A Comparison of Independent Samples t-Test, Approximate Randomization Test and Bootstrap Randomization Test to Departures from Population Normality: A Monte Carlo Study

by

Stephen Allen Miles
B.A., Queen's University, 1990

A Thesis Submitted in Partial Fulfilment of the Requirements for the Degree of

MASTER OF ARTS

in the Department of Psychology

We accept this thesis as conforming to the required standard

[redacted]
Dr. Helena Kadlec, Supervisor (Department of Psychology)

[redacted]
Dr. Michael A. Hunter, Departmental Member (Department of Psychology)

[redacted]
Dr. Kimberly Kerns, Departmental Member (Department of Psychology)

[redacted]
Dr. Bill McCarthy, External Examiner (Department of Sociology)

University of Victoria

Supervisor: Dr. Helena Kadlec

Abstract

A Monte Carlo study was undertaken to compare the two independent sample t-test (parametric) with the approximate randomization test and the bootstrap randomization test using data sampled from the Micceri distributions, a set of non-normal distributions. The three tests were compared with respect to Type I errors for varying sample sizes. Following the method of Sawilowsky and Blair (1992), Monte Carlo methods were used to sample the eight Micceri distributions with sample sizes of (5, 15), (7, 7), (10, 10), (10, 30), (20, 20), (20, 60), (60, 60). Observations were sampled independently with replacement from each of the eight distributions.

Replicating Sawilowsky and Blair (1992), the t-test showed marked non-robustness when matched with many of Micceri's distributions and the departures in Type I error from the five percent pre-set value of the study were greater than those found in previous studies. These results suggest that the probability estimates associated with the t-test are unreliable.

The present study extends these results to the alternative nonparametric methods of the bootstrap and approximate randomization tests as follows: When the distribution is extremely asymmetric, the bootstrap method outperforms both the t-test and approximate randomization test. Also, for non-symmetric distributions, pooled sample sizes of at least 30 to 40 would appear necessary to overcome the tendency of the t-test to underestimate the preset value when compared to the bootstrap test. For symmetric distributions (skew of an absolute value less than one), the t-test meets the criterion (i.e. it is within 10% of the preset value) for pooled sample sizes of 20 or more. The t-test lacks robustness for small (less than

20) unequal sample sizes from non-symmetric distributions and also is not robust for the very small and equal samples (7, 7) even for symmetric distributions. The approximate randomization test is consistently conservative for small sample sizes or unequal sample sizes. Whereas for the large sample sizes (60 ,60), this test produces a Type I error 4.6% of the time versus the 5% present value. The approximate randomization test never over estimates the preset value under the criterion.

Prior to further research, I recommend that anyone using small sample sizes (less than 20 subjects in either group) consider using both parametric and nonparametric tests. A frequency plot of the sample data should be examined for breaches of the symmetry assumption of the normal distribution. If such a breach does occur then nonparametric tests should be applied.

Examiners:

[redacted]

Dr. Helena Kadlec, Supervisor (Department of Psychology)

[redacted]

Dr. Michael A. Hunter, Departmental Member (Department of Psychology)

[redacted]

Dr. Kimberly Kerns, Departmental Member (Department of Psychology)

[redacted]

Dr. Bill McCarthy, External Examiner (Department of Sociology)

Table of Contents

List of Figures

INTRODUCTION

In the behavioural sciences, two philosophical orientations to hypothesis testing were developed in the early twentieth century: one by Sir Ronald Fisher and the other by Jerzy Neyman and Egon Pearson. Fisher argued that the focus should be on the data at hand when testing hypotheses; according to his philosophy, a real experiment is not subject to an infinite number of replications. Conversely, the Neyman-Pearson school contended that an experimental result obtained from a given sample could be compared to all possible relevant, albeit hypothetical outcomes. This latter approach, often referred to as classical inference, dominates modern day statistical testing (Camilli, 1990).

At the time of these developments, only the Neyman-Pearson model proved to be a viable option since it is conceptually and computationally simple, requiring only a pencil and a set of normal distribution tables; whereas the Fisherian approach (known today as the randomization model) which requires extensive computing power. Not surprisingly the latter approach was not embraced. Even Fisher himself used the Neyman-Pearson approach, as it represented the best approximation of his then unattainable randomization model. Thus, assumption of normality required by the Neyman-Pearson approach was originally adopted by psychologists and became engrained over time; few people examined whether the normality assumption was

actually being substantiated in real world data until Micceri (1989) examined it in detail.

Micceri's Distributions

Micceri (1989) obtained availability samples from publishers and users of psychological tests that represented real world populations. These data measured people's performance on various achievement and psychometric tests; measures many researchers assume are normally distributed . In fact, Pagano (1993) states that some researchers go as far as to say that a bell shaped distribution is guaranteed with these types of measures. However, depending on the type of violation of the underlying assumption of normality, the parametric statistical tests can show varying degrees of nonrobustness. Thus, there should be an awareness of the types of distributions associated with these measures.

Compounding the normality assumption problem is the complication resulting from statistical analyses performed without first examining the descriptive statistics and/or frequency plots. Micceri (1989) plotted and examined real psychological data sets, listing a number of conditions contributing to non-normally distributed data: (a) when items of a test correlate positively with each other (average correlations of .40 or greater), as they should to have the measurement method make sense, distributions of test scores are usually much flatter than normal, (b) the existence of sub-populations

within a target population having different abilities or attitudes, (c) ceiling or floor effects, (d) variability in the difficulty of items within a measure, and (e) treatment effects that change the mean, the variability and the shape of the distribution.

Among his 440 samples of achievement and psychometric measures with sample sizes varying from 190 to 10,893, Micceri found all the distributions to be significantly non-normal at an alpha of .01 with the Kolmogorov-Smirnov test of normality. He detailed several characteristics of distributions that depart from normality: (a) tail weights ranging from uniform to double exponential, (b) broad classes of symmetry (symmetric to very asymmetric), (c) varying number of modes, and (d) varying degrees of kurtosis. Eight non-normal distributions will be examined in greater depth here.

Micceri labelled these eight discrete distributions according to the shape and measures used: (1) Discrete mass at zero with gap, (2) Mass with gap, (3) Extreme asymmetry (psychometric measure), (4) Extreme asymmetry (achievement measure), (5) Extreme bimodality, (6) Multimodality and lumpy, (7) Digit preference, (8) Smooth symmetric. The eight frequency distributions are illustrated in Figures 1 to 8 respectively, and a complete listing of their descriptive statistics is shown in Table 1. Micceri's observed distributions provide evidence against the historical assumption of statistics that many underlying distributions for psychological variables are normal.

His own eight distributions were compiled into a Microsoft FORTRAN 5.0 library by Sawilowsky, Blair and Micceri (1990) for future research use.

The computational power which allows these distributions to be available contributes to a reexamination of the need for the normality assumption of the Neyman-Pearson approach. One cluster of new computer-intensive statistical methods is based on resampling theory; it includes exhaustive and approximate permutation tests and randomization test (Edington, 1969; Fisher, 1935; Feinstein, 1973; Kempthorne, 1955; May, Masson & Hunter, 1990), as well as bootstrapping methods (Efron, 1977; Efron & Diaconis, 1983; Mooney & Duval, 1993; Noreen, 1989; Simon & Bruce, 1991).

These methods test statistical hypotheses according to their underlying assumptions: Permutation analysis and bootstrapping methods assume random sampling from a population, whereas randomization tests assume that the subjects are randomly assigned to treatments. None of these nonparametric procedures make distributional assumptions that are required by the parametric, classical inference approach of Neyman-Pearson.

Nonparametric techniques are free from the two drawbacks that have plagued the classical inference approach to statistical testing since its genesis. The first drawback is the assumption that the data should conform to a normal distribution. This

assumption states that random fluctuations, or errors in the experimentally observed values of some quantity, are scattered symmetrically about the true value of the quantity. It is further presumed that the greater the error between the experimental value and the true value, the less likely the experimental value will be observed by chance (Diaconis & Efron, 1989).

The second drawback of the classical approach is the focus on statistical measures whose theoretical properties can be analysed mathematically (e.g., mean, standard deviation, correlation). Historically, arithmetic operations associated with statistical analyses were done by hand, and later by calculator; these analyses could be simplified with tables when their formulas have a concise analytical form. However, many other statistical measures such as the difference between two medians, may be relevant and do not have analytical formulas to describe them (Efron & Tibshirani, 1991).

The advantages of the nonparametric approaches are twofold: First, they are free from the restrictive normality assumption (which is rarely actualized in the behavioural sciences), and they can examine characteristics of the population distribution beyond the mean. The assumptions of the nonparametric method chosen depend not on a particular distribution, but either on random sampling or random assignment (Hays, 1988).

Although mathematical statisticians are aware that resampling models and normal curve models reflect different underlying philosophies, these philosophies are rarely conveyed in mainstream psychology. As well, textbooks and journal articles typically focus on the normal model (e.g., Hays, 1988; Kirk, 1982) and this model is taught extensively in mainstream psychology. Psychologists usually choose a nonparametric over a parametric test only when there is a need to overcome a *glaring* assumption violation and not from a philosophical realization (Camilli, 1990). Only recently have these trends begun to change. A number of researchers have published textbooks (e.g., Edington, 1969a, 1987; May, Masson & Hunter, 1990; Siegal & Castell, 1988) and journal articles (e.g., Blair & Higgins, 1985; Hunter & May, 1993; Onghena, 1994) specifically addressing the usefulness of nonparametric techniques, and classic articles on the subject are being revisited to reevaluate their relevance (e.g., Bradley, 1966; Fisher, 1935; Feinstein, 1973; Kempthorne, 1955, 1979; Pitman, 1937a, 1937b).

Intention of this Study

The purpose of this thesis is to discuss and compare nonparametric and parametric methods by examining their respective hypotheses and assumptions, as well as the validity of the conclusions drawn from them. A Monte Carlo study is undertaken to compare the two independent sample **t**-test (parametric) with the approximate

randomization test and the bootstrap randomization tests for testing two means. These tests use data sampled from the Micceri distributions. The three tests will be compared with respect to the probability of Type I errors at varying sample sizes. It is predicted that nonparametric techniques will be superior to their parametric counterparts, as non-parametric tests are often more accurate with nonnormal distributions. To begin, an overview of parametric and nonparametric philosophies is presented.

Parametric Normal Model

To review, the Neyman-Pearson or normal curve model assumes that: (a) all observations are randomly and independently sampled from their parent populations; (b) the population distributions from which samples are selected are normal; (c) all populations have the same variance; and (d) the data are measured on at least an interval scale (Hunter & May, 1993). These assumptions are discussed further in the next section.

The null hypothesis for two sample independent **t**-test (or **z**-test) is most often stated in terms of population parameters. The null hypothesis for this t-test is:

$$H_0: \mu_1 = \mu_2$$

The mean of population one equals the mean of population two. A probability of rejecting the null hypothesis is obtained by comparing the actual difference between

two population sample means with the set of all possible differences between two sample means when the null hypothesis is true. This can be conceptualized as an infinite set of replications that differ only with respect to the random selection of sample observations (Camilli, 1990). The underlying distribution may be unknown but the distribution of the sample mean can be estimated as a normal distribution if the conditions of the Central Limit Theorem are fulfilled. The theorem states that if random samples of equal size N are drawn from a population with a mean μ and a standard deviation σ, the sampling distribution of the sample mean will be approximately normally distributed when N is *large*. This approximation holds regardless of the population distribution and becomes more accurate as N increases (May, Masson & Hunter, 1990). For the two sample t-test, both samples, or the difference of the means must meet these requirements. It is a commonly accepted in psychology that in order to meet this criterion, the samples must be at least size 20. Under these conditions, one can find the probability of observing an experimental outcome under the null distribution by referring either to a table of **t** values as provided in most introductory textbooks or to a computer generated printout.

Although there is some evidence to suggest that the **t**-Test is robust to violations of the normality assumptions (Boneau,1960) it is often not robust to the extent conveyed in many textbooks. For example, Boneau (1960) states " the t-test is

functionally a distribution free test" (p. 60), and Box (1953) suggests that all parametric tests show a remarkable property of robustness to non-normality. This assessment persists in Hays' (1988) statement that, "the assumption of normality can be violated almost with impunity" (p. 303). With such statements being conveyed in the mainstream literature and by influential applied statisticians it follows that researchers who are not statistically expert feel little discomfort when their samples do not meet the normality assumption underlying their parametric test statistic. In other words, researchers are not inclined to examine the descriptive statistics if the **t**-test is viewed as a *distribution free test.* In effect, researchers analyze their data using high powered inferential statistical tests of parameters before they have examined the descriptive statistics (e.g., frequency distributions, variances, and skew) that carry the most pertinent information regarding the normality assumption. The popular belief that there is only one approach (e.g., parametric) for testing statistical hypotheses and that it is extremely robust under all conditions, has created a false sense of security.

Bradley (1978) thoroughly examined the robustness of normal curve procedures when faced with assumption violations . He found evidence of robustness but only under highly qualified conditions and involving the interplay of several factors: (1) the size of alpha, (as alpha decreased so did robustness); (2) location of the region of rejection (a two tailed, versus the more robust one tailed test); (3) for the smallest

sample, the size of the sample and the shape of the parent population; (4) for the larger samples, the absolute and relative size of the samples and the sizes, shapes, and variances of their parent populations. Bradley (1978) found that all of these factors interact at the highest level and can affect robustness depending on the characteristics of the data set.

Assumptions of the Normal Curve Model

Random sampling

Random sampling represents one of the most fundamental assumptions made in normal curve statistics; yet it is one of the most difficult to achieve in psychological research. The hypotheses tested in the parametric model are established in terms of population parameters. In order to obtain valid estimates of these parameters, random sampling *must* have occurred. If random sampling does not occur, there is no linkage between the population and the sample, therefore estimates obtained from the sample may be unreliable. Further, lack of random sampling could potentially introduce bias into the tests by affecting independence among the sample scores (Hays, 1988).

Hays (1988) writes that "some probability structure underlying the sample is a little price tag [to pay for] statistical inference.....Unless this assumption is at least reasonable, the probability results of inferential methods mean very little" (p. 54). He claims that researchers who knowingly use these normal curve procedures when they

have not randomly sampled their data are merely "*prettifying*" their study and adding little or nothing to its meaning.

In most experimental research in psychology, random sampling is all but impossible even with an unlimited budget. Researchers rarely generate exhaustive lists of their population(s) and then generate random numbers in order to select people. Moreover, it is unlikely that all the chosen people would agree to participate. Samples commonly used are convenience samples from introductory psychology courses or through some other means of recruitment, typically in a university setting.

Normally distributed data

Parametric tests and particularly t-tests, assume either of the following: the sample mean has a normal distribution when the sample data are normally distributed and the population variance is known; or, the sample mean has a t-distribution when the sample data are normally distributed and the population variance is unknown. However, if the sample data are not normally distributed then according to the Central Limit Theorem, random sampling from a non-normal population will still produce distributions for the sample mean that are approximately normal if the sample size is larger.

The problem that psychologists and others in the behavioural sciences need to face is twofold: First, samples are typically not random and not "large"; thus they can

not depend on the full benefit of the Central Limit Theorem. Second, samples are drawn from a population that may be extremely non-normal and prone to extreme values - a problem noted by Micceri (1989). Together, these can lead to questionable statistical outcomes in some instances.

