. Hodges and Lehmann attributed this difference to the different methods that were

used to interpolate the sample size of the t-test. The literature supports the theory that power efficiency will increase as the sample size increases. However, sampling error camouflages any possible trend of this nature for the U-test.

It is important to look at the values in an area instead of merely selecting one power efficiency value because of the variation in the values throughout Table 8. This applies to all of the power efficiency tables. Sampling error and the high power of the U-test tends to disguise the true relative efficiency. As is the case in most of the tables, the values in the first column ($\Theta = 0.2$) of Table 8 fluctuate more than the values in any other column or row. In fact, this column contains both the lowest and the highest values in the table.

As with the previous tests, the power efficiencies for certain large parameter combinations are 1.0. But with the U-test, certain values exceed 1.0. These are 1.033, 1.006, and 1.035 for $m = n = 14$, $\Theta = 0.2$; $m = n = 14$, $\Theta = 0.6$; and $m = n = 50$, $\Theta = 0.2$, respectively. Since such values are theoretically impossible, the only feasible explanation for the values is sampling error in the simulation process.

The asymptotic relative efficiency of .955 for the U-test appears to be the value around which most of the values fluctuate. The power efficiency for $m = n = 50$, $\Theta = 0.2$ is 1.035, which is not very close to .955. However, some of the values in the last two rows are quite close to this A.R.E. value. Certainly, the power efficiencies reflect the very high power of the U-test as compared to the t-test under conditions of normality. The results for $\alpha = .01$ suggest the same conclusion.

Table 9 contains the power efficiencies of the Mann-Whitney U test for one-tailed normal alternatives with a .01 significance level.

The efficiencies in Table 9 are very similar to those in Table 8, and most of the conclusions drawn in reference to Table 8 are germane to Table 9.

A number of similarities between the two tables is readily apparent. Again, no patterns in relative efficiency are evident as either sample size or shift alternative increase. This is obvious from the fluctuating means of each row and column. For certain combinations of large shift alternatives and sample sizes the efficiencies are 1.0, as before. And, as in Table 8, three values exceed 1.0. In addition, one value is equal to 1.0 which represents equal power between the U-test and the t-test for the same sized samples. The A.R.E. estimate is also similar in the two tables. For $m = n = 50$, $\Theta = 0.2$, the power efficiency is 1.006 which is, on a comparative basis, not very close to the A.R.E. However, two values in the last two rows are quite close to .955; the modified mean of the last row is .954. As with most of the other tables, the values in the $\Theta = 0.2$ column in Table 9 fluctuate the most. This column contains both the maximum and minimum values in the table.

There is too much fluctuation in the figures to determine if power efficiency increases or decreases as the significance level changes from .05 to .01. In analyzing the rank-sum test (U-test), Dixon (1954) had a power efficiency that increased slightly as the significance level increased. Just the opposite occurred for the sign test and the K-S test in the same study. Bradley (1968:109) perhaps expressed the opinion of most statisticians--that efficiency is expected to fall as the level of significance is increased. Most of the results of Milton (1970:37) followed this pattern.

Table 9

Empirical Power Efficiency of the Mann-Whitney U Test for Various Normal Shift Alternatives for Various Sample Sizes for $\alpha = .01$

Sample Size	Θ							Mean
m = n	0.2	0.4	0.6	0.8	1.0	2.0	3.0	
6	0.961	1.024	0.978	0.963	0.938	0.932	0.940	0.962
8	0.778	1.000	0.989	0.970	0.935	0.961	0.934	0.938
10	1.033	0.993	0.979	0.975	0.956	0.947	0.867	0.964
12	0.874	0.949	0.854	0.866	0.868	0.841	1.000	0.875*
14	0.857	0.936	0.938	0.948	0.936	0.917	1.000	0.922*
16	0.898	0.946	0.967	0.953	0.936	0.906	1.000	0.934*
18	0.951	0.968	0.917	0.909	0.947	1.000	1.000	0.938*
20	0.979	0.904	0.904	0.912	0.908	1.000	1.000	0.921*
30	0.986	0.770	0.938	0.970	0.926	1.000	1.000	0.918*
40	0.725	0.977	0.957	0.980	0.970	1.000	1.000	0.922*
50	1.006	0.908	0.922	0.992	0.942	1.000	1.000	0.954*
Mean	0.913	0.943	0.940	0.949	0.933	0.917*	0.914*	

Note: Power efficiencies for samples 6 through 20 were based on 2,000 test samples.
Power efficiencies for samples 30 through 50 were based on 1,000 test samples.
The significance level of the nonparametric test was randomized for samples 6-12.
The normal approximation for the U-test was used for samples 14 through 50.

* Modified mean excluding unbounded power efficiencies of 1.0.

Just less than one half of the efficiencies in Table 9 were 96 percent or higher. In contrast, the power of the sign test was considerably lower. It should not be surprising that the power of the U-test demonstrated superiority over the sign test, even though both are tests of location. The U-test utilizes more information in the data by incorporating the relative magnitude of the differences in addition to the direction of the differences. Because the U-test has such high power, the U-test may be preferred to the t-test in many situations--especially in those in which normal conditions are doubtful.

