COMPUTER-INTENSIVE METHODS FOR TESTING HYPOTHESES:

AN INTRODUCTION

Eric W. Noreen
University of Washington

A Wiley-Interscience Publication
JOHN WILEY & SONS
New York · Chichester · Brisbane Toronto Singapore

CHAPTER ONE

Introduction

The next few years are likely to be an exciting period for those involved in testing hypotheses. Recent dramatic decreases in the costs of computing now make revolutionary methods for testing hypotheses available to anyone with access to a personal computer. These methods are easy to understand, very general, and can often avoid troublesome assumptions that are required with conventional methods.

1.1 ASSESSING SIGNIFICANCE IN A HYPOTHESIS TEST

Three ingredients are usually required for a hypothesis test: a hypothesis, a test statistic, and some means of generating the probability distribution of the test statistic under the assumption that the hypothesis is true. The first ingredient, the hypothesis, should be suggested by substantive theory. For example, economists would predict that, all other things being equal, an increase in the supply of corn should lead to a decrease in the price of corn. Because of the logic of inference, the hypothesis is ordinarily stated negatively in the form of a null hypothesis which the researcher would like to reject. To take the above example, the null hypothesis might be that there is no relationship between the supply and the price of corn. The alternative hypothesis is that the price of corn is negatively related to its supply.

A test statistic is the second ingredient required for a hypothesis test. A test statistic can be any single-valued function of the data. For example, the average value of a variable across all cases is a single-valued function of the data. There are many possible test statistics in any given situation. Often test statistics are selected because they are familiar or because the distribution of the test statistic is known for a sufficiently structured null hypothesis.[1] However, a test statistic should be chosen because its value is most sensitive to the veracity of the substantive theory being tested. In other words, a test statistic should have the characteristic that the larger the value of the test statistic, the stronger the evidence of departure from the null hypothesis in the direction indicated by the substantive theory.[2] For example, economic theory suggests that as the supply of corn increases, its price should fall; that is, supply and price should be negatively correlated. A natural test statistic would be the negative of the correlation between the two variables. If the relationship between the price and supply of corn is perfect, the correlation will be -1 and the value of the test statistic will be +1. In cases where the relationship is strong but not perfect, the value of the test statistic will be positive, but less than +1.

The third ingredient required for a hypothesis test is some means of generating the probability distribution of the test statistic under the assumption that the null hypothesis is true. In conventional statistics, this is ordinarily accomplished by adding structure to the null hypothesis in such a way that it is possible to analytically derive the probability distribution. For the example of research involving the relationship between the price and the supply of corn, the simple null

[1] The distribution of a test statistic can be "known" in the following sense: if a null hypothesis is sufficiently structured, it may be possible to analytically derive the sampling distribution of the test statistic. For example, if the null hypothesis is that observations are randomly selected from a Normal population with zero mean and an unknown variance, then the sampling distribution of the standardized sample mean is the Student's t distribution.

[2] In this book the test statistic is always transformed so that a larger value is evidence of closer agreement with substantive theory.

hypothesis of interest is that the price of corn does not depend on the supply of corn. However, when the t test is used to assess the significance of the correlation, the null hypothesis is implicitly much more structured and complex than the researcher would like. Under the assumptions that the price of corn and the supply of corn are independently and Normally distributed random variables with constant means and variances, the probability distribution of the sample correlation can be analytically derived. Without this assumption (or some other restrictive assumption), the probability distribution of the correlation cannot be derived.

Note that the conventional t test is thus a test of the joint hypothesis that the two variables are independently and Normally distributed. The only part of this joint hypothesis that the researcher really cares about is that the variables are independent; the condition that the variables are Normally distributed with constant means and constant variances is added simply to allow the statistician to compute the probability distribution of the test statistic. As a consequence, however, if the null hypothesis is rejected, it may have been because the price of corn is related to the supply of corn or it may have been because the price of corn or the supply of corn is not Normally distributed.

Why invoke the Normality assumption when it is not necessary? Using computer-intensive methods, the simple hypothesis that the price and supply of corn are unrelated can be directly tested without having to add extraneous, but analytically convenient, conditions to the null hypothesis.

The process of assessing the significance of a test statistic is illustrated in Figure 1.1.