Homogeneity of variance

Parametric tests of difference between means assume equal variances in the populations being compared; therefore, when comparing two groups, the variances in the two samples must be equal in order for the **t**-test to be valid (Hays, 1988). Statistical programs like SPSS adjust the degrees of freedom to account for unequal variances but there is no concise theoretical basis for this adjustment. However, in real world data that may be distributed in a manner not consistent with the normal distribution, the group with the largest mean is also likely to have the largest variance (Micceri, 1989). Bradley (1978) found that even the robust t-statistic can be affected when faced with this assumption violation, and here again the results obtained may be unreliable.

Random assignment for independent samples t-test

This assumption underlies almost all experimental research and should be the easiest to achieve, regardless of the design or sample. It asserts that subjects have an equal and independent probability of being assigned to each condition in the

experiment. This allows the researcher to eliminate any potential for bias and to ensure independence of observations. Further, random assignment of subjects to conditions allows the researcher to make causal inferences because it virtually ensures that no factor other than the differences in experimental conditions could have caused scores to vary across conditions (May, Masson, & Hunter, 1990). Although these conditions are frequently realized in experimental psychological research, they are often impossible to meet with the attribute variables of applied psychology, since it is not possible to randomly assign people to head injured versus non-injured groups, or to gender-specific groups.

Scale of measurement

Most parametric tests assume that the dependent variables have been measured on at least an interval scale and preferably on a ratio scale (May, Masson, & Hunter, 1990). The z or t test require interval scale as a minimum. Many types of experiments conducted in psychology, however, have variables (e.g., preference) that are measured on ordinal scales which are then transformed into an interval or ratio scale of measurement through manipulations or assumptions. This poses problems when making an inference from the original data. When interpreting the effect the researcher must take into account the manipulation of the original variables measured.

Non-Parametric Models

Apart from random assignment, it is often very difficult to meet many of the conditions underlying the assumptions of normal curve statistics; consequently tests may not be as robust as many researchers believe (Bradley, 1978). Methods for testing statistical hypotheses that are not encumbered by underlying distributional assumptions are clearly needed. Fortunately such methods exist although they have become practical only in the last fifteen years with the availability of inexpensive computer power. For instance, the randomization model (Fisher, 1935; Pitman, 1937a, 1937b; Kempthorne 1955, 1979; Edington, 1966, 1986) includes approximate and exhaustive randomization tests as well as permutations tests. Another type of nonparametric model is bootstrapping. These nonparametric tests are differentiated by their assumptions and uses.

Randomization Tests

Fisher (1935) did not accept the Neyman-Pearson philosophy which compared actual results to theoretical outcomes (Camilli, 1990). He argued that in reality, an experiment can not be subject to an infinite number of identical replications. If performed, replications would certainly not be identical because features of the design and the sample change over time. Fisher's approach was to ignore the data that had not been observed, or more specifically, to restrict the sample space to the data at hand (Camilli, 1990). Thus Fisher advocated the randomization model as it focused on the

available data. He permuted this data into a referent distribution that was based on only those outcomes that have a possibility of being observed. He first suggested tests based on permutations of observations as an alternative to the Neyman-Pearson approach. Fisher's approach was in sharp contrast to the accepted practice at the time (Bradbury, 1987). Since then, Pitman (1937a, 1937b), Kempthorne (1955), Edington (1969a), and others have devoted attention to the model, developing it further. But like Fisher, they were confronted with the lack of computing power to perform their analyses.

Randomization tests do not directly test hypotheses about the values of population parameters. Experimental psychologists are most often interested in demonstrating a relationship between a predictor and a criterion or the effects of a specific treatment on a dependent variable; thus, they are often not interested in estimating population parameters. Based on the assumption of random assignment, the null hypothesis of the randomization method expresses these relationships in terms of the dissimilarities between samples, such as the presence of different means.

This could be partially stated as:

$$H_o: \bar{x}_1 - \bar{x}_2 = 0$$

The difference between two sample means is equal to zero. Other possible sample distribution differences should also be considered like the difference between group medians (May & Hunter, 1993). A normal curve test, e.g., a t-test, that has not met the assumption of random sampling would test the partial hypothesis of equal sample means only as there is no linkage between the sample and the population.

The number of people promoting the use of resampling techniques in the behavioural sciences has increased with the recent accessibility of computer technology (e.g., Edgington, 1986; Siegal & Castellan, 1988; May, Masson & Hunter, 1990; Hunter & May, 1993; Noreen, 1993; Onghena, 1994), although this group is still relatively small compared with the normal curve model advocates. This latter group accepts the historical necessity of the normal curve model despite the new research on methods. Users of statistics are also resistant to change because of a reliance on statistical packages. However, as computer technology continues to develop and enhanced randomization techniques emerge, advocates of randomization methods may become more mainstream. With repeated exposure, users of statistics may become more discerning of the limitations and advantages of both methods; thus their choice of

statistical test will be based on the data and the theory they wish to assess.

A randomization test for two samples pools the samples and then permutes the pooled sample to obtain all possible two sample outcomes. A test statistic such as the difference between sample medians or perhaps even sample means is computed for each shuffle. The test statistic's frequency distribution over all possible arrangements serves as the empirical null distribution or randomization distribution. As with the normal curve model, the probability of the obtained (or observed) sample statistic when the null hypothesis is true is found by adding together the probability of outcomes equal to, or more extreme than the obtained outcome. The number of permutations that meet the extreme criteria are then divided by the number of possible permutations, establishing a probability value (p-value). This approach contrasts the purely theoretical calculations of the normal distribution which is an integral of a density function. This p-value is then compared to a preset alpha level typically .01 or .05. If the obtained p value is less than the alpha value the researcher can conclude that there is an effect of manipulation (Siegal & Castellan, 1988).

When the sample size is increased, an approximate randomization test is used because permuting the data would simply take too long (with increasing sample size the number of permutations can soar to over one hundred million). These approximate tests have proved to be good estimations of the exhaustive tests (May, Masson &

Hunter, 1990).

Bootstrapping Tests

The bootstrap test is another computer-intensive nonparametric technique. It was first developed by Efron (1977) when computing power was becoming affordable. Since then, many variations of the bootstrap test have been and continue to be developed. The current task for some applied researchers is to make probability based inferences about a population of interest based on a sample estimate of that population. Bootstrapping is a nonparametric technique that allows one to make such inferences without strong distributional assumptions. Although it estimates population parameters, this techniques differs from the traditional parametric approaches in that it utilizes large numbers of repetitive computations to estimate the shape of a statistical sampling distribution (Noreen, 1989). This allows a researcher to make inferences about the populations when analytic formulas are unavailable or when the distributional assumptions are untenable.

Bootstrapping, like randomization, uses the data at hand to generate an empirical estimate of the sampling distribution of a statistic. Instead of permuting the data, bootstrapping re-samples the data with replacement thousands of times in order to generate an empirical referent distribution. Researchers typically bootstrap the data approximately 2000 to 5000 iterations to obtain a meaningful referent distribution

(Mooney & Duval, 1993; Noreen, 1993).

Both bootstrapping and parametric techniques have the same underlying purpose: to estimate the sampling distribution of a statistic in order to make inferences about population parameters (Mooney & Duval, 1993). The key difference lies in the way each approach obtains its respective sampling distributions. While the parametric approaches use a priori assumptions about the shape of the sampling distributions, bootstrapping estimates the sampling distribution of a statistic by relying on an analogy between the samples and the populations. For instance, to generate an empirical estimate of the sampling distribution of the sample means, bootstrapping uses the sample data as if it were the population and applies Monte Carlo sampling (Mooney & Duval, 1993). In addition, the bootstrap technique performs a test of the similarities between the two sampling distributions.

Bootstrapping is best used in cases when the sampling distribution is unknown (Noreen, 1989). For example, one might need to examine the difference between two sample medians or a least squares regression coefficient where the residuals are non-normal.

Bootstrap Randomization Test

The bootstrap methods pioneered by Efron (1977) can be modified. One of these hybrids is the Bootstrap Randomization Test (Hinkley, 1988; Noreen, 1989).

This method is a hybrid of the randomization and bootstrap approaches. It can be used to test a hypothesis that the data are random samples from the populations in which the variables are stochastically independent; furthermore, it can be used to determine whether the marginal distributions of the variables in the samples satisfactorily approximate the marginal distributions of the population variables (Noreen, 1989).

As previously suggested, bootstrap randomization is very similar to the approximate randomization test. However, instead of sampling (the sample) without replacement, bootstrap randomization samples (the sample) with replacement to minimize the difference between the actual and the estimated group distribution, it does this over 5000 bootstrap iterations (Hinkley, 1988). For example, two samples are drawn of size n. These are then combined and this combined sample is bootstrapped (sampled with replacement) into two new samples of size n. Now a randomization test is executed on the bootstrapped samples by pooling the samples permuting them to obtain all possible two sample outcomes. A test statistic such as the difference between to means is computed for each shuffle. The test statistic's frequency distribution over all possible arrangements serves as one empirical null distribution to find the probability of the observed sample statistic when the null hypothesis is true. With the bootstrap randomization test this randomization test is repeated for each bootstrapped samples (5000 times).within the original sample. Thus it is computing 5000 randomization tests

within a single sample.

The approximate randomization test and the bootstrap randomization test will usually agree on the criteria for the rejection of their respective null hypotheses, although the hypotheses that each tests are stated somewhat differently. In the case of the approximate randomization test, the null hypothesis is that the variables are independent in the samples. More specifically, it states that all permutations of each variable are equally likely. The null hypothesis for the bootstrap randomization test is that the data are random samples from a population in which the variables are independent and the marginal distributions of the variables in the sample approximate those in the population (Noreen, 1989). The bootstrap randomization test is most advantageous when the researcher wants to make a direct inference about a characteristic of a population (i.e. that the variables are independent in the population).

With the bootstrap randomization test, as with the normal curve model, there is no guarantee that the sample satisfactorily approximates the population. Nevertheless, with the bootstrap randomization test, there is some relationship other than a theoretical assumption, between the sample and the population. This technique is in its developmental stages and more research is needed to validate its findings. Noreen (1989) suggests running an approximate randomization test whenever the bootstrap randomization test is performed. He found little difference with respect to their

rejection of the null hypothesis. However, he used data based on mathematically generated distributions that may or may not be representative of real world data as are Micceri's distributions (1989).

Assumptions of the Non-Parametric Models

The advantage of these methods is that they make *no* distributional assumptions and in particular, they do not require the assumption of normality. For the two methods examined here, the assumptions are: first, randomization tests (both exhaustive and approximate) require the experimenter to randomly assign subjects to conditions; second, depending on the choice of randomization test, some require the data to be measured on an interval or ratio scale. Finally, bootstrapping methods, because they estimate population parameters, require random sampling from a population of interest (Noreen, 1989).

Problems of Result Interpretation

Many researchers view randomization tests as restrictive because only inferences of cause and effect can be made and there is no assurance of generalization from their sample to a referent population (Bradbury, 1987). This restriction is significant in psychology, although sampling theory was originally developed for agricultural research (Mook, 1983). In the field of agriculture, researchers experiment with new combinations in their samples in order to predict that a particular mixture in

a specific environment with a certain crop will increase the average yield in a similar real life situation. Like agricultural research, survey studies in psychology were first developed from the parametric model with the goal of generalizing to a population. Thus the preoccupation with generality appears to be a convention grown out of an agricultural context and is often not related to the type of research done in psychology today (Mook, 1983).

The focus in psychology is the need to reflect a substantive theory that explains differences between two or more groups rather than between two or more population parameters. If we are testing an hypothesis that reflects group differences and not parameter estimates then why are parametric methods used? If external validity is required of psychological research it can be attained through replication in different settings rather than with parametric methods. Indeed, regardless of the sample size or its randomness, researchers should avoid making sweeping generalizations based on a single study that is tested with parametric methods.

Comparison of Parametric and Non-Parametric Tests

Two criteria are commonly cited as a basis for arguments against the use of

non-parametric tests: Low power-efficiency and non-robustness. These arguments will be examined in the following sections.

Power-Efficiency

The power of a statistical test is defined as the probability of correctly rejecting a null hypothesis. Power-efficiency is determined by the ratio of the sample sizes from two statistical tests designed to achieve the same power. For example, compare the independent samples **t**-test to its non-parametric counterpart, the Mann-Whitney U-test. If the U-test needs 30 subjects to achieve a power of .90 while the **t**-test requires 35 subjects, the power efficiency of the **t**-test compared to the U test will be 30/35 or 85%. Generally speaking, the more subjects that are required to obtain the same power as another test, the lower the former's power-efficiency (May, Masson & Hunter, 1990).

Noreen (1989) found that randomization tests produce the same results as their parametric counterparts when the normal assumptions are met, that is they had the same power-efficiency. He constructed data sets that met all the assumption criteria for a parametric test (**t**-Test) and concluded that the approximate randomization test performance could not be distinguished from that of the **t**-Test.

The power of the **t**-test and Mann Whitney U-test were compared by Sawilowsky and Blair (1992) using data from Micceri's extreme asymmetry

(psychometric) distribution. Using a Monte Carlo simulation and the extremely asymmetric (psychometric) distribution, they found that for an effect size of .5 and an alpha of .05, the U test rejected H_0 at a rate of .723; whereas the **t**-test rejected it at a rate of .495. This example indicates that nonparametric tests can have a power-efficiency greater than 100% when the data are non-normal and contain outliers.

Robustness

Bradley (1978) and Micceri (1989) point out that the majority of the literature on robustness is devoted to mathematically derived distributions, not empirical ones. In fact, over all the years that statistics have been used only a few researchers have examined robustness in real world distributions (e.g. Stigler, 1977; Hill & Dixon, 1982; Sawilowsky & Blair, 1992). In these real world situations, the statistics displayed radically different properties as compared to simulated environments. Unequal sample sizes, small and large sample sizes and symmetry violations are effects that influence the robustness of the test. For example Sawilowsky and Blair (1992) found both the independent and dependent **t**-test to be nonrobust to Type I error when confronted with data that was non-normal. The degree of nonrobustness evidenced in their study was, at times, more severe than had been previously reported by researchers utilizing mathematically generated distributions. However, the **t**-test was found to be robust under the following circumstances: (a) the sample sizes were equal, (b) sample

sizes were fairly large, and (c) the tests were two-tailed. In psychological research, particularly applied psychology, these conditions are sometimes difficult to meet.

Availability of Parametric and Non-Parametric Models

The parametric model is limited by its underlying assumptions. Although some of these procedures will provide reliable estimates even when assumptions are violated (e.g., normality, and random sampling, provided the sample size is large 20>) , there is still cause for concern when a breach of assumptions, based only on experiential evidence without a consistent theory on the effects, is assumed to be harmless of such a breach.

The nonparametric model also suffers from limitations; although its strongest impediment is its relative obscurity. Most introductory statistics classes include only the parametric model (e.g. May & Hunter, 1993). Although recently there have been some statistical textbooks that address nonparametric tests (Edington, 1987; Seigal & Castellan, 1988; May, Masson & Hunter, 1990), this trend is still in its infancy.

Another factor limiting the use of nonparametric statistics is their unavailability among software packages commonly used for significance testing. The majority of researchers in psychology use packages such as SPSS, SYSTAT, SAS, and BMDP, of these only SYSTAT 7.0 for windows has bootstrapping and permutation tests

available. As each new version is released more and more nonparametric tests are being incorporated into the packages, thus making them accessible to all researchers. Prior to the release of SYSTAT 7.0 for windows their availability was certainly not mainstream, they included programs that were mainly written for pedagogical purposes like: NPSTAT (May, Masson & Hunter, 1989) STATXACT (Mehta & Patel, 1991), CANOCO (ter Braak, 1988) and RESAMPLING STATS (Simon & Bruce, 1991).