CHAPTER V

SUMMARY AND CONCLUSIONS

The final chapter is divided into two parts. A summary of the simulation technique and the results of the study is presented first. This is followed by the conclusions drawn from the results of the simulation.

SUMMARY

Two-sample statistical tests are often used in business problems to examine the hypothesis of equality between populations or, more specifically, to examine the hypothesis of equality between population means. When the researcher is faced with such a problem, a decision must be made as to what type of test to apply. This decision should be based primarily upon what particular test is most appropriate. The appropriateness of a test is based upon what assumptions the researcher can justifiably make concerning the underlying populations. If normality can be assumed, then a parametric test is appropriate. However, if the researcher has reason to believe that normal conditions do not exist, then a nonparametric test is suitable. Thus, from both theoretical and practical standpoints, a criterion is needed to evaluate these two types of tests.

The most common method for comparing statistical tests is on the basis of their relative powers. This comparison is usually made in the form of power efficiency, which is the ratio of the sample sizes of a

parametric test and a nonparametric test that are required to equate the powers of the two tests. This ratio is usually computed under the assumption of normality. In this case, the underlying assumptions of both tests are satisfied.

The main purpose of this study was to determine the power efficiency of three nonparametric tests for a wide range of parameters. Such information would provide analysts with an a priori estimate of comparative powers for alternative tests, thus enabling them to make an enlightened choice among tests. The nonparametric tests, which include the sign test, the Kolmogorov-Smirnov test, and the Mann-Whitney U test, were chosen on the basis of their wide applicability to business problems. The test that is appropriate for the same type of problems when its parametric assumptions are met is Student's t-test--the paired t-test in the case of the sign test, and the t-test for independent samples for the Kolmogorov-Smirnov test and the Mann-Whitney U test. In this study, the power of each nonparametric test was compared with the power of its t-test equivalent.

Power is a function of the significance level, sample size, and the true difference between the hypothesized mean and the population mean. In order to do a thorough analysis of the power efficiency of each test it is necessary to evaluate a wide range of these parameter combinations. Each test was investigated for one-tailed significance levels of .05 and .01. Equal samples of size $m = n = 6(2)20, 30, 40, 50$ and location-shift alternatives $\Theta = 0.0(0.2)1.0, 2.0, 3.0$ comprise the parameters that were studied. Restrictions on computer time limited the analysis to these parameters.

A simulation technique was used since it permitted greater flexibility in analyzing a wider range of parameters than the more standard deterministic studies of the past. The investigation was made with a simulation process based on a Monte Carlo procedure of generating random normal deviates. Equal samples were considered drawn from normal distributions with variances equal to one; the first sample being drawn from a distribution with $\mu = 0$ and the second sample from a distribution with $\mu = \Theta$. The possibility of tied values in the samples was ignored since the pseudo-random number generator was capable of generating up to 2^{29} numbers before repeating. Two thousand separate samples were tested for each set of parameters for samples 6 to 20 and 1,000 repetitions for samples 30 to 50. Power was obtained by establishing a decision rule and determining the number of rejections in the total number of test samples.

The three nonparametric tests that were analyzed are based on discrete distributions. Therefore, it was necessary to interpolate the power of the nonparametric tests for an exact significance level of .05 or .01. Linear interpolation was performed for certain small samples of each test. In addition, as is done in applied research, the normal approximation was applied to the sign test and the Mann-Whitney U test when the samples were large enough to justify the approximation.

Because of the nature of the problem, it was possible to divide the findings into two categories--probability of a Type I error and power efficiency. The initial results concerned the probability of a Type I error. These represent the outcomes of simulating test performance with distributions that had equal means ($\Theta = 0.0$).

According to the empirical probabilities of a Type I error, the sign test and the t-test performed similarly for sample sizes 6 to 20. However, for samples 30 to 50, the empirical probabilities for the sign test were generally lower than the corresponding probabilities for the t-test. In fact, two of the sign test values for $\alpha = .05$ were not enclosed by a 95 percent confidence interval which indicates the presence of sampling error and, perhaps, a slight bias. The decrease in the number of samples from 2,000 to 1,000 would increase sampling error and a non-interpolated significance level appears to have introduced a downward bias into the results. Also, the normal approximation was applied to samples of these sizes.

The empirical probabilities of a Type I error for the Kolmogorov-Smirnov test point to an interesting property of the test. The test procedure, which follows Siegel (1956:127-136) and Roscoe (1969:214-218), was initially based upon two empirical cumulative frequency distributions that included n class intervals. However, the results reflected an extreme bias. In anticipation of obtaining more representative results, the test procedure was revised to include $n + m$ classes where n and m represent the number of elements in the two samples. The results improved substantially, but were not completely satisfactory. The final results, which are presented in Chapter IV, were generated from tests including $2(n + m)$ classes. These results can be compared with Appendix B to determine how the performance of the K-S test improved.

For both $\alpha = .05$ and $\alpha = .01$, only four of 22 empirical probabilities for the K-S test failed to be enclosed by a 95 percent confidence interval. In contrast, only one of the probabilities for the t-test suffered a similar malady. For the K-S test, most of the problem occurred

in the larger samples (m = n = 30-50). These samples were subjected to only 1,000 test repetitions and the significance level was not interpolated. These characteristics, coupled with the problem in the establishment of classes, gave most of the values a downward bias. However, the means for each column reflect a fairly close overall performance for the two tests.