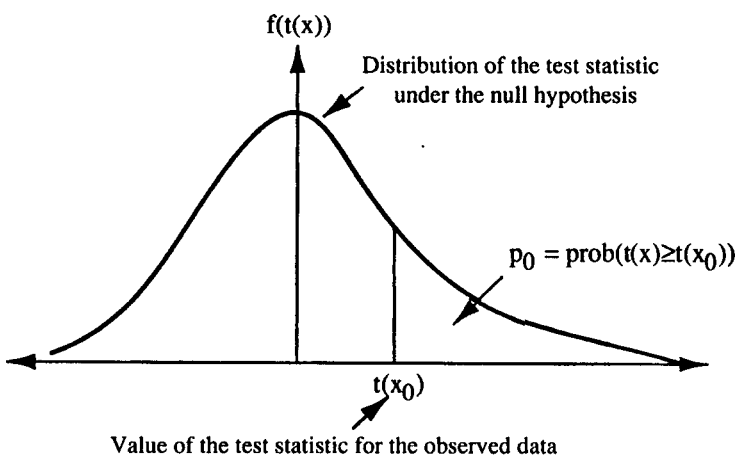

Figure 1.1 Testing the null hypothesis

In Figure 1.1, $t(x_0)$ is the value of the test statistic for the observed data x_0. The probability density function of the test statistic under the null hypothesis is $f(t(x))$. This probability density function can be estimated using computer-intensive methods, or it may be possible to analytically derive $f(t(x))$ if additional conditions are imposed in the null hypothesis. The probability, assuming that the null hypothesis is true, that the test statistic would have been as large or larger than $t(x_0)$ is the area to the right of $t(x_0)$ under the probability density function $f(t(x))$. This area is denoted by $p_0 = \text{prob}(t(x) \geq t(x_0))$. If this area is small, then given the value of the test statistic actually observed, it is unlikely that the null hypothesis is true.

It is good practice for researchers to report this exact significance level p_0. However, in many fields it is common to report only the value of the test statistic and whether or not the test statistic is significant at a conventional rejection level such as 0.05 or 0.10. The significance level of the test – that is, the exact value of p_0 for the test – is more informative. Kempthorne and Doerfler [1969, p. 239] argue:

> *It is obvious that in choosing a fixed level, such as the conventional 5%, merely reporting that the significance level is less than or equal to 5% is a condensation of the data which results in loss of information, however that may be defined. Our attitude to a situation in which the significance level is 0.024 should be different from one in which the significance level is 0.045, for instance.*

Or, perhaps more importantly, our attitude to a situation in which the attained significance level is 0.16 should be different from one in which the significance level is 0.96. Further, Edgington [1970, p. 110] argues that "there is not necessarily any particular p_0 value that would cause [the researcher] to switch to a definite disbelief, and so the accept-reject dichotomy is inappropriate."

Nevertheless, it is common practice to say that the null hypothesis is rejected if the attained significance of the test p_0 is less than a conventional rejection level such as 0.05 or 0.10.[3] A careful researcher would say in such a case that the evidence is consistent with the theory. If the probability p_0 is greater than the prespecified rejection level for the test, then the null hypothesis is not rejected. Suppose, for example, that p_0 is 0.42. The implication is that if the null hypothesis is true, the test statistic would be as large as it was 42% of the time. Thus, the probability that a value as large as the actual value of the test statistic would have been observed, even though the null hypothesis is true, is too large to be able to reject the null hypothesis. Narrowly, the sole function of the computer-intensive methods discussed in this book (and, for that matter, of

[3] Pearson [1937] traces the notion of a prespecified rejection level to agricultural research, where rejection of the null hypothesis often would lead to costly repetition of the experiment. A preset rejection level permitted precise control of the frequency with which costly repetitions would be undertaken when the null hypothesis is in fact true.

conventional methods) is to estimate the probability p_0.

There is an ecclesiastical analogy to the process of testing a hypothesis. When an individual is proposed for beatification (the first step toward canonization as a saint) in the Roman Catholic Church, two officials of the Congregation of Rites in Rome are appointed to the case. One official, the "advocate of the cause," is charged with preparing a brief in favor of beatification that deals with the life of, and miracles attributed to, the candidate for beatification. The other, the "promoter of the faith" (popularly known as "the devil's advocate"), is responsible for preparing a brief against the cause. In particular, miracles attributed to the candidate are carefully scrutinized to determine if they could have some natural cause.