Significance tests using bootstrapping are so new that little research has been done to evaluate their performance. In research using known outcomes, Noreen (1989) found bootstrapping tests to be unreliable with small sample sizes. These limitations are associated with the extension of the hypothesis to distribution comparisons included in the bootstrap randomization method; they should begin to disappear as more theoretical evaluations occur. As psychologists familiarize themselves with available software packages for performing nonparametric tests on non-normal data they should test more appropriate hypothesis rather than convenient or traditional versions.

Empirical Investigation

Sawilowsky and Blair (1992) were the first to sample the Micceri distributions and to examine the Type I error properties of the independent samples **t**-test when departures from population normality occurred. Among the samples from the eight

distributions, they found that the probability estimates were either liberal or conservative (with rejection rates near alpha occurring in only about half of the instances). Provided the sample sizes were equal and large, only the following distributions produced probability estimates that would be close to those obtained under normal curve theories: (a) Smooth symmetric, (b) Extreme bimodality, (c) Digit Preference (d) Multimodal lumpy.

Present Research

An investigation of classical and nonparametric methods on real world data is clearly overdue. In this study, I analyze samples from Micceri's distributions using an Approximate Randomization Test, Independent Samples **t**-Test and a Bootstrap Randomization Test. I compare these three tests using eight Micceri distributions with various sample sizes in order to verify whether a sample size of "at least 20" is required for the Central Limit Theorem.

The following three results are expected: First, with respect to the estimates of the probability values for Type I error, the bootstrap randomization test should outperform both the randomization test and **t**-test where the two population means are the same. The bootstrap algorithm uses sampling with a replacement and, therefore, should outperform the randomization test that does not include this feature. The bootstrap algorithm should perform especially well for small sample sizes since

sampling with replacement would allow the occurrence of highly skewed distributions as so often happens in psychology.

Second, the bootstrap test should perform with increasing accuracy of meeting the nominal Type I error rates as the sample sizes expand to better reflect the asymmetries of the original distributions. When the distributions are extremely asymmetric the bootstrap and randomization tests should outperform the t test with respect to Type I error probability estimates.

Third, the independent samples **t**-test should meet preset Type I error rates when sample sizes are equal and large (30+) and when a two tailed test is used (Sawilowsky and Blair, 1992). Yet, regardless of the robustness of the **t**-test's performance in specific cases, it is still in violation of many of its assumptions. Until all assumption violations are known, the results may be unreliable. The probability estimates obtained by **t**-Test may be especially unreliable when the Micceri distributions are involved.

METHOD

Apparatus and Software

The computer program for this simulation was written using MATLAB 4.0 for windows and Microsoft C++ by Tom Allen. A Microsoft FORTRAN program supplied by Sawilowsky, Blair and Micceri (1990) was used to sample with replacement from

each of the eight Micceri distributions. The software was executed on a 486-66 MHz personal computer.

Procedure

As a preliminary software test, the three methods (independent two sample t-test, randomization test, and bootstrap randomization test), were compared using three conditions with samples of size (n_1 , n_2) = (10, 10), (10, 15) and (20, 20) sampled from normal distributions. In each instance, the calculated alpha value was exactly the nominal alpha value of .05.

Following the method of Sawilowsky and Blair (1992), Monte Carlo sampling was used to obtain the 10, 000 samples from each of the eight distributions characterized by Micceri (1989) as being representative of the types of data found in his study. For each sample, observations were sampled independently with replacement from each of the eight distributions.

The study proceeded as follows: Samples of size (n_1 , n_2) = (5, 15), (7, 7), (10, 10), (10, 30), (20, 20), (20, 60), (60, 60) were generated. The independent samples **t**-test was computed on each sample using an algorithm from NPSTAT (May, Masson & Hunter, 1993). An approximate randomization test (algorithm from NPSTAT) and bootstrap randomization test (algorithm by Noreen, 1989) were performed on the samples. The approximate randomization test was set to 5000 permutations (previous

research has found that with only 2000 permutations, the approximate randomization test will produce a probability value that approximates the exhaustive test within several decimal places) (May, Masson & Hunter, 1990). The bootstrap randomization test was slightly more complicated as it incorporates the randomization and bootstrap into one test. It first bootstraps the sample from the chosen Micceri distribution, by sampling with replacement two groups of size (n_1 , n_2). These two groups are then combined and permuted into two groups of size (n_1 , n_2). For each permutation the difference of the means is calculated and compared to the sample difference, exactly the same procedure as the randomization test. The only difference is instead of just permuting two groups (n_1 , n_2) it first samples the sample with replacement and then permutes the groups. The notion is that sampling with replacement will emulate the population more effectively over 5000 iterations than simply taking one sample and computing a test on it.

The program kept track of the outcomes with probability values at 0.05 (the nominal alpha value) or less (i.e., rejected H_o at alpha of .05). The sum of these was divided by 10,000 repetitions to give an obtained alpha value for each method and for each type of distribution.

RESULTS

The probability value calculated in the Monte Carlo simulations is the

proportion of Type I errors over 10,000 repetitions. The preset value, or nominal alpha of .05 is the assumed probability of a Type I error; this represents the probability that the null hypothesis, H_o, is rejected, given that H_o is true. This value is used for rejecting the null hypothesis in each repetition. The calculated probability obtained for each test is the actual probability of rejecting the null hypothesis over repeated trials; that is, the proportion out of 10,000 repetitions. In order for a researcher to be confident in making inferences from the results of a single experiment or repetition, the actual and the a priori or preset alpha values should be equal. Since there is no precedent the criterion of equality is interpreted as plus or minus .003, therefore values of .047 to .053 are acceptable.

Results will be discussed for each Micceri distribution from which data were sampled.

Smooth Symmetric (Figure 1)

The Micceri distribution closely approximates the normal distribution although there are slight discrepancies within the body, and in contradiction to its name, its tails are slightly asymmetric. This close approximation to the normal is evidenced in the skew of only 0.01.

As illustrated in Figure 9, the randomization test is always more conservative than the other tests; that is, it underestimates the nominal probability value of .05,

except for the samples of (10, 30) were it is slightly more liberal. As sample size increases, the randomization test approaches the nominal value. The bootstrap randomization test and the **t**-test have similar patterns to each other; thus meeting the criterion of equality except at the smallest sample size of (7, 7).

Discrete Mass at Zero (Figure 2)

Again the Micceri distribution closely resembles the normal distribution although there are variations in the tails. The skew also is close to normal at -0.03.

The results graphed in Figure 10 indicate that the randomization test is the most conservative. The bootstrap is slightly liberal while the t-test is slightly moderate in meeting the criterion of equality at (5,15). The bootstrap is equivalent to the **t**-test at meeting the criterion except at (10,30) where the t-test is liberal while the bootstrap meets the equality criterion.

Discrete Mass at Zero with Gap (Figure 3)

This Micceri distribution is extremely non-normal with a skew of 1.65. This is representative of what might be expected, for example, in a drug trial study.

The results for the discrete mass at zero with gap are shown in Figure 11. The three tests demonstrate different behaviours for samples from this distribution. The

randomization test is extremely conservative (α less than .025) for a pooled sample size of 40 or less. Although less so than the randomization test, the **t**-test is still extremely conservative when either one of the sample sizes are less than 20 with the obtained probability of less than .04; otherwise, the t-test meets the equality criterion. Conversely, with obtained probability of .0615, the bootstrap is too liberal for the extremely small sample (7, 7) but it meets the equality criterion in each of the other sample sizes.

Extreme Asymmetry (achievement) (Figure 4)

This Micceri distribution is extremely non-normal with a skew of -1.33. As illustrated in Figure 12, the randomization test is again the most conservative but only slightly more than the **t**-test. The bootstrap randomization test meets the criterion of equality for all sample sizes. All three tests meet the criterion for large sample sizes of at least 20 subjects per group.

Extreme Asymmetry (psychometric) (Figure 5)

This Micceri distribution is again extremely nonnormal with a skew of 1.64. As shown in Figure 13, the **t**-test and randomization test are both extremely conservative for small sample sizes, with little discrepancy between their obtained probability values. The **t**-test is more conservative for unequal sample sizes while the randomization test is

more conservative for equal sample sizes. The bootstrap meets the criterion of equality for all conditions. All methods meet the equality criterion for combined sample sizes of 40 or more.

Digit Preference (Figure 6)

This Micceri distribution is slightly skewed at -0.07. The results graphed in Figure 14 show that the randomization test meets the equality criterion for all sample sizes except (7, 7) where it is conservative. The **t**-test is also conservative for sample size (7,7) but less so than the randomization test. The **t**-test and the bootstrap meet criterion of equality for all sample sizes except (10, 30) where they are both liberal.

Extreme Bimodality (Figure 7)

Visually, this Micceri distribution appears very nonnormal; although its skew is only -0.08. The randomization test is extremely conservative only meeting the criterion of equality for the largest sample size (60, 60). The **t**-test meets the criterion for all conditions except (10, 10) where it is too liberal. As shown in Figure 15, the bootstrap test only fails to meet criterion for sample sizes (7, 7) and (10, 10) where it is too liberal.

Multi Modal Lumpy (Figure 8)

This Micceri distribution has multiple modes that make it non-normal; but the skew is only 0.19. The results graphed in Figure 16 indicate that all three tests meet the

equality criterion for samples having either or both sizes equal to or greater than 10. The only deviation is the randomization test which is too conservative at (20, 20). At (5, 15), only the bootstrap test meets the criterion of equality, while the **t**-test is conservative and the randomization test is even more so. At (7, 7) the randomization test is extremely conservative with obtained probability of .0319, the **t**-test meets the criterion of equality with obtained probability of .0521 and the bootstrap is too liberal with obtained probability of .056.

DISCUSSION

The above findings are similar to those found by Sawilowsky and Blair (1992). Generally, the t-test shows marked non-robustness when faced with many of Micceri's distributions; in fact, the departures in our study are slightly greater than those found by Sawilowsky and Blair (1992). The discrepancies in the present study may be the result of the larger number of iterations, leading to greater accuracy. However, Sawilowsky and Blair (1992) also found that the **t**-test is remarkably robust when certain conditions are met: (a) when sample sizes are equal, (b) sample sizes are fairly large (25>) and (c) when the tests are two-tailed. The present study confirms these findings: here the **t**-test is also adversely affected by the non-normal data sets, but especially for unequal sample sizes and small samples. Thus the probability estimates associated with such data may be unreliable.

The two nonparametric techniques (approximate randomization test and bootstrap randomization test) were hypothesized to be superior to the t-test because they make no assumption regarding the shape of the sample distribution. The results indicate that this hypothesis requires further elaboration. The results for the three methods can be summarized as follows:

Bootstrap randomization

For sample sizes of (7, 7), the results for the bootstrap randomization test depend on the shape of the population distribution. If the bell shape of the normal distribution is approximated - in the smooth symmetric, discrete mass at zero and digit preference - the bootstrap method is slightly conservative or meets the equality criterion. Overall it provides results that are consistently close to the preset nominal value of .05 for the Type I error rates.

If the distribution is extreme asymmetric, either positively (psychometric) or negatively (achievement), the bootstrap test meets the criterion of equality and outperforms both the t-test and approximate randomization test. For the most non-normal distribution (the discrete mass at zero with gap), the bootstrap test is closest to the criterion, however the test is still too liberal when small sample sizes are used.

t-Test

The t-test meets the criterion of equality for pooled sample sizes of 20 or more

among symmetric distributions with skew of an absolute value less than one. This appears to be a result of the Central Limit Theorem. For non-symmetric distributions, such as the two extreme asymmetric distributions, combined sample sizes of 30 to 40 are necessary to overcome the conservative outcomes of this test. The **t**-test is overly conservative for the discrete mass at zero with gap distribution (the most non-normal of the distributions) until a sample size of 20 per group is attained. The **t**-test lacks robustness for small (less than 20) unequal sample sizes from non-symmetric distributions as well as for very small samples (7, 7) regardless of the symmetry of the distribution.

Approximate Randomization Test

The approximate randomization test outcomes are the most difficult to categorize. As a test, it is consistently conservative for small sample sizes or unequal sample sizes. Even for the very large sample sizes (60, 60), this test produces a 4.6% probability of Type I error. The approximate randomization test is never liberal against the criterion of equality.

Comparison of the Three Methods

The bootstrap test is the method of choice for small sample sizes because it provides probability estimates that are closest to the nominal alpha level. However, when this test is employed, the researcher must be aware of the hypothesis that is being

tested; that is, the samples must have the same conditional distributions as well as the same sample means. The most thorough procedure for small samples would be to run both the t-test and the bootstrap randomization test. If the results agree, one can be reasonably confident of the reliability of the probability estimates. If the results of the two tests differ, then a combined sample size of at least 20 per group will be required.

Even larger sample sizes should be considered if there is a possibility that the underlying distribution will have a discrete mass at zero with gap. For example, this may occur in tests for a rare dysfunction among the general population or in an experiment involving drug trials. This particular distribution causes a severe lack of robustness in the t-test for samples as large as (10, 30). Extreme care must be exercised when using the t-test as unequal sample sizes of this type are common in situations such as the above mentioned clinical studies. A large sample size would help validate any of the tests, as well it would provide reliable frequency distributions and better estimates of the descriptive statistics. In the rare dysfunction situation, a large sample size would be difficult to obtain; thus a bootstrap randomization test is preferred.

The consistently conservative outcomes of the approximate randomization test indicate that it is the least likely of the three tests to commit a Type I error. The problem with its use is that the preset error rate differs from the actual error rate or p-

value although not in a consistent manner. This makes interpretation of the results difficult. In addition, like the bootstrap randomization test, this randomization test also tests the similarity of the two conditional distributions of the independent samples. Unlike the bootstrap randomization test, however, the approximate randomization test does not have a consistent pattern which could be compared to the t-test. This precludes combining the two as a check on each other.

If dealing with small sample sizes, it is helpful that the results of both the t-test and bootstrap randomization test are reliable. The present research indicates that using the measurement of skewness (establishing whether the samples are distributed in a symmetric fashion) allows a basic rule of thumb. If they are symmetric (with the skewness less than 1.0), then the estimates of probability of Type I error obtained by the t-test and the bootstrap method will agree; if they are very asymmetric then the estimates will diverge and the researcher should be alerted to the necessity of a larger sample. If large samples are not feasible, then a minimum combined sample size of at least twenty should be considered and the bootstrap test should be used. The present research points out the need for symmetry rather than depending on the normal distribution with a skew and kurtosis close to zero. A U-shaped distribution (e.g., the extreme bimodality distribution), is symmetric but in an upside down version of the normal curve and the t-test meets the equality criterion for all sample sizes in this case.

Finally the time required to calculate these tests is now short enough to be practical and becomes shorter with each new computer processor release. The t-test and approximate randomization test are the fastest while the bootstrap randomization test required more time as it was computationally the most intensive test. Today's fast personal computers are capable of processing the bootstrap randomization test and the approximate randomization test without significantly impinging on the researcher's time.

Practical Considerations

Based on the present findings, statistical software manufacturers should make nonparametric tests more widely available. Also, there must be a motivation and awareness of these alternative techniques; therefore, researchers must be aware of the underlying assumptions of all tests and test hypotheses appropriate for the statistical tests in use. Researchers also must satisfy their audience that a test's underlying assumptions hold for their particular data set. When submitting a paper for publication, a first step should be to provide, as part of the results, a frequency plot or a complete list of descriptive statistics including the skew.