The empirical probabilities of a Type I error in Table 3 reveals that the Mann-Whitney U test and the t-test performed in a very similar fashion. Not only were all of the probabilities for the U-test within 95 percent confidence limits, but neither the U-test nor the t-test demonstrated any superiority or consistency over the other test.

The second, and the primary set of findings were the power efficiency results. Power efficiencies are presented in tables for each nonparametric test for α = .05 and .01. All of the values were subjected to a certain amount of sampling error and, as a result, tend to fluctuate.

Previous research has provided a limited number of values that could be used as guidelines for determining the accuracy of the results. Much of the previous work has, however, been done with asymptotic relative efficiency which only provides a lower limit to the power efficiency function. The asymptotic relative efficiency of the sign test is .637 and that for the Mann-Whitney U test is .955. Although the Kolmogorov-Smirnov test does not have the characteristics necessary for determining a true asymptotic relative efficiency, it is believed that the value lies somewhere between .637 and 1.0 as explained by Bradley (1968:291).

It was necessary to make comparisons with the asymptotic and deterministic findings in knowledge of certain stringent conditions. The first of these has to do with the methodology utilized to generate the

efficiencies. When different mathematical procedures are used to obtain relative efficiency, the outcomes may be quite contrasting values. Therefore, certain differences were to be expected when comparing the results of this simulation study with those of a deterministic study which is usually based on the integration of power functions. Another conditional factor involves the equality of parameters that prevail in the comparison. To legitimately compare power efficiencies, the parameters of α, n, and θ must be equal. However, because no values existed that fulfilled these criteria, the theoretical bounds were violated in order to make certain comparisons that normally are questionable. It is believed that, in many cases, despite slight differences in parameters, the value being compared would change very little for the parameters to agree.

The power efficiencies of the sign test for $\alpha = .05$ reflect results that are quite consistent with the isolated values that have been found in previous studies. Power efficiency decreased fairly steadily, from values around .80, as the sample size and shift alternative increased, to values generally between .50 and .60.

The efficiencies in all of the tables exhibited fluctuations which indicated the presence of sampling error. This was mostly the result of too few samples. A parameter for the sign test for which the power efficiencies exhibited an unusual trend was n = 30. The power efficiencies for n = 30 were substantially lower than they should have been, as indicated by the surrounding values. It should be pointed out that n = 30 was the smallest sample that was based on 1,000 tests, for which the significance level was not interpolated, and for which the normal approximation to the sign test was applied.

Two characteristics were revealed by all of the tests. First, the relative efficiencies in the first column ($\Theta = 0.2$) were particularly plagued by variation. Secondly, the efficiencies for large samples combined with large shift alternatives tended to be 1.0.

The trends in the power efficiencies of the sign test for $\alpha = .01$ appear to be smoother than the power efficiencies for $\alpha = .05$. As with $\alpha = .05$, power efficiency decreased as the sample size increased and as the location shift alternative increased. The values for the shift alternative 0.2 were substantially smoother than the respective values for $\alpha = .05$.

By comparing the performance of the sign test for $\alpha = .05$ with the sign test for $\alpha = .01$, another point was obvious. The power efficiencies were generally higher for $\alpha = .01$. Most authorities agree that power efficiency should increase as the significance level decreases. Power efficiency was around 90 percent for small samples and small shift alternatives, decreasing to about 75 percent for the medium-sized samples and shift alternatives, and finally, the values very close to an A.R.E. of .637 for the large samples.

In calculating the Kolmogorov-Smirnov test, the study followed the procedure outlined by Siegel (1956) because it is felt that his classical text furnishes the guidelines for a majority of the analyses using nonparametric tools. This procedure is based upon the establishment of empirical cumulative frequency distributions involving an arbitrary number of classes. As outlined previously, the final results of the K-S test were based on $2(n + m)$ classes.

For $\alpha = .05$, the power efficiency of the K-S test was in the neighborhood of 50 percent for the smaller samples and location-shift

alternatives and increased to around 70 percent for large parameter values. Previous investigations of the K-S test gave conflicting evidence as to whether the efficiency should increase or decrease for more distant differences in means. The values for $m = n = 30$-50 indicate that only 1,000 samples and the failure to interpolate the significance level caused the power efficiency to fluctuate more than usual.

The increase in power efficiency as the probability of a Type I error changed from .05 to .01 was predicted, as with the sign test. A majority of the efficiencies for the K-S test for both significance levels were in the area of 70 to 75 percent which is in agreement with the findings of Dixon, Milton, and Knott, among others.

A majority of the values for the Mann-Whitney U test for $\alpha = .05$ exceeded 96 percent which indicates how close the power of the U-test is to the power of the t-test. There was no evidence of any trends in the power efficiencies of the U-test for increasing sample sizes or shift alternatives. The lowest value obtained was .816 which represents a relatively small sacrifice in power when the U-test is used instead of the t-test. Some of the power efficiencies exceeded 1.0 which can only be attributed to sampling error.

The power efficiencies of the Mann-Whitney U test for $\alpha = .01$ were very similar to the values obtained for $\alpha = .05$. The relative efficiencies fluctuated primarily between 90 and 100 percent, which coincides with the results for $\alpha = .05$, and also with the findings of asymptotic theory. The values that were obtained fluctuated irregularly and prevented any patterns from emerging.