In research, the "cause" is the theory and the researcher ideally combines the functions of both advocate of the cause and devil's advocate. A carefully constructed research design marshals evidence in favor of the theory while controlling for the possibility that the evidence is explained by other causes. It is customary to pay particular attention to the possibility that evidence that appears to support the theory does so only by chance. A random sample may not be representative of the population from which it is drawn; two unrelated variables may just happen to be aligned in a way that makes them appear related. A significance test provides the likelihood that the evidence would support the theory by chance alone and therefore serves one of the functions of the devil's advocate.

1.2 WHICH COMPUTER-INTENSIVE METHOD SHOULD BE USED?

Depending on the nature of the hypothesis, a significance test provides information about one of two types of random influences. The first type of hypothesis is concerned with a characteristic of the population from which a random sample is drawn.[4] The second type of hypothesis is concerned with the relationship among the variables, and in this type of hypothesis test the observations may or may not be a random sample from a population.[5] Frequently, the research question is whether one set of variables is related to another set of variables in a predictable way. For example, theory might suggest that large values of one variable are expected to lead to large values of another variable. A large number of cases in which this relationship is observed would provide evidence consistent with the theory. The concern in such a test is that the larger values of the two variables may be aligned by chance. When testing the relationship between two variables, the alignment of variables relative to each other across cases is the presumed source of random variation.

[4] Conventional parametric statistical tests are concerned with testing this type of hypothesis.

[5] Conventional nonparametric tests are concerned with testing this type of hypothesis. While a very subtle point, the observed relationship among the variables could be interpreted as a random sample of size one from the population of all possible relationships.

Computer-intensive methods can be used to assess the significance of a test statistic under either type of hypothesis. Two generic computer-intensive methods are considered in this book: Monte Carlo sampling and approximate randomization tests. Monte Carlo sampling is used to test a hypothesis concerning the population from which a random sample is drawn. Randomization is used when the hypothesis is concerned with the relationship among variables.

To illustrate the first type of hypothesis, suppose a senatorial candidate commissions an opinion poll in which 54% of a small random sample of registered voters report that they intend to vote for him. If the sample size is small relative to the number of potential voters, the fact that 54% of the sampled voters favor the candidate is not in itself of much interest. The election will be decided by all voters, not just the voters in the sample. The candidate is interested in the sample only insofar as it is informative about voters in general – not just the sampled voters. But a majority of a given random sample of the voters may favor the candidate even if most voters do not. The Monte Carlo sampling method may be used in such a situation to assess the likelihood that the apparently favorable survey result is due to chance.

To illustrate the second type of hypothesis, suppose a self-proclaimed connoisseur of gin claims to be able to distinguish from among eight brands of domestic and imported gins in a blind tasting.[6] And suppose that the connoisseur does, in fact, successfully identify six out of eight brands. He may, of course, have simply been lucky. The randomization method can be used to assess the likelihood that luck, rather than skill, accounts for the gin drinker's apparent success.

1.3 IMPLICATIONS FOR SELECTION OF A TEST STATISTIC

To reiterate, the computer-intensive methods described in this book can be used to assess the significance of virtually any test statistic under the most minimal assumptions. That is, once a test statistic has been selected and its value computed for the observed data, the methods discussed in this book can be used to assess how unusual that value of the test statistic is under an appropriate null hypothesis. Monte Carlo sampling can be used when the hypothesis concerns a parameter of the population from which a random sample has been drawn. A randomization test can be used when the null hypothesis is that one variable is unrelated to another – whether or not the observations constitute a random sample.[7]

---

[6] CONSUMER REPORTS [July 1967 ,p. 381] reports that "despite abundant advertising of the superiorities of imported gins, our experts were baffled when it came to pinning down the origin of a particular sample. They also had difficulty in making marked flavor and aroma distinctions between the brands." Copyright 1967 by Consumers Union of United States, Inc., Mount Vernon, NY 10553. Excerpted by permission.

[7] As Kempthorne [1966, pp. 14-15] points out, randomization tests also do not require

The ability to assess the significance of virtually any well-defined test statistic is liberating. Test statistics can be selected based on which test statistic is best able to discriminate between the null and alternative hypotheses. Out of necessity, researchers have fallen into the habit of selecting a test statistic from the large (but nevertheless limited) set of test statistics for which the sampling distributions are known. Sometimes this has been like fitting a square peg in a round hole. Substantive theory in a field may suggest that the most powerful test statistic is one whose distribution is not known to the researcher.[8,9] However, using the computer-intensive techniques in this book, the researcher can select a test statistic that will maximize the ability to discriminate between the null and alternative hypotheses and be secure in the knowledge that the significance of that test statistic can be assessed.