Academic psychologists must inform their students of the limitations of parametric statistics as well as the necessity to validate the underlying assumptions rather than simply accepting the assumed robustness as justification for use. As well,

psychologists must decide whether tests of parameters are more important than a comparison of the sample distributions. This judgement is intrinsic to the choice of the statistical procedure; convention is no longer a justification for the emphasis on parametric over nonparametric tests.

Directions for Future Research

Due to their skewness, the Micceri Distributions affect adversely the robustness of the t-test (the greater the skew, the less robust the t-test). To provide a more exhaustive set of skew values and varying degrees of deviation from the normal distribution, further investigations into the effects of skew and non-normality on the t-test should be performed through transformations of these distributions.

Both the bootstrap and randomization methods test the hypothesis that the conditional distribution functions are similar. If two sample distributions are similar then the shape and the descriptive statistics are also similar to the population; if they differ, the failure may be in shape and descriptive statistics. Additional research should examine the effect of differences in skew and variance on these two tests.

The power of the three tests should be compared as a function of the effect size between two samples from the same distribution in order to maintain the same conditional distributions within two samples that give different means. These comparisons would not only examine the robustness of the t-test but provide a basis on

which to evaluate the applicability of the two nonparametric tests as tests of differences of means.

Since the Micceri distributions occur frequently in some areas of psychology, psychologists, rather than mathematical statisticians, should take responsibility for future research in this area. The precedent of examining parameter estimates rather than comparing distributions has led to an over dependence on the normal distribution. Like the belief in the robustness of the t-test, the need to test parameters in psychological research needs to be reexamined. It is time for psychologists to reevaluate how we use statistics.

REFERENCES

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

Bradbury, I. (1987). Analysis of variance versus randomization tests-a comparison. British Journal of Mathematical and Statistical Psychology, 40, 177-187.

Bradley, J.V. (1978). Robustness? British Journal of Mathematical and Statistical Psychology, 31, 144-152.

Bradley, J.V. (1980). The nonrobustness of z, t, and F tests: A large-scale sampling study. Bulletin of the Psychonomic Society, 15, 333-336.

Camilli, G. (1990). The test of homogeneity for a 2x2 contingency tables: A review of some personal opinion on the controversy. Psychological Bulletin, 66, 252-262.

Diaconis, P., & Efron, B. (1983). Computer-intensive methods in statistics. Scientific American, 48, 116-130.

Edgington, E.S. (1969). Approximate Randomization Tests. Journal of Psychology, 57, 445-449.

Edgington, E.S. (1987). Randomization Tests (2nd edition). New York: Marcell Dekker.

Efron, B. (1977) . Bootstrap methods: another look at the jackknife. Annals of Statistics, 7, 1-26.

Efron, B., & Tibshirani, R. (1991) . Statistical data analysis in the computer age. Science, 253, 390-395.

Fisher, R.A. (1935) . The design and analysis of experiments. Edinburgh: Oliver and Boyd.

Hays, W. L. (1988) . Statistics (4th ed.). New York: Holt, Rhinehart, & Winston.

Hinkley, D.V. (1988) . Bootstrap Methods. Journal of the Royal Statistical Society, 50(3), 321-337.

Hunter, M. A., & May, R. B. (1993) . Some Myths Concerning Parametric and Nonparametric Tests. Canadian Psychology, 34 (4), 384-389.

Kempthorne, O. (1979) . The randomization theory of experimental inference. Journal of the American Statistical Association, 50, 115-145.

Kirk , R. E. (1982) . Experimental Design. Pacific Grove, CA: Brookes/Cole.

May, R.B., Masson, M. E. J., & Hunter, M. A. (1990) . Application of statistics in behavioural research. New York: Harper & Row.

May, R.B., Masson, M. E. J., & Hunter, M. A. (1993) . NPSTAT (version 3.7) [computer program]. Department of Psychology, University of Victoria.

Mehta, C., & Patel, N. (1991) . StatXact version 2.0 [computer program] . Cambridge, MA : Cytel Corporation.

Micceri, T. (1989) . The unicorn, the normal curve, and other improbable creatures. Psychological Bulletin, 105 (1), 156-166.

Mook, D.G. (1983) . In defense of external validity. American Psychologist, 38, 379-387.

Mooney, C.Z., & Duval, R.D. (1993) . Bootstrapping: A nonparametric approach to statistical inference. Newbury Park: Sage.

Neyman, J., & Pearson, E. S. (1928) . On the use and interpretation of certain test criteria for purposes of statistical inference. Part I. Biometrika, 20A, 175- 240.

Noreen, E.W. (1989) . Computer Intensive Methods for Testing Hypotheses. Wiley.

Pagano, R.R. (1990) . Understanding Statistics in the Behavioural Sciences. St Paul MN: West.

Pitman, E.J.G. (1937a) . Significance tests which may be applied to samples from any populations. Journal of the Royal Statistical Society (Series B), 4, 119-130.

Pitman, E.J.G. (1937b) . Significance tests which may be applied to samples from any populations II. Journal of the Royal Statistical Society, 4, 225-232.