CONCLUSIONS

The results obtained from simulating the probability of a Type I error indicate that, in general, each nonparametric and parametric test was operating under similar test conditions, and, therefore, valid findings were produced in the study. It is evident that the Mann-Whitney U test performed closest to the t-test in rejecting a true hypothesis. The U-test was followed closely by the sign test, and then the Kolmogorov-Smirnov test.

A slight bias is noticed in the empirical Type I error probabilities for the sign test and the Kolmogorov-Smirnov test for the larger samples ($m = n = 30$-50). This can be explained by the fact that only 1,000 tests repetitions were performed on these sample sizes and the significance level was not interpolated.

The performance of the Kolmogorov-Smirnov test showed marked improvement after the number of class intervals was increased from n to $2(m + n)$. The reason for the improvement in the performance of the K-S test is fairly straightforward. The test validity is based upon the assumption of a continuous underlying distribution. Thus, when the data are not continuous or are assigned to too few classes (as in the initial case), the test loses much of its power. Therefore, the researcher is cautioned to establish at least $2(n + m)$ class intervals to maintain the validity of the K-S test. Increasing the number of intervals in the simulation beyond $2(n + m)$ would have undoubtedly improved the test performance, but it was deemed impractical from an applied standpoint.

The major contribution of the study consists of the power efficiency tables that cover a wide range of parameter values. As expected

in any simulation process, a certain amount of sampling error was present which caused some random fluctuations in the results. It is advisable for the analyst to investigate the particular power efficiency that is of interest, to determine if it seems to contain a disproportionate share of error. If so, then it is recommended that a mean be computed from the value of interest and the surrounding values to obtain a more representative estimate of efficiency.

Fluctuating power efficiencies were particularly evident for the smallest location-shift alternative ($\Theta = 0.2$). The reason for this is twofold. First, as Dixon (1953) pointed out, applying linear interpolation to the integer sample sizes of the t-test to equate powers (which was the process used in this study) is inaccurate for shift alternatives approaching zero. Secondly, a given change in power when power is low, which is usually the case for shift alternatives near zero, affects the power efficiency ratio more than when the same change is made to power values that are close to one.

Another characteristic that is evident in all of the tables is the efficiencies of 1.0 for the large parameter combinations. The reason for this occurrence was that, as the mean difference grew larger, the power of both the nonparametric and the parametric tests approached 1.0. This was especially true as the sample size increased, resulting in a ratio of identical sample sizes. The asymptotic relative efficiency concept prevents this from happening as $n \to \infty$ by restricting the shift alternative such that $\Theta \to 0.0$. Thus, the powers of both tests are bound from unity. But the simulation process followed practical operations by letting the powers approach unity.

The relative efficiency of the sign test decreased from approximately 80 percent for the smaller parameter combinations (n and Θ) to around 50-60 percent as the parameters increased. The power efficiencies for $\alpha = .01$ were generally higher than those for $\alpha = .05$. These findings support and extend the few isolated results of previous deterministic and asymptotic studies.

For the smaller parameters, the power efficiencies of the Kolmogorov-Smirnov test were approximately 50 percent--somewhat less than this writer anticipated. The power efficiency increased to 70-75 percent as the mean difference and sample size increased. As expected, the efficiencies generally increased as the significance level decreased.

The evidence suggested that the K-S test would outperform the sign test for all parameter values. This proved not to be true for the smaller parameters. Most of the fault undoubtedly stemmed from the class interval problem that was previously mentioned. The power of the K-S test relies upon the assumption of continuity and if this assumption is violated by creating too few classes then performance suffers. There is also the possibility that the characteristics of the K-S test were a factor. The K-S test is designed to detect differences of any sort between populations, whereas the sign test and the Mann-Whitney U test are designed specifically to detect differences in location. Thus, when the K-S test is applied to local normal alternatives, it will usually perform less powerfully than a test designed to concentrate on a difference in location. Since the K-S test incorporates the magnitude as well as the direction of the difference in means, its power increases relatively greater than the sign test as Θ or n increases.

The Mann-Whitney U test power efficiencies fluctuated primarily in the 90-100 percent bracket which reflects a relatively high power for a nonparametric test.

The values for the U-test fluctuated irregularly and prevented any patterns from emerging for changes in the sample size, shift alternative, or significance level. Essentially, the problem is that the powers of the U-test and the t-test were so close that the sampling process was not refined enough to expose the minute differences in power. This problem could possibly have been reduced by increasing the number of samples. However, that would have involved an inordinate amount of computer time.

Increasing the number of samples would have achieved several improvements. It would have eventually eliminated the power efficiencies of 1.0 that were attained for large sample sizes and large location-shift alternatives. It also would have improved the consistency of the results for the large samples (30 to 50). Making a general comparison between the results for samples 6 to 20 and those for samples 30 to 50 reveal that both increasing the number of samples and randomization of the significance level improved the consistency of the power efficiencies. In particular, the results for the sample size of 30 demonstrated that more than 1,000 test repetitions would have been beneficial.