---

the assumption that the observations are realizations of a continuous random variable. As a practical matter, all empirical observations are measured on a discrete rather than a continuous scale and hence Normal theory derivations of sampling distributions, which assume the variables are continuous, may be suspect.

[8] Some statisticians may not be sensitive to this problem. Statistics texts and journals are full of derivations of sampling distributions for new and novel test statistics. As a practical matter, however, very few empirical researchers have the means to access this literature. An empirical researcher is likely to be familiar with only a few test statistics for which the sampling distributions under sufficiently structured null hypotheses are known. Furthermore, Chung and Fraser [1958], Klauber [1971], Tsutakawa and Yang [1974], Boyett and Shuster [1977], Noreen and Sepe [1981], Bowen, Noreen, and Lacey [1981], and Blanchard, Chow, and Noreen [1986], among others, provide examples of situations in which the conventional parametric sampling distribution of the test statistic suggested by substantive theory is not known.

[9] Pearson [1937] argues that an advantage of conventional parametric methods over randomization methods is that the choice of the most powerful test statistic can be made endogenous with conventional methods. However, Lehmann and Stein [1949] and others have pointed out that the key in making the choice of a powerful test statistic endogenous is in the detailed specification of an alternative hypothesis. Lehmann and Stein derive optimum randomization test statistics for a variety of alternative hypotheses.

CHAPTER TWO

Approximate Randomization Tests

Randomization is used to test the generic null hypothesis that one variable (or group of variables) is unrelated to another variable (or group of variables). Significance is assessed by shuffling one variable (or set of variables) relative to another variable (or set of variables). Shuffling ensures that there is in fact no relationship between the variables. If the variables are related, then the value of the test statistic for the original unshuffled data should be unusual relative to the values of the test statistic that are obtained after shuffling.

2.1 THE BASIC IDEA OF RANDOMIZATION TESTS

A randomization test can be used to test the hypothesis that there is a specified stochastic relationship between one set of random variables and another set of random variables. Usually, the null hypothesis is simply that one set of variables is unrelated to another set of variables. For example, suppose a researcher is interested in whether transfer students perform differently than other students at the University of Washington in the sophomore level introductory managerial accounting course. There are reasons to suspect that the performance of these students might differ systematically. One argument is that only those students who have proven themselves at another school (usually a community college) will be admitted to the university as transfer students. And the preparation for the introductory managerial accounting course provided in a community college is not the same as at the University of Washington. Additionally, some maintain that the best students graduating from high school tend to matriculate directly into the University of Washington. At any rate, the alternative hypothesis is that the performance of transfer students differs from the performance of nontransfer students. The null hypothesis in this case is that performance (measured by grade) in the introductory managerial accounting course is independent of (i.e., is unrelated to) whether the individual is a transfer student.[1]

To test this conjecture, data were collected concerning the grades received by juniors who completed the introductory managerial accounting course during one quarter. These data are displayed in Table 2.1. (At the University of Washington grades are given in increments of 1/10 of a grade point.)

Table 2.1
Grades in introductory managerial accounting

Transfer students: (mean = 2.85)
3.8, 1.8, 1.0, 3.6, 3.3, 2.7, 3.7, 2.5, 3.8, 2.2, 2.5, 3.4, 2.8

Nontransfer students: (mean = 2.57)
4.0, 2.5, 3.6, 2.5, 3.6, 1.7, 2.8, 2.6, 2.7, 2.5, 2.6, 2.2, 2.5, 2.3, 1.3, 3.2, 2.6, 1.0, 2.6, 0.0, 2.8, 3.0, 2.5, 3.1, 4.0, 2.9, 2.7, 3.9, 3.4, 3.6, 3.1, 0.7, 0.7, 2.2

Out of 47 juniors taking the course, 13 were transfer students. The absolute value of the difference between the average grades of the two groups (transfer and nontransfer students) is a natural choice for the test statistic. The average

[1] If x and y are stochastically independent random variables, then for a given data set all permutations of the observed values of y relative to the observed values of x were equally likely. This is the fundamental notion that is exploited in randomization tests.