Figure 1. Smooth Symmetric Distribution

Mean = 13.19 Standard deviation = 4.91 Skew = 0.01

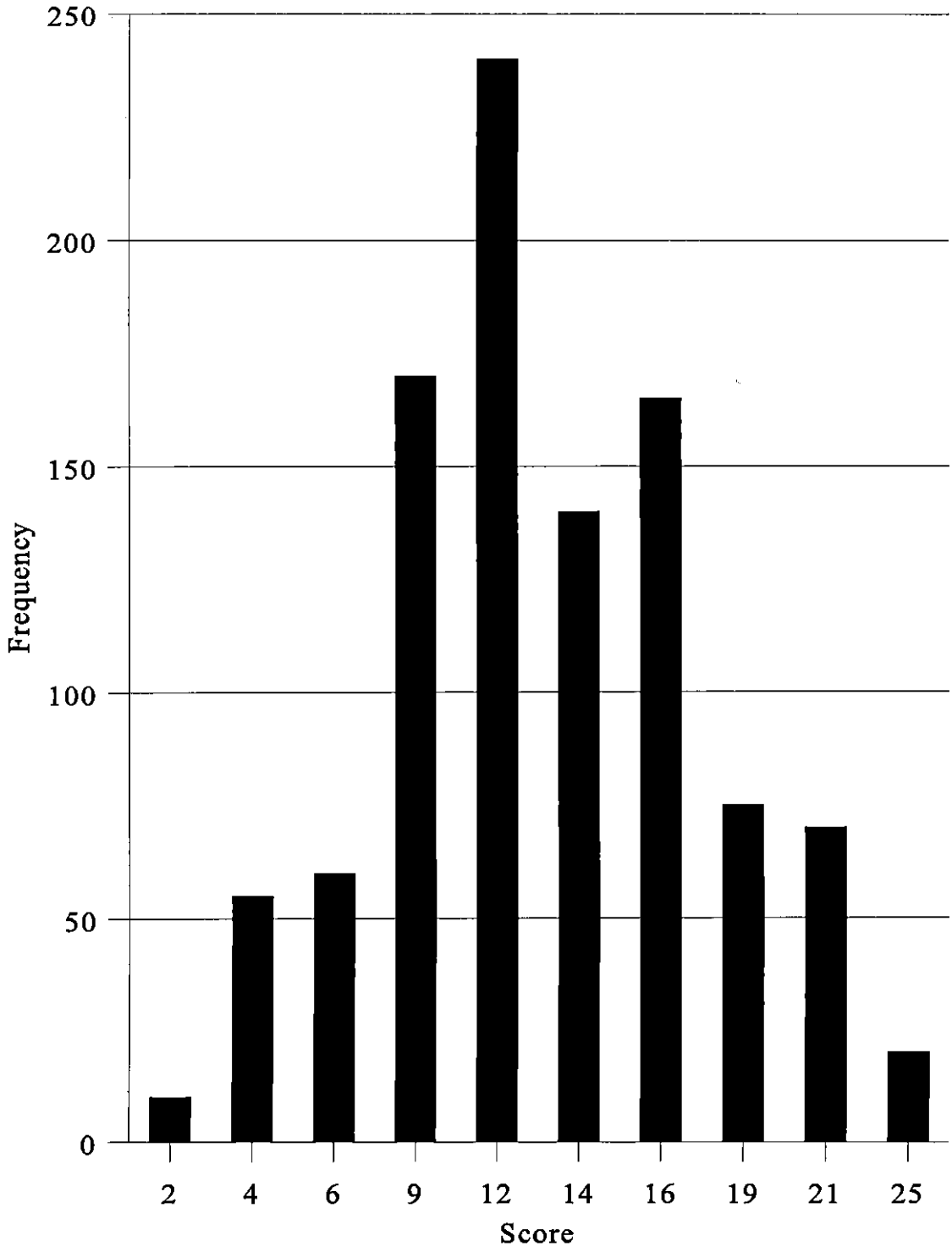


Figure 2. Discrete Mass at Zero Distribution

Mean = 12.92 Standard deviation = 4.42 Skew = -0.03

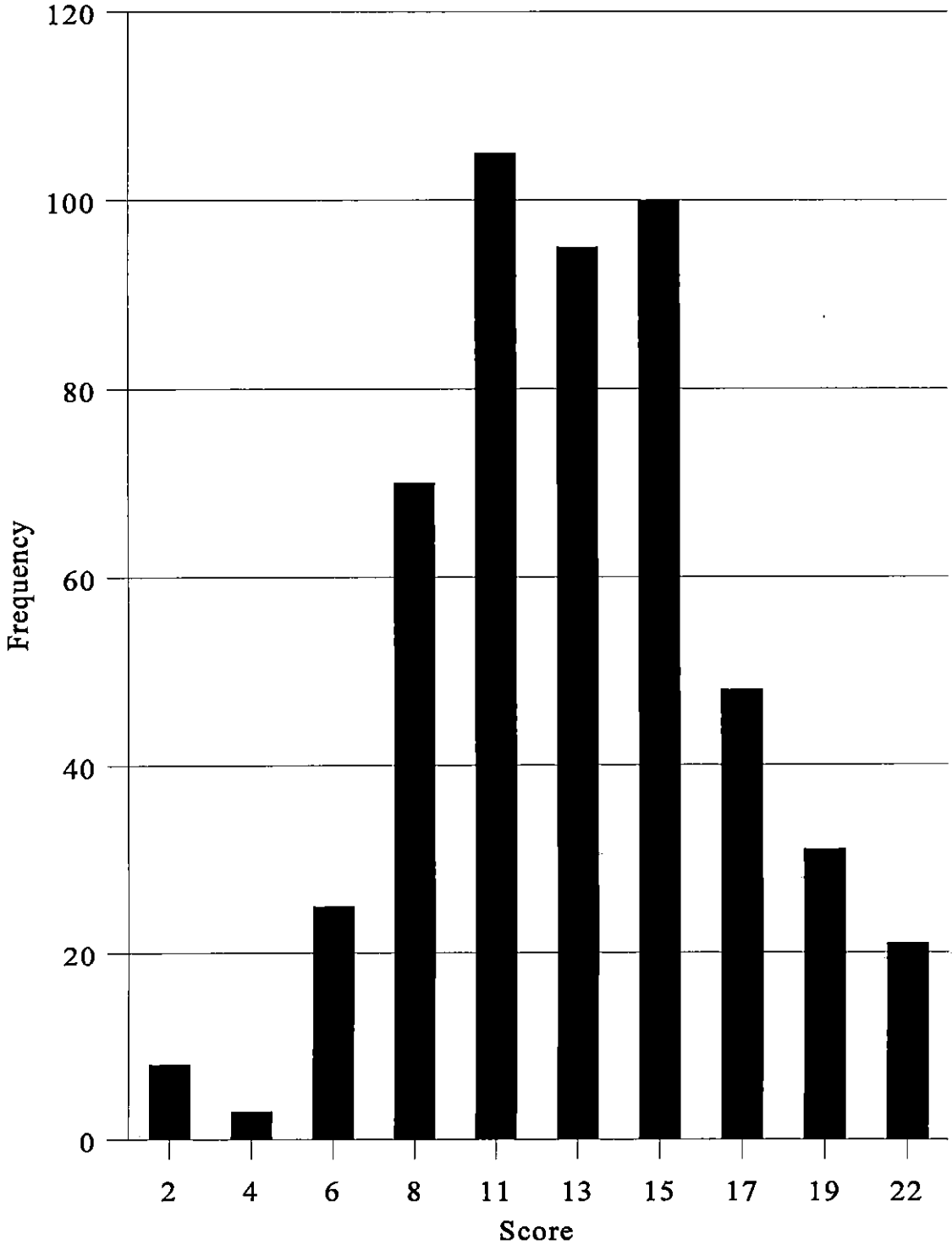


Figure 3. Discrete Mass at Zero with Gap Distribution

Mean = 1.85 Standard deviation = 3.80 Skew = 1.65

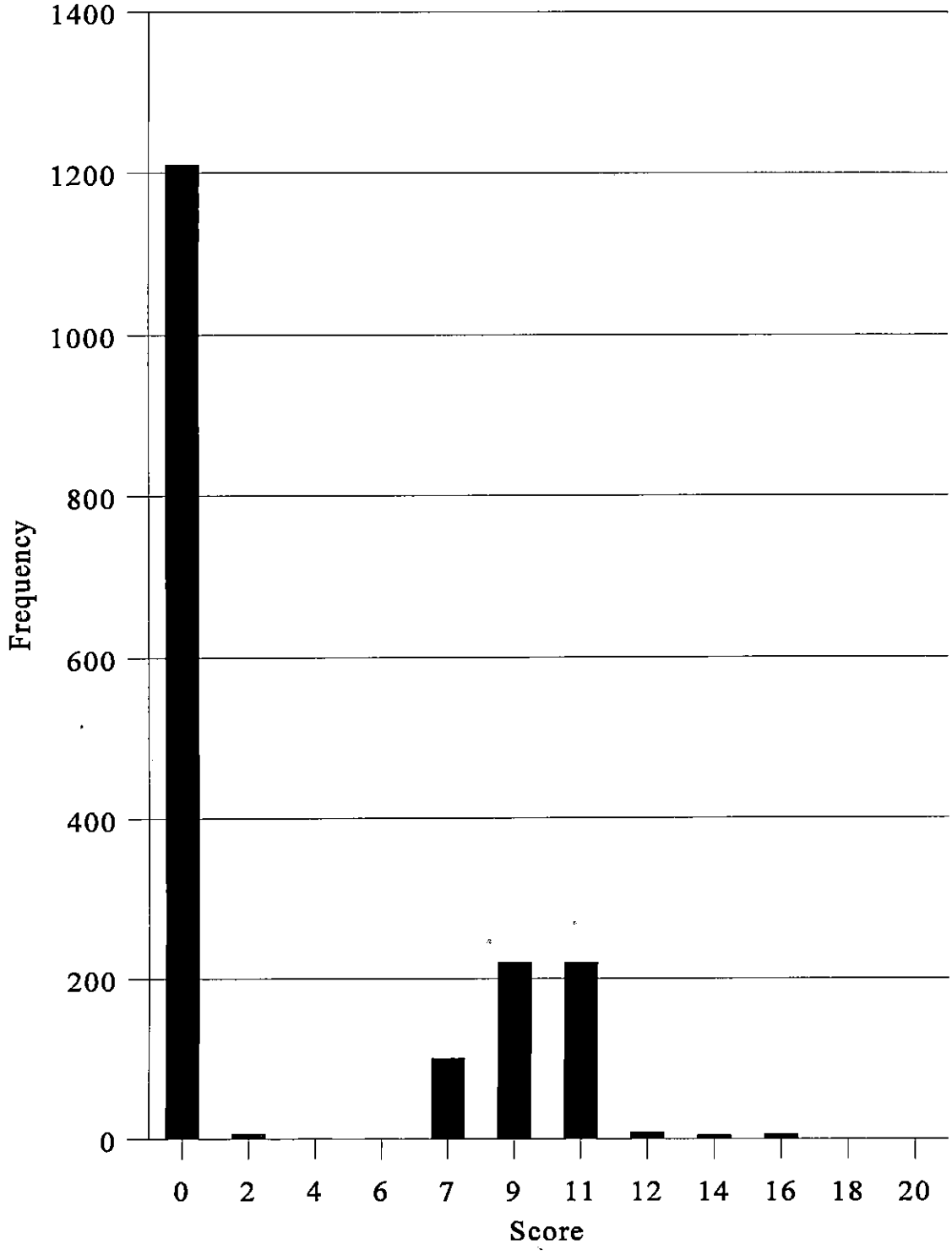


Figure 4. Extreme Asymmetry(Achievement) Distribution

Mean = 24.50 Standard deviation = 5.79 Skew = -1.33

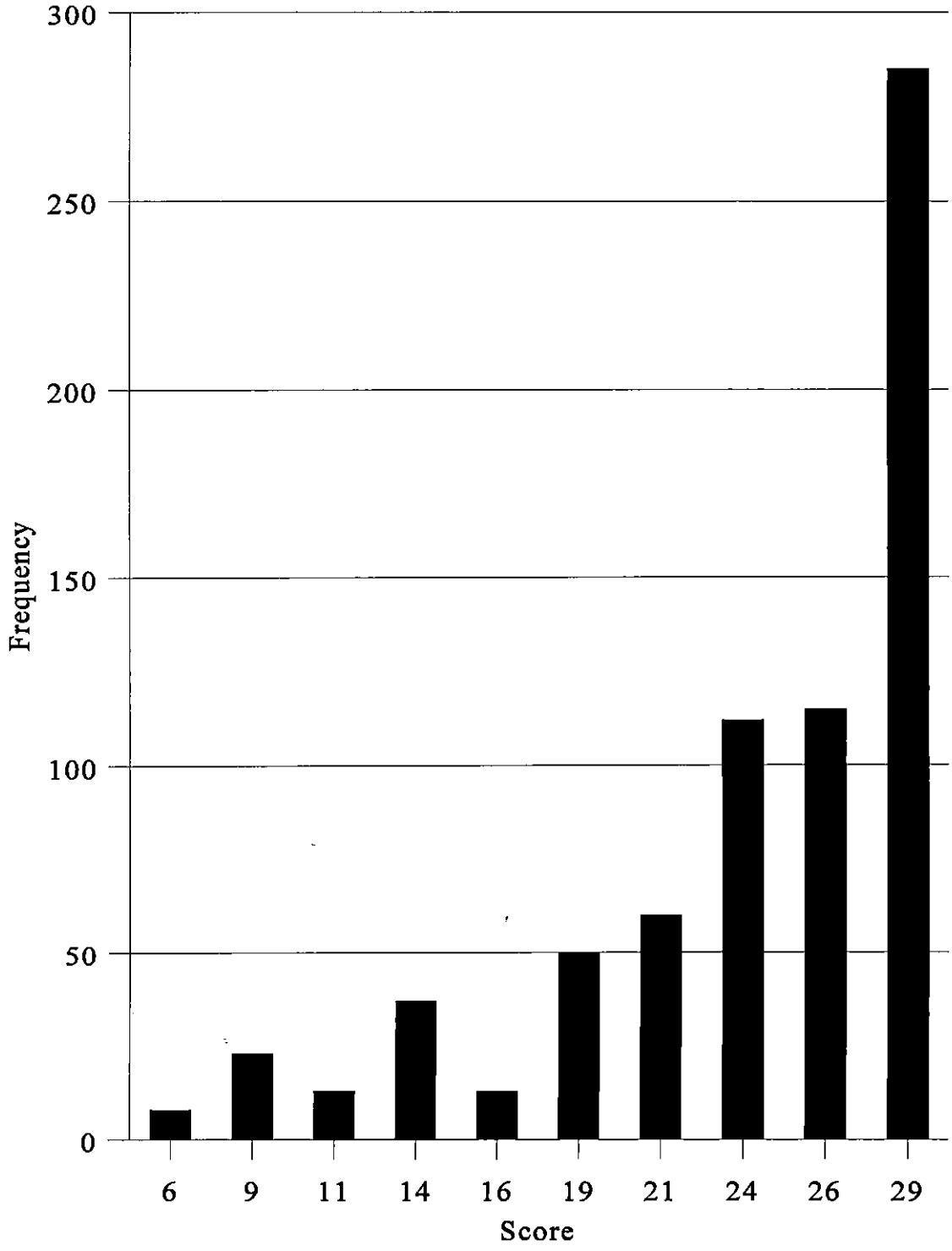


Figure 5. Extreme Asymmetry(Psychometric) Distribution

Mean = 13.67 Standard deviation = 5.75 Skew = 1.65

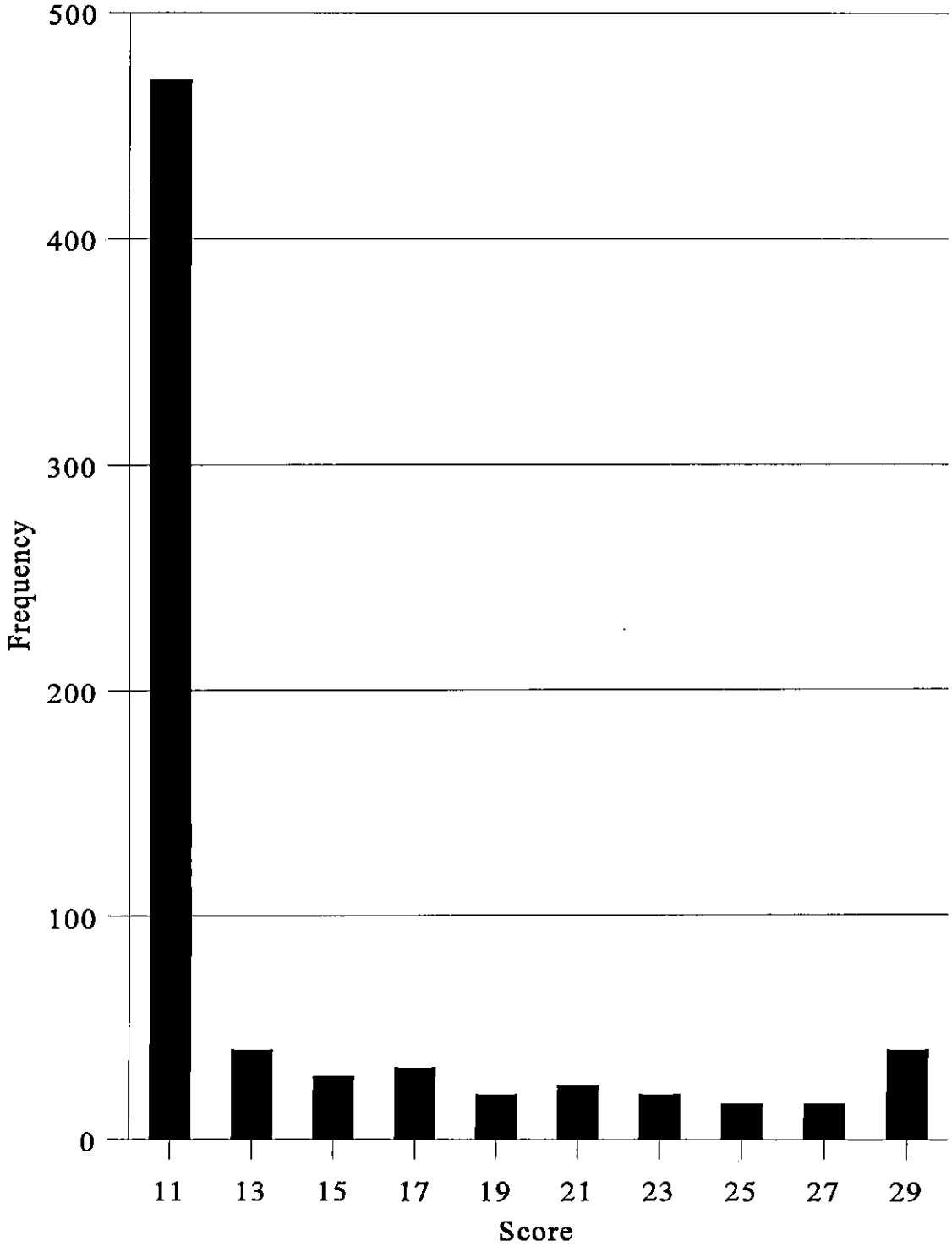


Figure 6. Digit Preference Distribution

Mean = 536.95 Standard deviation = 37.64 Skew = -0.07

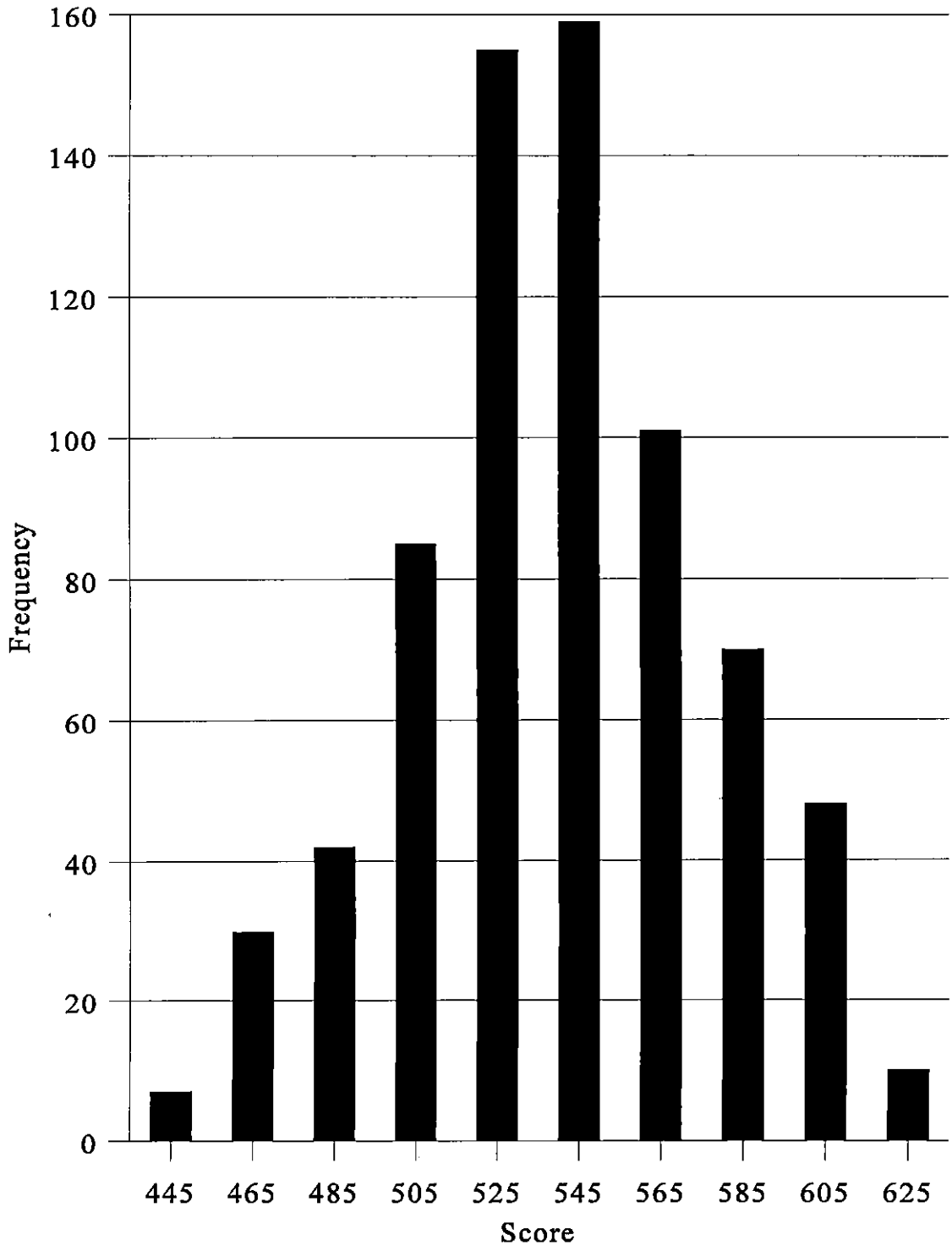


Figure 7. Extreme Bimodality Distribution