The power efficiency results of this simulation study reveal a power hierarchy for the two-sample tests that were investigated. As expected, the performance of the t-test was superior to all of the tests studied, because under conditions of normality the t-test is the most powerful test for detecting a difference in central tendency. The power of the U-test was obviously very close to that of the t-test. The U-test is recommended over the t-test in all cases for testing the hypothesis of

equal means, except those in which the underlying distributions can be safely assumed to be normal. The K-S test is preferred to the sign test when large samples or large location-shift alternatives are encountered. However, when small samples or alternatives are involved the evidence of this study favors the sign test, especially when one considers how easy the sign test is to compute.

While conducting this study, a number of questions arose which are beyond the scope of the present study but are certainly worthy of attention. One of the more obvious avenues of further research is the investigation of power efficiency under non-normal conditions. Although some research has been done in this area, the choice of various types of underlying distributions, skewness, and kurtosis is quite extensive. Another area that is worthy of investigation is the effect on power efficiency of unequal variances in the underlying populations. Finally, there are many more nonparametric tests other than the three investigated in the present study. These also should be analyzed to determine their power efficiencies.

REFERENCES

Anderson, N. H. (1961). Scales and statistics: parametric and nonparametric. Psychological Bulletin, 58, 305-316.

Auble, D. (1953). Extended tables for the Mann-Whitney statistic. Bulletin of the Institute of Educational Research at Indiana University, 1, 1-33.

Bahadur, R. R. (1960a). On the asymptotic efficiency of tests and estimates. Sankhya: The Indian Journal of Statistics, 22, 229-252.

_____ (1960b). Stochastic comparison of tests. Annals of Mathematical Statistics, 31, 276-295.

_____ (1960c). Simultaneous comparison of the optimum and sign tests of a normal mean. Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. eds. I. Olkin, and others. Stanford: Stanford University Press. 79-88.

_____ (1967). Rates of convergence of estimates and test statistics. Annals of Mathematical Statistics, 38, 303-324.

Barton, D. E. (1957). A comparison of two sorts of test for a change of location and applicable to truncated data. Journal of the Royal Statistical Society, B, 19, 119-124.

Basu, D. (1956). The concept of asymptotic efficiency. Sankhya: The Indian Journal of Statistics, 17, 193-196.

Birnbaum, Z. W. (1952). Numerical tabulation of the distribution of Kolmogorov's statistic for finite sample size. Journal of the American Statistical Association, 47, 425-441.

_____ (1953). Distribution-free tests of fit for continuous distribution functions. Annals of Mathematical Statistics, 24, 1-8.

_____ and Hall, R. A. (1960). Small sample distributions for multisample statistics of the Smirnov type. Annals of Mathematical Statistics, 31, 710-720.

Blomqvist, N. (1950). On a measure of dependence between two random variables. Annals of Mathematical Statistics, 21, 593-600.

Blum, J. R. and Fattu, N. A. (1954). Nonparametric methods. Review of Educational Research, 24, 467-487.

Blyth, C. R. (1958). Note on relative efficiency of tests. Annals of Mathematical Statistics, 29, 898-903.

Boneau, C. A. (1960). The effects of violations of assumptions underlying the t-test. Psychological Bulletin, 57, 49-64.

_____ (1962). A comparison of the power of the U and t-tests. Psychological Review, 69, 246-256.

Bradley, J. V. (1968). Distribution-Free Statistical Tests. Englewood Cliffs: Prentice-Hall, Inc.

Bradley, R. A. (1967). Topics in rank-order statistics. Proceedings of the Fifth Symposium on Mathematical Statistics. Vol. 1. Eds. L. le Cam and J. Neyman. Berkeley: University of California Press, 593-607.

Capon, J. (1965). On the asymptotic efficiency of the Kolmogorov-Smirnov test. Journal of the American Statistical Association, 60, 843-853.

Claypool, P. L. (1970). Linear interpolation within McCornack's table of the Wilcoxon matched pair signed rank statistic. Journal of the American Statistical Association, 65, 974-975.

Cochran, W. G. (1937). The efficiencies of the binomial series tests of significance of a mean and of a correlation coefficient. Journal of the Royal Statistical Society, 100, Part I, 69-73.

Cohen, J. (1969). Statistical Power Analysis for the Behavioral Sciences. New York: Academic Press.

Conover, W. J. (1971). Practical Nonparametric Statistics. New York: John Wiley & Sons, Inc.

Darling, D. A. (1957). The Kolmogorov-Smirnov, Cramer-von Mises tests. Annals of Mathematical Statistics, 28, 823-838.

Dixon, W. J. (1953). Power functions of the sign test and power efficiency for normal alternatives. Annals of Mathematical Statistics, 24, 467-473.

_____ (1954). Power under normality of several nonparametric tests. Annals of Mathematical Statistics, 25, 610-614.

_____ and Massey, F. J., Jr. (1969). Introduction to Statistical Analysis. 3rd ed. New York: McGraw-Hill Book Co.

_____ and Mood, A. M. (1946). The statistical sign test. Journal of the American Statistical Association, 41, 557-566.

_____ and Teichroew, D. (1954). Some sampling results on the power of nonparametric tests against normal alternatives (abstract). Annals of Mathematical Statistics, 25, 175.