(mean) grade earned by the transfer students was 2.85, while the mean grade earned by nontransfer students was 2.57, so the difference is 0.28 in favor of the transfer students. Is this difference statistically significant?

The null hypothesis is that grades are unrelated to whether a student is a transfer student. The distribution of the absolute value of the difference in mean grades under this null hypothesis could be constructed in the following manner. First, copy all of the grades onto individual notecards. Second, shuffle the cards. Third, take the first 13 cards from the top of the deck. Arbitrarily label this stack of 13 cards "transfer students" and the stack consisting of the remaining 34 cards "nontransfer students." Fourth, compute and then record the absolute value of the difference between the mean grade for the "transfer students" and the mean grade for the "nontransfer students." Repeat steps two through four many times. In this manner, an empirical distribution can be constructed for the absolute value of the difference in mean grades under the null hypothesis that grades are unrelated to whether a student has transferred from another college. Shuffling the cards and arbitrarily treating the first 13 cards dealt as "transfer students" ensures that the null hypothesis is true, that is, shuffling the cards ensures that grades are unrelated to transfer status. The null hypothesis is rejected if, relative to this empirical distribution, the actual difference of 0.283 is unusual.

The results of shuffling the cards thousands of times are displayed in Figure 2.1.

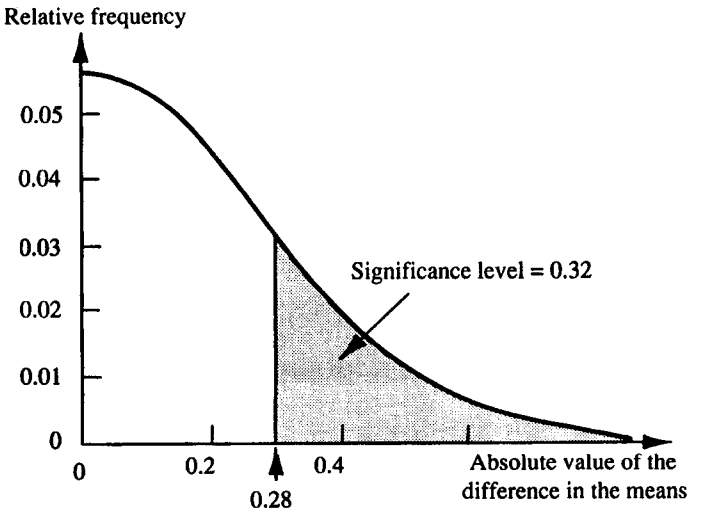

Figure 2.1 Histogram of the absolute value of the difference in mean grades between transfer and nontransfer students

Recall that the value of the test statistic for the original unshuffled data was 0.28. As illustrated in Figure 2.1, the value of the pseudostatistic was at least 0.28 in about 32% of the shuffles. Since it would have been reasonably likely (i.e., probability ≈ 0.32) to have obtained a value of the test statistic as large as 0.28 even though there is no relationship between grades and transfer status, the null hypothesis of no relationship is not rejected.

There are several striking aspects of this approach to assessing the significance of the test statistic. The null hypothesis is very simple – grades are independent of whether a student has transferred from another college. No assumptions are made concerning the distribution of the grades. Furthermore, the data are not a random sample from some population.

2.2 EXACT VERSUS APPROXIMATE RANDOMIZATION TESTS

Since this procedure for assessing the significance of a test statistic involves randomizing the ordering of one variable relative to another, it is called a "randomization" test. When all possible orderings (permutations) of the variables relative to each other are exhaustively listed, the test is called an "exact randomization" test.[2] When the procedure involves randomly shuffling one variable relative to another as in the above example, the test is called an "approximate randomization" test.

The following example should bring out the distinction between an exact and an approximate randomization test.

At a party, a self-proclaimed expert insisted on instructing everyone within hearing concerning the finer points of vodka. He claimed that there were substantial and obvious differences in quality between the finest imported vodkas from Poland and Russia and the premium and budget brands of domestic vodkas.