Mean = 2.97 Standard deviation = 1.69 Skew = -0.08

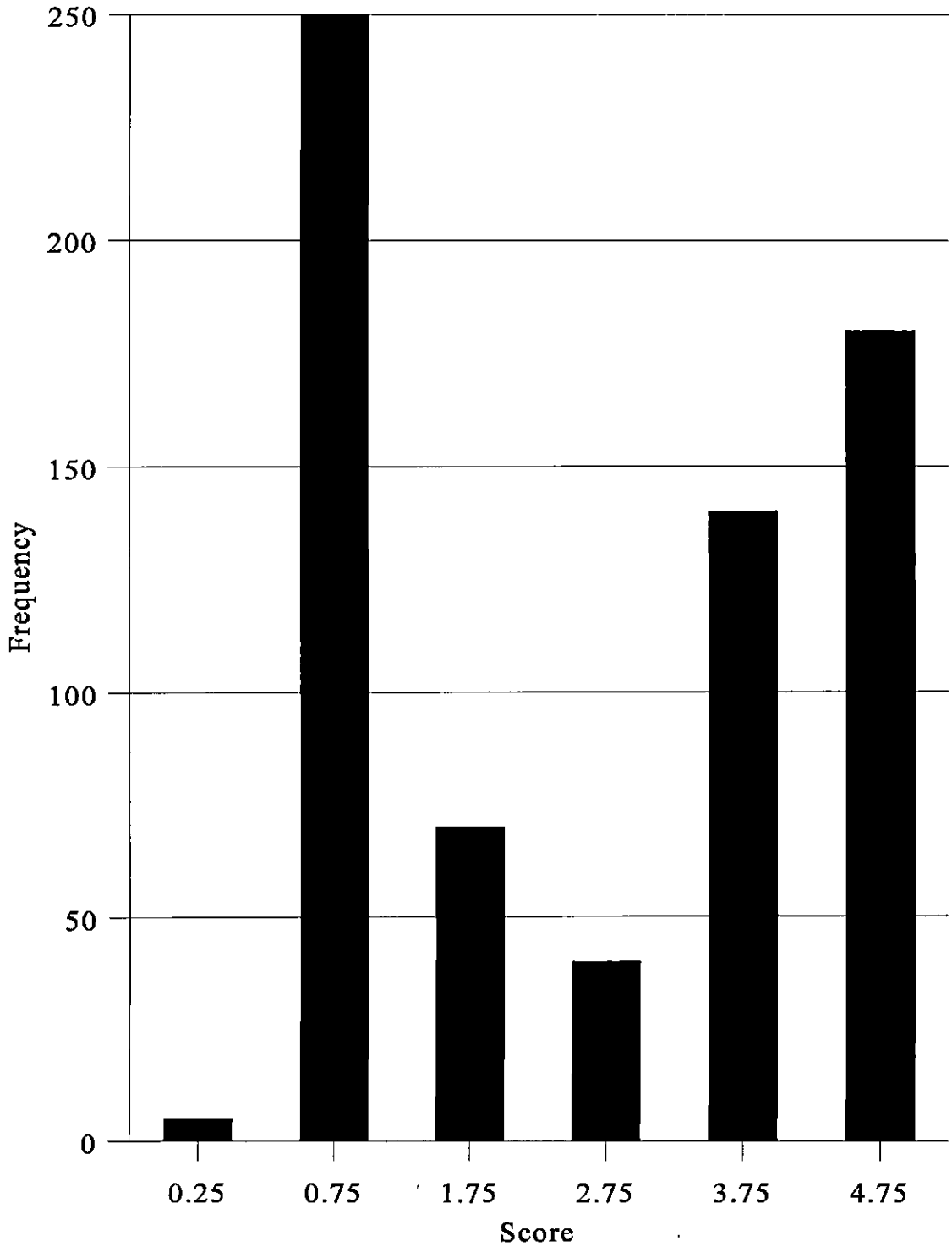


Figure 8. Mulltimodal Lumpy Distribution

Mean = 21.15 Standard deviation = 11.90 Skew = 0.19

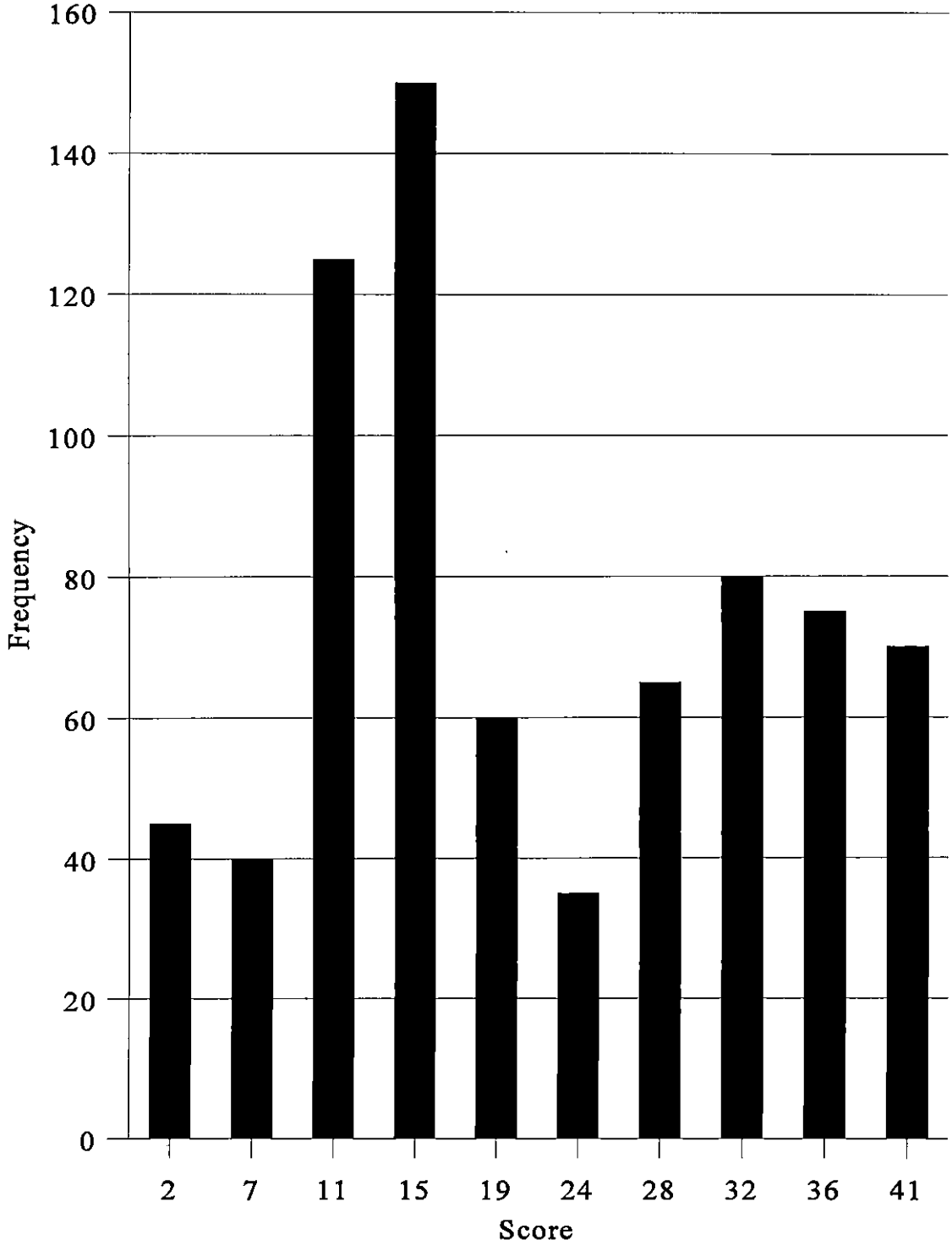


Figure 9. Probabilities of type I error obtained by the three methods for various sample sizes for the Smooth Symmetric distribution

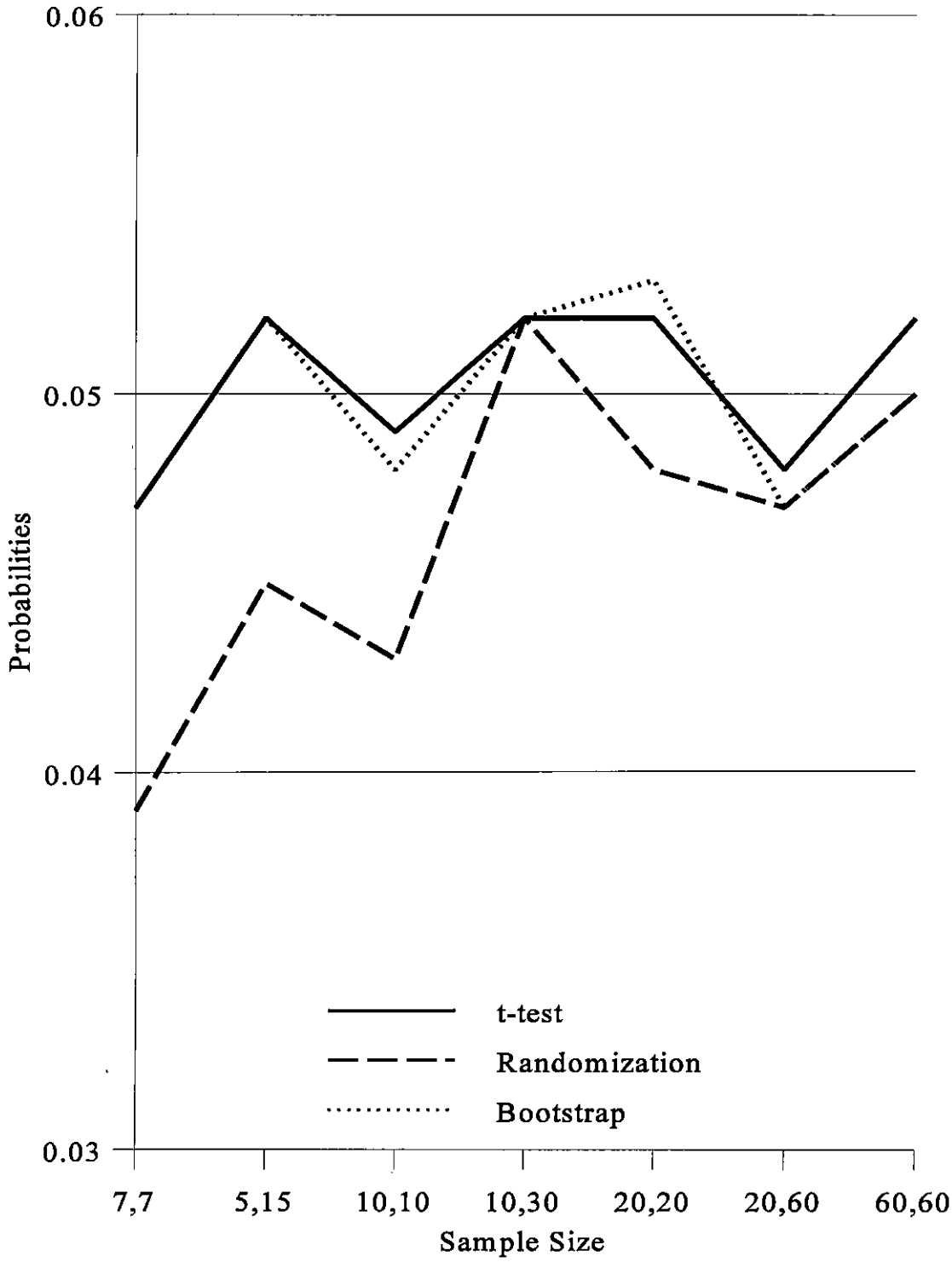


Figure 10. Probabilities of type I error obtained by the three methods for various sample sizes for the Discrete Mass at Zero distribution

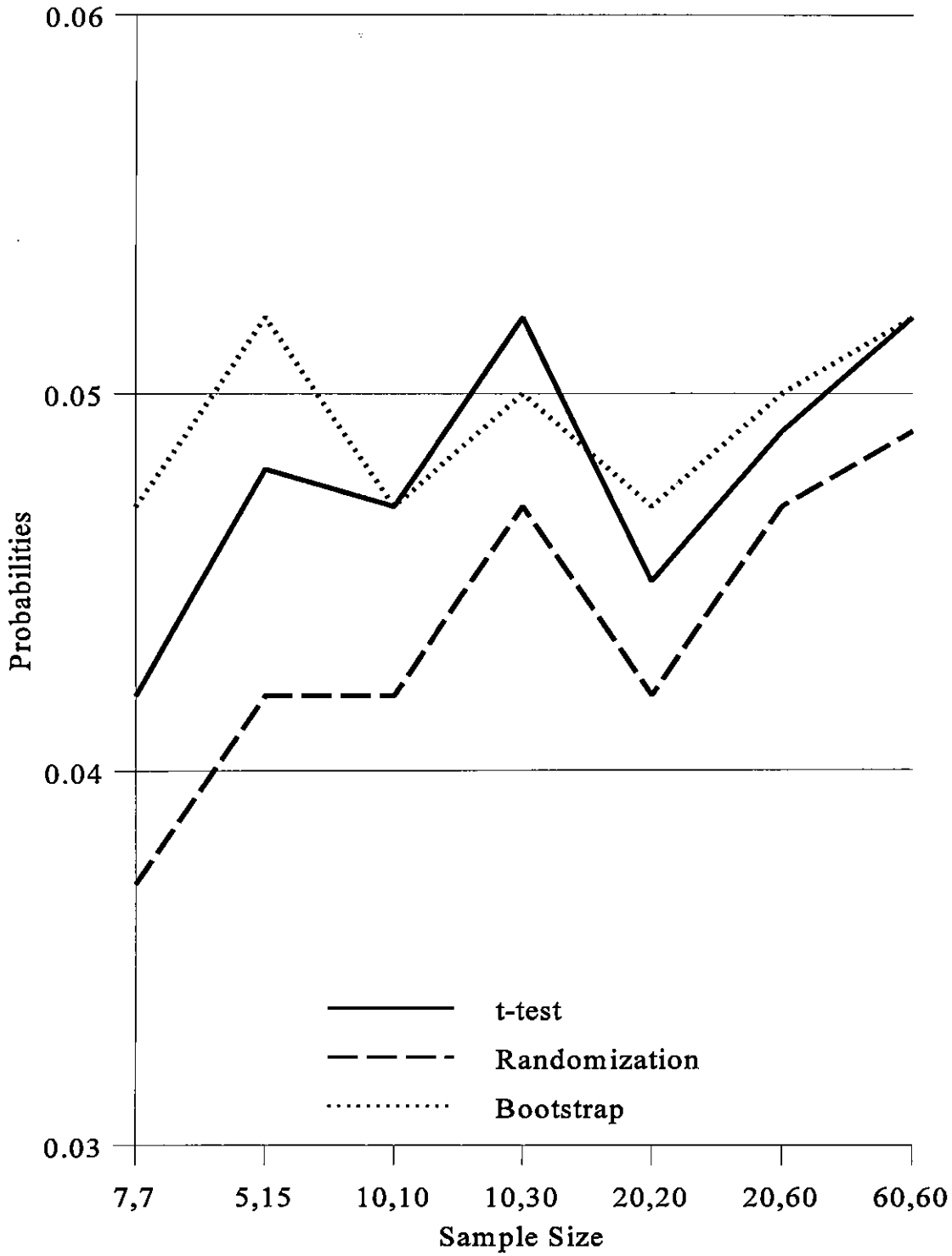


Figure 11. Probabilities of type I error obtained by the three methods for various sample sizes for the Discrete Mass at Zero with Gap distribution

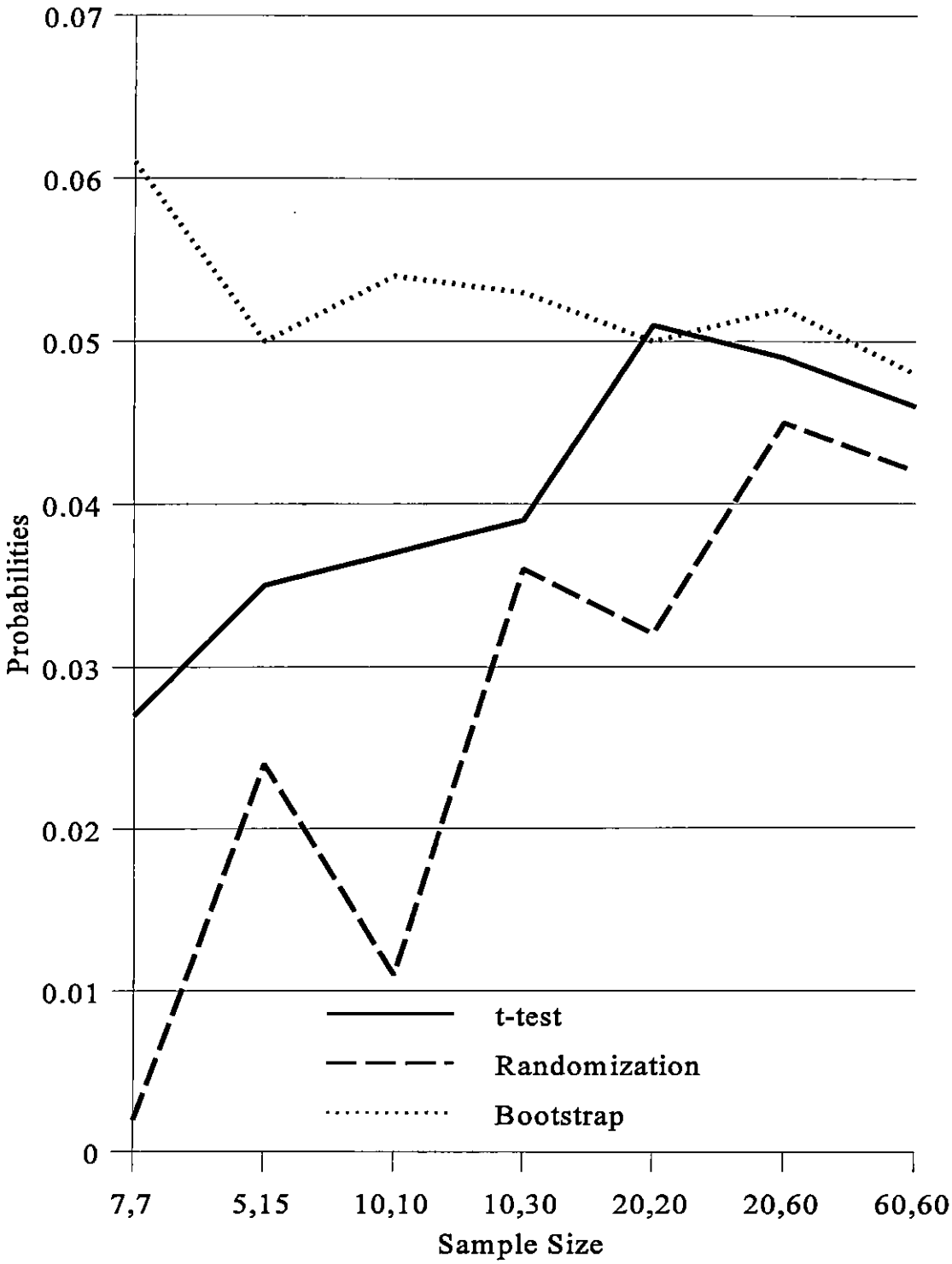


Figure 12. Probabilities of type I error obtained by the three methods for various sample sizes for the Extreme Asymmety(Achievement) distribution