Epstein, B. (1955). Comparison of some non-parametric tests against normal alternatives with an application to life testing. Journal of the American Statistical Association, 50, 894-900.

Fraser, D. A. S. (1957). Nonparametric Methods in Statistics. New York: John Wiley & Sons, Inc.

Gaito, J. (1959). Non-parametric methods in psychological research. Psychological Reports, 5, 115-125.

_____ **(1960).** Scale classification and statistics. Psychological Bulletin, 67, 277-278.

Gibbons, J. D. (1963). On the power of rank tests on the equality of two distribution functions (abstract). Annals of Mathematical Statistics, 34, 355.

_____ **(1964).** Effect of non-normality on the power function of the sign test. Journal of the American Statistical Association, 59, 142-148.

_____ **(1971).** Nonparametric Statistical Inference. New York: McGraw-Hill Book Co.

Gleser, L. J. (1964). On a measure of test efficiency proposed by R. R. Bahadur. Annals of Mathematical Statistics, 35, 1537-1544.

Goodman, L. A. (1954). Kolmogorov-Smirnov tests for psychological research. Psychological Bulletin, 51, 160-168.

Hájek, J. (1969). A Course in Nonparametric Statistics. San Francisco: Holden-Day, Inc.

Harshbarger, T. R. (1971). Introductory Statistics: A Decision Map. New York: Macmillan Co.

Hemelrijk, J. (1961). Experimental comparison of Student's and Wilcoxon's two sample tests. Quantitative Methods in Pharmacology. Ed. H. De Jonge. Amsterdam: North-Holland Publishing Co., 118-134.

Hodges, J. L., Jr. and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. Annals of Mathematical Statistics, 27, 324-335.

_____ **(1961).** Comparison of the normal scores and Wilcoxon tests. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, ed. J. Neyman. Berkeley: University of California Press, 307-317.

Hoeffding, W. and Rosenblatt, J. R. (1955). The efficiency of tests. Annals of Mathematical Statistics, 26, 52-63.

Hoel, P. G. (1962). Introduction to Mathematical Statistics. 3rd ed. New York: John Wiley & Sons, Inc.

Hollander, M. (1967). Asymptotic efficiency of two nonparametric competitors of Wilcoxon's two sample test. Journal of the American Statistical Association, 62, 939-949.

International Business Machines Corporation Application Program. System/360 Scientific Subroutine Package (SSP). Reference manual H20-0205-3, Version III, 4th ed.

Jacobson, J. E. (1963). The Wilcoxon two-sample statistic: tables and bibliography. Journal of the American Statistical Association, 58, 1086-1103.

Jeeves, T. A. and Richards, R. (1950). A note on the power of the sign test (abstract). Annals of Mathematical Statistics, 21, 618.

Klotz, J. (1967). Asymptotic efficiency of the two sample Kolmogorov-Smirnov test. Journal of the American Statistical Association, 62, 932-938.

Knott, M. (1970). The small sample power of one-sided Kolmogorov tests for a shift in location of the normal distribution. Journal of the American Statistical Association, 65, 1384-1391.

Kolmogorov, A. N. (1933). Sulla determinatione empirica di una legge di distributione. Giornale dell' Instituto Italiano degli Attuari, 4, 83-91.

_____ (1941). Confidence limits for an unknown distribution function. Annals of Mathematical Statistics, 12, 461-463.

Lee, S. W. (1966). The Power of the One-sided and One-sample Kolmogorov-Smirnov Test. Unpublished Master's thesis, Kansas State University.

Lehmann, E. L. (1953). The power of rank tests. Annals of Mathematical Statistics, 24, 23-43.

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. Journal of the American Statistical Association, 62, 399-402.

Mc Cornack, R. L. (1965). Extended tables of the Wilcoxon matched pair signed rank statistic. Journal of the American Statistical Association, 60, 864-871.

Mac Stewart, W. (1941). A note on the power of the sign test. Annals of Mathematical Statistics, 12, 236-239.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. Annals of Mathematical Statistics, 18, 50-60.

Massey, F. J., Jr. (1950a). A note on the estimation of a cumulative distribution function by confidence intervals. Annals of Mathematical Statistics, 21, 116-119.

_____ (1950b). A note on the power of a nonparametric test. Annals of Mathematical Statistics, 21, 440-443.

_____ (1951a). The distribution of the maximum deviation between two sample cumulative step functions. Annals of Mathematical Statistics, 22, 125-128.

_____ (1951b). The Kolmogorov-Smirnov test for goodness of fit. Journal of the American Statistical Association, 46, 68-78.

_____ (1952a). Distribution table for the deviation between two sample cumulatives. Annals of Mathematical Statistics, 23, 435-441.

_____ (1952b). Correction to "A note on the power of a nonparametric test." Annals of Mathematical Statistics, 23, 637-638.

Miller, L. H. (1956). Table of percentage points of Kolmogorov statistics. Journal of the American Statistical Association, 51, 111-121.

Milton, R. C. (1964). An extended table of critical values for the Mann-Whitney (Wilcoxon) two-sample statistic. Journal of the American Statistical Association, 59, 925-934.

_____ (1970). Rank Order Probabilities: Two-Sample Normal Shift Alternatives. New York: John Wiley and Sons, Inc.

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. Annals of Mathematical Statistics, 25, 514-522.

Moses, L. E. (1952). Non-parametric statistics for psychological research. Psychological Bulletin, 49, 122-143.

Neave, H. R. and Granger, C. W. J. (1968). A Monte Carlo study comparing various two-sample tests for differences in mean. Technometrics, 10, 509-522.

Noether, G. E. (1955). On a theorem of Pitman. Annals of Mathematical Statistics, 26, 64-68.

_____ (1958). The efficiency of some distribution-free tests. Statistica Neerlandica, 12, 63-73.

_____ (1967). Elements of Nonparametric Statistics. New York: John Wiley & Sons, Inc.

_____ (1971). Introduction to Statistics: A Fresh Approach. Boston: Houghton Mifflin Co.

Owen, D. B. (1962). Handbook of Statistical Tables. Reading, Mass.: Addison-Wesley Publishing Co.

_____ (1965). The power of Student's t-test. Journal of the American Statistical Association, 60, 320-333.

Pitman, E. J. G. (1948). Lectures in Non-Parametric Statistical Inference. New York: Columbia University.

Roscoe, J. T. (1969). Fundamental Research Statistics for the Behavioral Sciences. New York: Holt, Rinehart and Winston, Inc.

Savage, I. R. (1953). Bibliography of nonparametric statistics and related topics. Journal of the American Statistical Association, 48, 844-906.

_____ (1957). Nonparametric statistics. Journal of the American Statistical Association, 52, 331-344.

_____ (1962). Bibliography of Nonparametric Statistics. Cambridge: Harvard University Press.

_____ (1969). Nonparametric statistics: a personal review. Sankhya: The Indian Journal of Statistics, 31, 107-144.

Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. New York: McGraw-Hill Book Co.

_____ (1957). Nonparametric statistics. The American Statistician, 11, 13-19.

Smirnov, N. (1939). On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin Mathematique de l'Universite de Moscou, 2, fasc. 2.

_____ (1944). Approximate laws of distribution of random variables from empirical data. Uspehi Matem. Nauk, 10, 179-206.

_____ (1948). Table for estimating the goodness of fit of empirical distributions. Annals of Mathematical Statistics, 19, 279-281.

Smith, K. (1953). Distribution-free statistical methods and the concept of power efficiency. Research Methods in the Behavioral Sciences. eds. L. Festinger and D. Katz. New York: Dryden Press, 536-577.

Stevens, S. S. (1946). On the theory of scales of measurement. Science, 103, 677-680.

_____ (1968). Measurement, statistics, and the schemapiric view. Science, 161, 849-856.

Stuart, A. (1954a). Asymptotic relative efficiencies of distribution-free tests of randomness against normal alternatives. Journal of the American Statistical Association, 49, 147-157.

_____ (1954b). The asymptotic relative efficiencies of tests and the derivatives of their power functions. Skandinavisk Aktaurietidskrift, 37, 163-169.

_____ (1957). The measurement of estimation and test efficiency. Bulletin de L'Institut International de Statstique, 36, 70-86.

Student (1908). The probable error of a mean. Biometrika, 6, 1.

Sundrum, R. M. (1954). On the relation between estimating efficiency and the power of tests. Biometrika, 41, 542-544.

Teichroew, D. (1954). A table giving a probability associated with order statistics in samples from two normal populations which have the same variance but different means. Technical Report, U. S. Department of Commerce, National Bureau of Standards.

Toothaker, L. E. (1969). An Empirical Investigation of the Permutation T-Test As Compared to Student's T-Test and the Mann-Whitney U-Test. Unpublished Doctoral dissertation, The University of Wisconsin.

Tsao, C. K. (1957). Approximations to the power of rank tests. Annals of Mathematical Statistics, 28, 159-172.

Tsutakawa, R. K. (1968). An example of large discrepancy between measures of asymptotic efficiency of tests. Annals of Mathematical Statistics, 39, 179-182.

van der Laan, P. and Oosterhoff, J. (1965). Monte Carlo estimation of the powers of the distribution-free two-sample tests of Wilcoxon, van der Waerden and Terry and comparison of these powers. Statistica Neerlandica, 19, 265-275.

van der Vaart, H. R. (1950). Some remarks on the power function of Wilcoxon's test for the problem of two samples I, II. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 53, 494-520 (also Indagationes Mathematicae, 12, 146-172).

_____ (1953). An investigation on the power function of Wilcoxon's two sample test if the underlying distributions are not normal. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 56, 438-448 (also Indagationes Mathematicae, 15, 438-448).

van der Waerden, B. L. (1952). Order tests for the Two-sample Problem and Their Power. Proceedings Koninklijke Nederlandse Akademic van Wetenschappen, Series A, 55, 453-458 (also Indagationes Mathematicae, 14, 453, 458).

van der Waerden, B. L. (1953a). Testing a distribution function. Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 56, 201-207 (also Indagationes Mathematicae, 15, 201-207).

_____ (1953b). Order tests for the two-sample problem (second and third communications). Proceedings Koninklijke Nederlandse Akademie van Wetenschappen, Series A, 56, 303-316 (also Indagationes Mathematicae, 15, 303-316).

Walker, H. M. and Lev, J. (1953). Statistical Inference. New York: Holt, Rinehart and Winston, Inc.