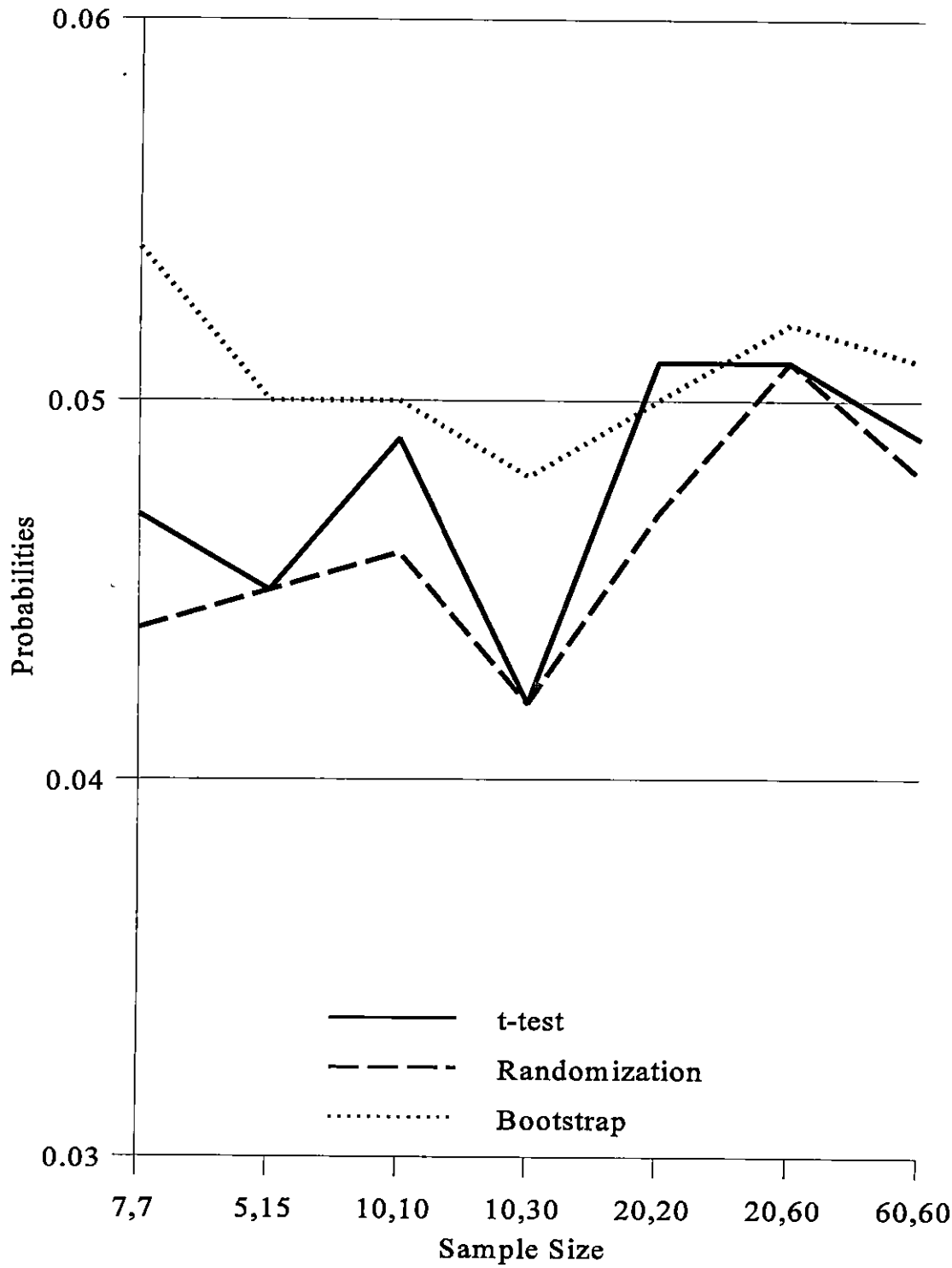


Figure 13. Probabilities of type I error obtained by the three methods for various sample sizes for the Extreme Asymmetry(Psychometric) distribution

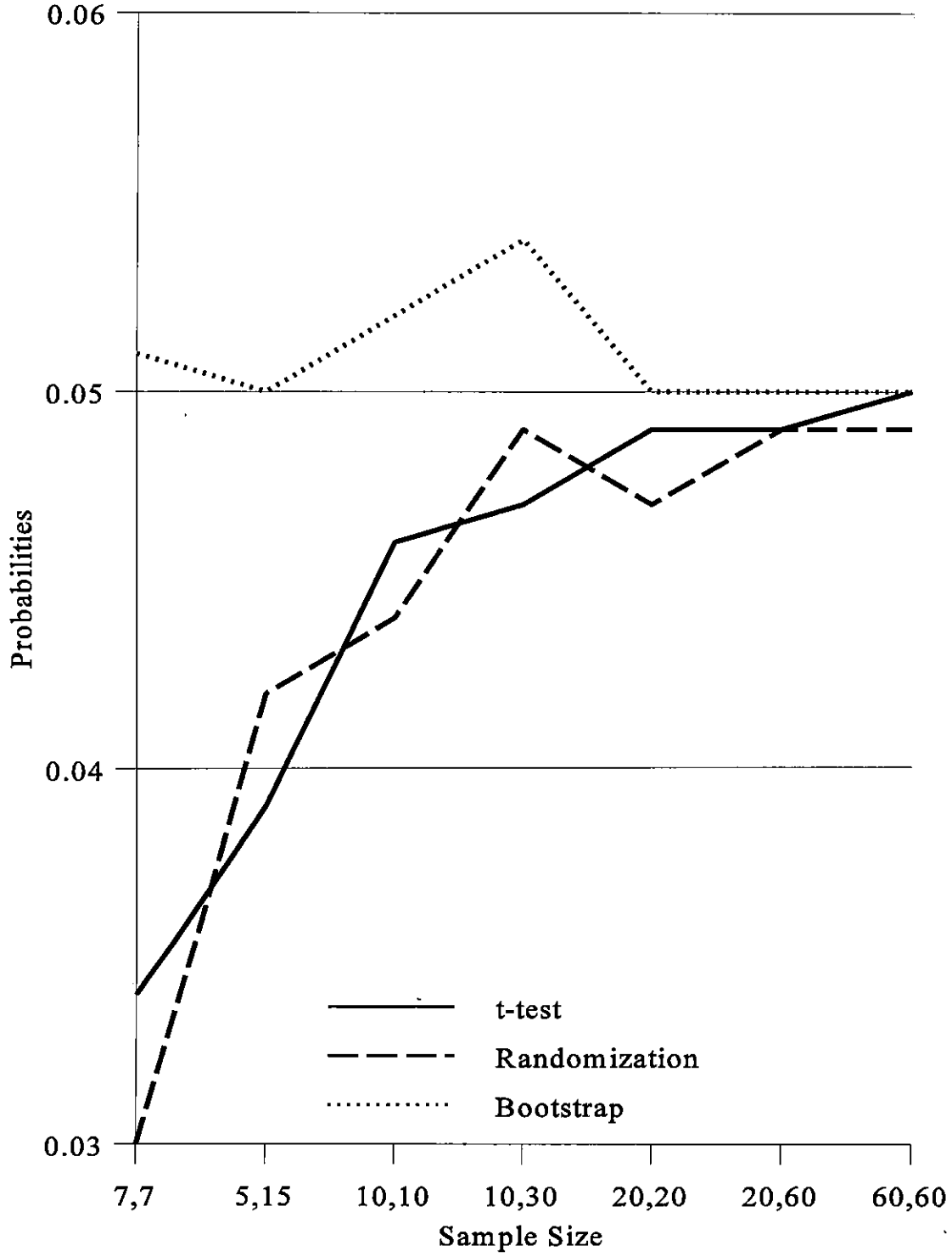


Figure 14. Probabilities of type I error obtained by the three methods for various sample sizes for the Digit Preference distribution

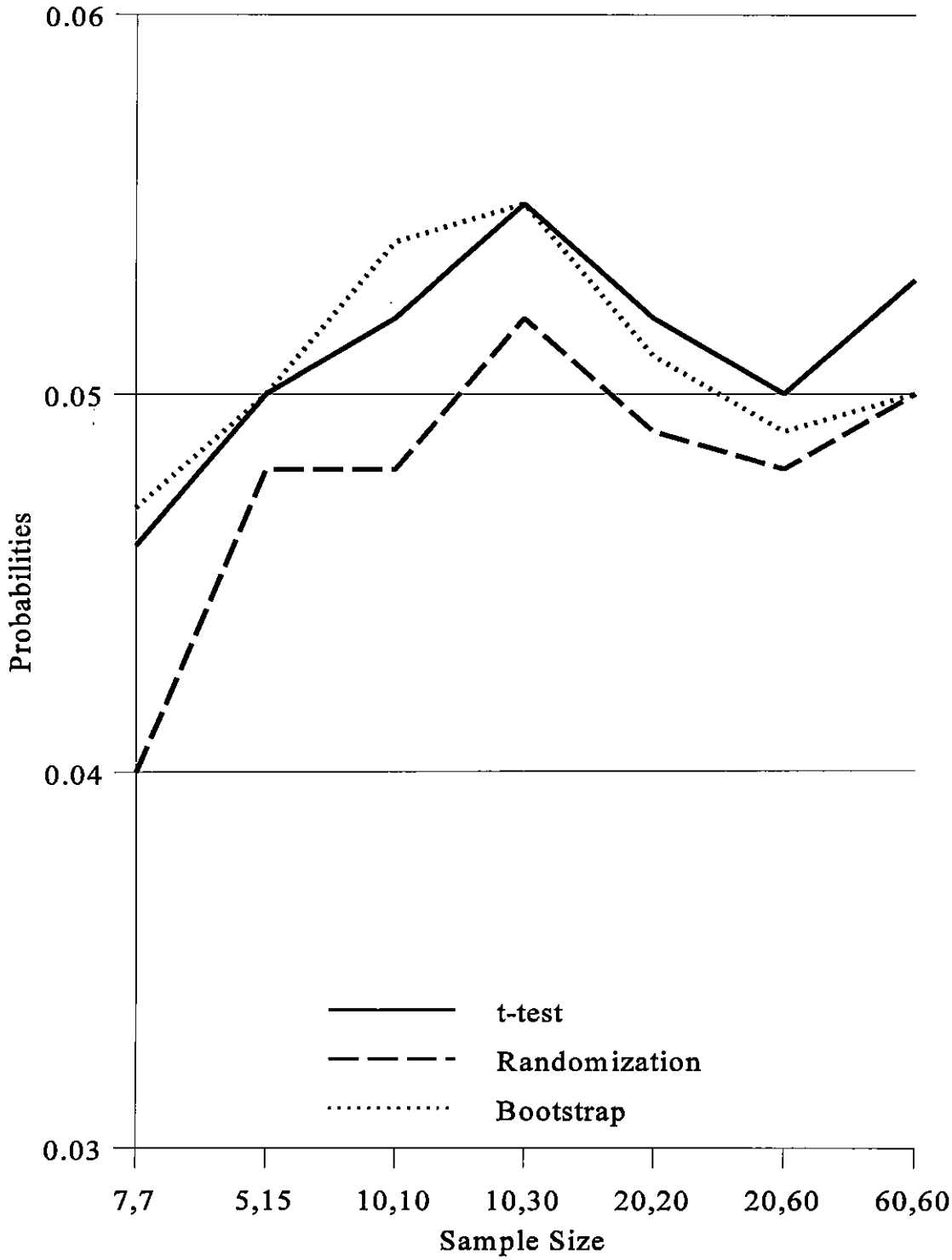


Figure 15. Probabilities of type I error obtained by the three methods for various sample sizes for the Extreme Bimodality distribution