Walsh, J. E. (1946). On the power function of the sign test for slippage of means. Annals of Mathematical Statistics, 17, 358-362.

_____ (1949). Some significance tests for the median which are valid under very general conditions. Annals of Mathematical Statistics, 20, 64-81.

Wetherill, G. D. (1960). The Wilcoxon test and non-null hypothesis. Journal of the Royal Statistical Society, 22, 402-418.

Whitney, D. R. (1948). A Comparison of the Power of Non-Parametric Tests and Tests Based on the Normal Distribution Under Non-Normal Alternatives. Unpublished Doctoral dissertation, Ohio State University.

Wilcoxon, F. (1945). Individual comparison by ranking methods. Biometrics Bulletin, 1, 80-83.

Witting, H. (1960). A generalized Pitman efficiency for nonparametric tests. Annals of Mathematical Statistics, 31, 405-414.

Appendix A

Table of Critical Values of D in the Kolmogorov-Smirnov Test for Two Samples of Equal Size*

N	ONE-TAIL TEST		TWO-TAIL TEST	
	$\alpha = .05$	$\alpha = .01$	$\alpha = .05$	$\alpha = .01$
5	4	5	5	5
6	5	6	5	6
7	5	6	6	6
8	5	6	6	7
9	6	7	6	7
10	6	7	7	8
11	6	8	7	8
12	6	8	7	8
13	7	8	7	9
14	7	8	8	9
15	7	9	8	9
16	7	9	8	10
17	8	9	8	10
18	8	10	9	10
19	8	10	9	10
20	8	10	9	11
21	8	10	9	11
22	9	11	9	11
23	9	11	10	11
24	9	11	10	12
25	9	11	10	12
26	9	11	10	12
27	9	12	10	12
28	10	12	11	13
29	10	12	11	13
30	10	12	11	13
31	10	12	11	13
32	10	13	11	13
33	10	13	12	14
34	11	13	12	14
35	11	13	12	14
36	11	13	12	14
37	11	14	12	14
38	11	14	12	15
39	11	14	12	15
40	11	14	13	15

*Adapted from Massey (1951a) and Birnbaum and Hall (1960).

Appendix B

Empirical Probability of a Type I Error for the Kolmogorov-Smirnov Test and the t-test for Various Sample Sizes Using n Class Intervals

Sample Size m = n	α = .05		α = .01	
	K-S test	t-test	K-S test	t-test
6	.0313	.0585	.0071	.0130
8	.0313	.0500	..0028	.0090
10	.0262	.0530	.0054	.0115
12	.0316	.0480	.0054	.0100
14	.0374	.0525	.0013	.0080
16	.0271	.0495	.0068	.0120
18	.0337	.0510	.0078	.0090
20	.0363	.0515	.0074	.0100
30	.0260	.0480	.0010	.0060
40	.0270	.0440	.0030	.0070
50	.0240	.0440	.0010	.0030

Note: Probabilities for samples 6 through 20 were based on 2,000 test samples.

Probabilities for samples 30 through 50 were based on 1,000 test samples.

The significance level of the nonparametric test was randomized for samples 6 through 20.

The chi-square approximation for the Kolmogorov-Smirnov test was used for sample size 50.

VITA

Travis Hillman Willis was born on November 9, 1940, in Ft. Worth, Texas. He was raised in Baton Rouge, Louisiana, where he completed his elementary and secondary education at the University Laboratory School on the campus of Louisiana State University. After graduating from high school in 1958, he entered Louisiana State University and graduated with a B.S. degree in general business in 1962. After working for a year in Sear's executive training program, he entered the U. S. Coast Guard Officer Candidate School in Yorktown, Virginia, in September, 1963. He was commissioned Ensign in January, 1964, and after completing a service school in dangerous cargo and port security, he was initially assigned as supply officer at the New Orleans Captain of the Port. While stationed in New Orleans, he attended night school at Tulane University. In 1965 he was transferred to Marathon, Florida, where he was assigned to duty as commanding officer of Coast Guard Station Marathon. Three years active duty was culminated as administrative officer at the Coast Guard Base Miami Beach, Florida. Mr. Willis was recently selected for promotion to the rank of Lt. Commander in the ready reserve. In January, 1967, he entered graduate school at Memphis State University and graduated with an M.B.A. degree in management in 1968. Then he entered graduate school at Louisiana State University and is presently a candidate for the Ph.D. degree in quantitative methods. At L.S.U. he held a teaching assistantship and was awarded an L.S.U. dissertation-year fellowship. He is presently an Assistant Professor in the Department of Management at Memphis

State University. Mr. Willis is married to the former Ilona Eaton and has two daughters, Ann Hillary and Kristin Cara.

EXAMINATION AND THESIS REPORT

Candidate: Travis Hillman Willis

Major Field: Quantitative Methods

Title of Thesis: A Simulation Study of the Power Efficiency of Certain Nonparametric Statistical Tests for Normal Alternatives

Approved:

Vincent E. Carpenter

Major Professor and Chairman

Max Goodrich

Dean of the Graduate School

EXAMINING COMMITTEE:

James E. Willis

Roger L. Burford

G. Randolph Rice

Eugene C. McCarron

Date of Examination:

July 10, 1972