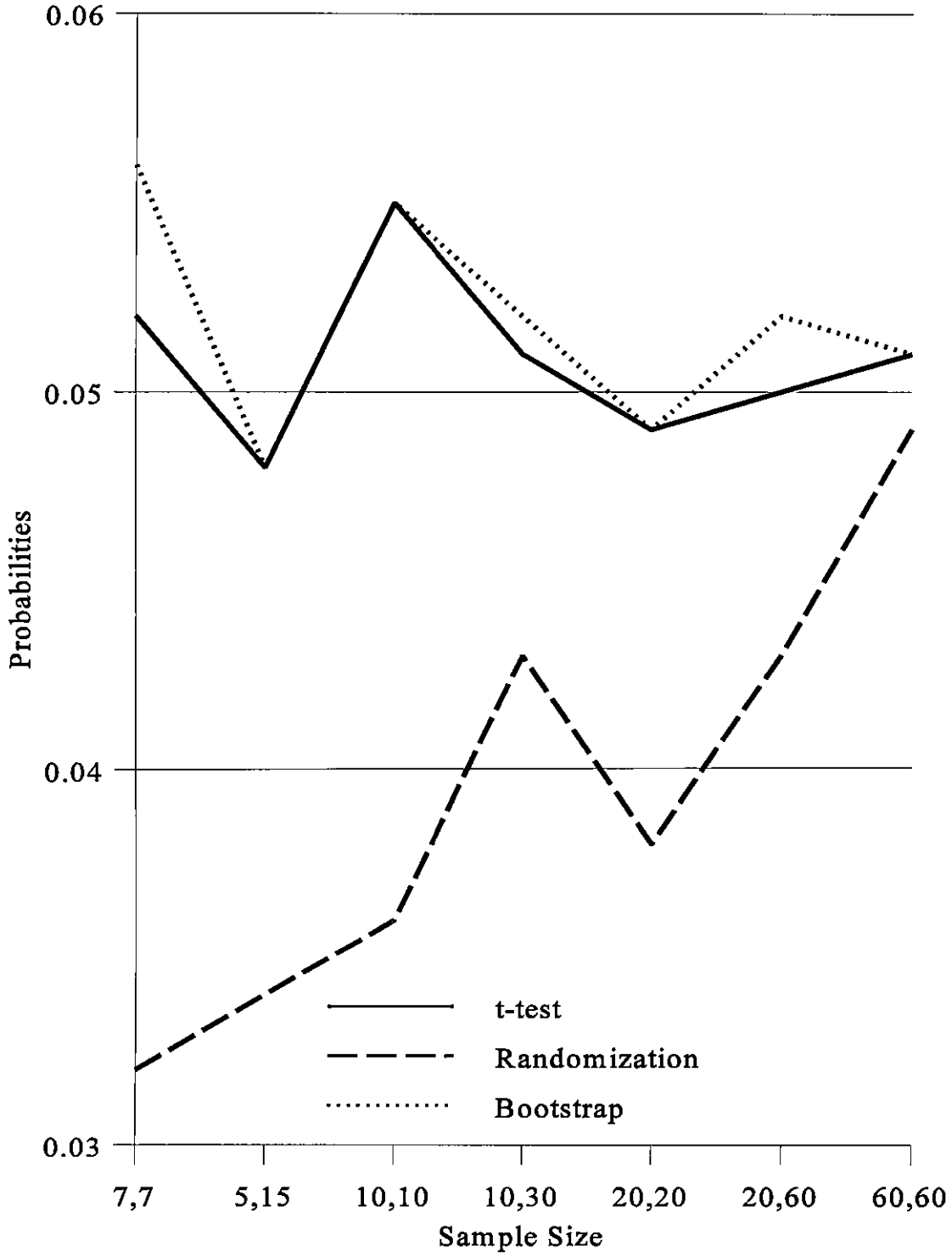
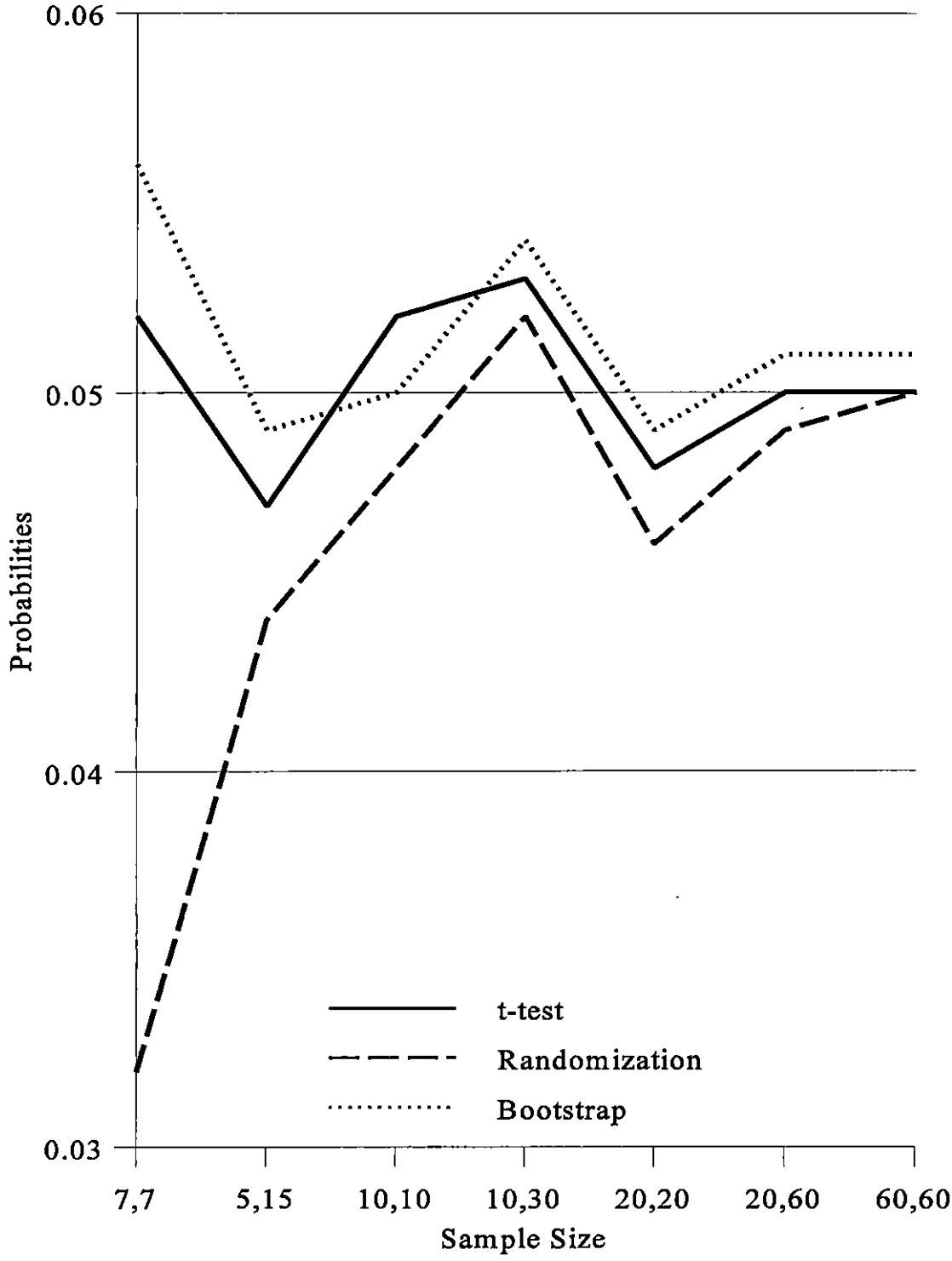


Figure 16. Probabilities of type I error obtained by the three methods for various sample sizes for the Multimodal Lumpy distribution



Appendix A
Complete Output

Sample Permutations = 5000
Boot Samples = 5000
Iterations = 10000
Sample Size (7, 7)

Distribution 1-Smooth Symmetric
t-test percent alpha = 0.046700
rnd percent alpha = 0.038600
boot percent alpha = 0.046800

Distribution 2-Discrete Mass at Zero
t-test percent alpha = 0.041800
rnd percent alpha = 0.036700
boot percent alpha = 0.046800

Distribution 3-Discrete Mass at Zero with Gap
t-test percent alpha = 0.016600
rnd percent alpha = 0.002800
boot percent alpha = 0.061500

Distribution 4-Extreme Asymmetry (Achievement)
t-test percent alpha = 0.046800
rnd percent alpha = 0.044000
boot percent alpha = 0.053800

Distribution 5-Extreme Asymmetry (Psychometric)
t-test percent alpha = 0.034200
rnd percent alpha = 0.030400
boot percent alpha = 0.051500

Distribution 6-Digit Preference
t-test percent alpha = 0.045800
rnd percent alpha = 0.040300
boot percent alpha = 0.047200

Distribution 7-Extreme Bimodality
t-test percent alpha = 0.052100
rnd percent alpha = 0.031900
boot percent alpha = 0.056000

Distribution 8-Multimodal Lumpy
t-test percent alpha = 0.052100
rnd percent alpha = 0.031900
boot percent alpha = 0.056000

Sample Permutations = 5000
Boot Samples = 5000
Iterations = 10000
Sample size (5, 15)

Distribution 1-Smooth Symmetric
t-test percent alpha = 0.051700
rnd percent alpha = 0.045300
boot percent alpha = 0.052300

Distribution 2-Discrete Mass at Zero
t-test percent alpha = 0.047700
rnd percent alpha = 0.041600
boot percent alpha = 0.051600

Distribution 3-Discrete Mass at Zero with Gap
t-test percent alpha = 0.034500
rnd percent alpha = 0.024300
boot percent alpha = 0.049500

Distribution 4-Extreme Asymmetry (Achievement)
t-test percent alpha = 0.045000
rnd percent alpha = 0.044900
boot percent alpha = 0.050100

Distribution 5-Extreme Asymmetry (Psychometric)
t-test percent alpha = 0.039200
rnd percent alpha = 0.041500
boot percent alpha = 0.049800

Distribution 6-Digit Preference
t-test percent alpha = 0.050400
rnd percent alpha = 0.047700
boot percent alpha = 0.051200

Distribution 7-Extreme Bimodality
t-test percent alpha = 0.047500
rnd percent alpha = 0.033700
boot percent alpha = 0.048200

Distribution 8-Multimodal Lumpy
t-test percent alpha = 0.046900
rnd percent alpha = 0.044300
boot percent alpha = 0.048900

Sample Permutations = 5000
Boot Samples = 5000
Iterations = 10000
Sample Size (10, 10)

Distribution 1-Smooth Symmetric
t-test percent alpha = 0.049000
rnd percent alpha = 0.042500
boot percent alpha = 0.047900

Distribution 2-Discrete Mass at Zero
t-test percent alpha = 0.047300
rnd percent alpha = 0.042000
boot percent alpha = 0.0471

Distribution 3-Discrete Mass at Zero with Gap
t-test percent alpha = 0.037200
rnd percent alpha = 0.011200
boot percent alpha = 0.053700

Distribution 4-Extreme Asymmetry (Achievement)
t-test percent alpha = 0.048800
rnd percent alpha = 0.046200
boot percent alpha = 0.049600

Distribution 5-Extreme Asymmetry (Psychometric)
t-test percent alpha = 0.046400
rnd percent alpha = 0.043700
boot percent alpha = 0.052100

Distribution 6-Digit Preference
t-test percent alpha = 0.052200
rnd percent alpha = 0.047600
boot percent alpha = 0.053900

Distribution 7-Extreme Bimodality
t-test percent alpha = 0.055200
rnd percent alpha = 0.036100
boot percent alpha = 0.055100

Distribution 8-Multimodal Lumpy
t-test percent alpha = 0.052000
rnd percent alpha = 0.048800
boot percent alpha = 0.050200

Sample Permutations = 5000
Boot Samples = 5000
Iterations = 10000
Sample Size (10, 30)

Distribution 1-Smooth Symmetric
t-test percent alpha = 0.051900
rnd percent alpha = 0.052900
boot percent alpha = 0.051600

Distribution 2-Discrete Mass at Zero
t-test percent alpha = 0.051900
rnd percent alpha = 0.046900
boot percent alpha = 0.049600

Distribution 3-Discrete Mass at Zero with Gap
t-test percent alpha = 0.038500
rnd percent alpha = 0.035600
boot percent alpha = 0.052800

Distribution 4-Extreme Asymmetry (Achievement)
t-test percent alpha = 0.042300
rnd percent alpha = 0.042000
boot percent alpha = 0.047800

Distribution 5-Extreme Asymmetry (Psychometric)
t-test percent alpha = 0.047000
rnd percent alpha = 0.049100
boot percent alpha = 0.053600

Distribution 6-Digit Preference
t-test percent alpha = 0.055200
rnd percent alpha = 0.052500
boot percent alpha = 0.055300

Distribution 7-Extreme Bimodality
t-test percent alpha = 0.050900
rnd percent alpha = 0.042700
boot percent alpha = 0.052300

Distribution 8-Multimodal Lumpy
t-test percent alpha = 0.053400
rnd percent alpha = 0.052200
boot percent alpha = 0.053500

Sample Permutations = 5000
Boot Samples = 5000
Iterations = 10000
Sample Size (20, 20)

Distribution 1-Smooth Symmetric
t-test percent alpha = 0.051900
rnd percent alpha = 0.048200
boot percent alpha = 0.053100

Distribution 2-Discrete Mass at Zero
t-test percent alpha = 0.045200
rnd percent alpha = 0.042200
boot percent alpha = 0.046600

Distribution 3-Discrete Mass at Zero with Gap
t-test percent alpha = 0.051600
rnd percent alpha = 0.032400
boot percent alpha = 0.050400

Distribution 4-Extreme Asymmetry (Achievement)
t-test percent alpha = 0.051200
rnd percent alpha = 0.047400
boot percent alpha = 0.050100

Distribution 5-Extreme Asymmetry (psychometric)
t-test percent alpha = 0.049200
rnd percent alpha = 0.047000
boot percent alpha = 0.049500

Distribution 6-Digit Preference
t-test percent alpha = 0.051800
rnd percent alpha = 0.048700
boot percent alpha = 0.050600

Distribution 7-Extreme Bimodality
t-test percent alpha = 0.048700
rnd percent alpha = 0.038200
boot percent alpha = 0.048800

Distribution 8-Multimodal Lumpy
t-test percent alpha = 0.048300
rnd percent alpha = 0.045800
boot percent alpha = 0.047800

Sample Permutations = 5000
Boot Samples = 5000
Iterations = 10000
Sample Size (20, 60)

Distribution 1-Smooth Symmetric
t-test percent alpha = 0.048800
rnd percent alpha = 0.046800
boot percent alpha = 0.048600

Distribution 2-Discrete Mass at Zero
t-test percent alpha = 0.049000
rnd percent alpha = 0.047200
boot percent alpha = 0.050100

Distribution 3-Discrete Mass at Zero with Gap
t-test percent alpha = 0.048700
rnd percent alpha = 0.045300
boot percent alpha = 0.052000

Distribution 4-Extreme Asymmetry (Achievement)
t-test percent alpha = 0.051300
rnd percent alpha = 0.050700
boot percent alpha = 0.051500

Distribution 5-Extreme Asymmetry (Psychometric)
t-test percent alpha = 0.0492350
rnd percent alpha = 0.049300
Boot percent alpha = 0.050500

Distribution 6-Digit Preference
t-test percent alpha = 0.049500
rnd percent alpha = 0.047700
boot percent alpha = 0.047800

Distribution 7-Extreme Bimodality
t-test percent alpha = 0.050100
rnd percent alpha = 0.043400
boot percent alpha = 0.052200

Distribution 8-Multimodal Lumpy
t-test percent alpha = 0.050100
rnd percent alpha = 0.049400
boot percent alpha = 0.051200

Sample Permutations = 5000
Boot Samples = 5000
Iterations = 10000
Sample Size (60, 60)

Distribution 1-Smooth Symmetric
t-test percent alpha = 0.051700
rnd percent alpha = 0.049200
boot percent alpha = 0.050300

Distribution 2-Discrete Mass at Zero
t-test percent alpha = 0.051900
rnd percent alpha = 0.048600
boot percent alpha = 0.051800

Distribution 3-Discrete Mass at Zero with Gap
t-test percent alpha = 0.046100
rnd percent alpha = 0.042100
boot percent alpha = 0.048000

Distribution 4-Extreme Asymmetry (Achievement)
t-test percent alpha = 0.049300
rnd percent alpha = 0.047600
boot percent alpha = 0.051400

Distribution 5-Extreme Asymmetry (psychometric)
t-test percent alpha = 0.049800
rnd percent alpha = 0.048700
boot percent alpha = 0.049600

Distribution 6-Digit Preference
t-test percent alpha = 0.053300
rnd percent alpha = 0.049600
boot percent alpha = 0.050001

Distribution 7-Extreme Bimodality
t-test percent alpha = 0.051300
rnd percent alpha = 0.048800
boot percent alpha = 0.051200

Distribution 8-Multimodal Lumpy
t-test percent alpha = 0.050800
rnd percent alpha = 0.049600
boot percent alpha = 0.051300

VITA

Surname: Miles Given Names: Stephen Allen

Place of Birth: Nairobi, Kenya Date of Birth: November 28, 1967

Educational Institutions Attended:

University of Victoria	1993 to 1997
University of Victoria	1991 to 1993
Queens University	1987 to 1990

Degrees Awarded:

B.A.	Queens University	1990

Honours and Awards:

Ontario Scholar	1986
University of Victoria Teaching Fellowship	1993 to 1994

PARTIAL COPYRIGHT LICENSE

I hereby grant the right to lend my thesis to users of the University of Victoria Library, and to make single copies only for such users or in response to a request from the Library of any other university, or similar institution, on its behalf or for one of its users. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by me or a member of the University designated by me. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Title of Thesis:

A Comparison of Independent Samples t Test, Approximate Randomization Test and Bootstrap Randomization Test to Population Non-normality: A Monte Carlo Study.

Autho 

Stephen Allen Miles

10 May 1997