A skeptic proposed a test. The host just happened to have four different bottles of vodka. One was a Russian import, one a Polish import, one a heavily advertised domestic brand sold at a premium price, and one a generic label budget vodka that the host had poured into a crystal decanter. The vodka connoisseur was shown the four bottles and was then blindfolded. He was told that he would be presented with four different glasses of vodka, one poured from each of the four bottles. His task was to identify which glass was poured from which bottle by taste alone. The host marked the glasses and poured vodka into each of them. The connoisseur then tasted each of the glasses in turn and attempted to

[2] While Edgington [1969] may have originated the terms "exact" and "approximate" randomization tests, Fisher introduced the idea of an exact randomization test in his 1935 book. The term "permutation tests" is often used in the statistics literature to refer to randomization tests. Unfortunately, statisticians also use the term randomization test to refer to a postexperimental procedure to adjust significance levels when a probability distribution is not continuous.

identify which glass was poured from which bottle. The results of the taste test appear in Table 2.2.

Table 2.2
The vodka expert

	Glass 1	Glass 2	Glass 3	Glass 4
Actual contents:	Polish	Premium US	Russian	Budget US
Expert's opinion:	Polish	Premium US	Budget US	Russian

Is the "expert" really an expert?[3]

The null hypothesis is that the expert's opinion is independent of the actual contents of the glass. All of the possible identifications that the expert could have made are listed in Table 2.3. Each one of these possible identifications is a permutation of the order of Polish, Premium US, Russian, and Budget US vodkas.

If the null hypothesis is true and the expert's opinion of the contents of the glasses has nothing to do with the actual contents of the glasses, then each of these permutations was equally likely. Since there are a total of 24 possible permutations and there are seven in which two or more of the glasses are correctly identified, the probability of the expert correctly identifying at least two out of four glasses, given that the expert in fact cannot discriminate among the vodkas, is 0.29 (= 7/24).[4]

[3] Fisher's tea lady is the inspiration for this example.

[4] Some would argue that this test is not interpretable unless there was explicit randomization of the order of presentation of the glasses in the experiment. Whenever possible, such experimental randomization should be followed. Unfortunately, however, many researchers do not have the luxury of randomly assigning treatments in experiments. Does this invalidate the hypothesis test? For example, in the vodka tasting experiment above, the host may attempt to help or hinder the expert by the order in which the vodkas are presented. And, the expert may try to "psyche out" the host. Nevertheless, any such activity by the host and the expert are based on the contents of the glasses and so the test is still a valid test of the null hypothesis that the expert's opinion is independent of the contents of the glasses. The difficulty comes in interpretation of the results. If the null hypothesis is rejected, it may be because the expert successfully psyched out the host rather than because he can discriminate among vodkas. Campbell and Stanley [1963] discuss interpretation of the results of quasi-experiments in which the experimenter has little control over treatments.

Table 2.3
Enumeration of the possible opinions of the vodka expert

	Glass 1	Glass 2	Glass 3	Glass 4	# correct
*	Polish	Premium US	Russian	Budget US	4
*	Polish	Premium US	Budget US	Russian	2
	Polish	Budget US	Premium US	Russian	1
*	Polish	Budget US	Russian	Premium US	2
*	Polish	Russian	Premium US	Budget US	2
	Polish	Russian	Budget US	Premium US	1
*	Premium US	Polish	Russian	Budget US	2
	Premium US	Polish	Budget US	Russian	0
	Premium US	Russian	Budget US	Polish	0
	Premium US	Russian	Polish	Budget US	1
	Premium US	Budget US	Polish	Russian	0
	Premium US	Budget US	Russian	Polish	1
	Russian	Polish	Premium US	Budget US	1
	Russian	Polish	Budget US	Premium US	0
	Russian	Premium US	Budget US	Polish	0
*	Russian	Premium US	Polish	Budget US	2
	Russian	Budget US	Premium US	Polish	0
	Russian	Budget US	Polish	Premium US	0
	Budget US	Polish	Premium US	Russian	0
	Budget US	Polish	Russian	Premium US	1
*	Budget US	Premium US	Russian	Polish	2
	Budget US	Premium US	Polish	Russian	1
	Budget US	Russian	Polish	Premium US	0
	Budget US	Russian	Premium US	Polish	0

2.3 THE APPROACH IN APPROXIMATE RANDOMIZATION TESTS

The foregoing example used the exact randomization method; all possible permutations of the variables relative to each other were listed and the test statistic was computed for each permutation. Exact randomization is feasible, however, with present computer technology only for very small data sets.

Suppose the expert agreed to discriminate among the vodkas by smell alone and 16 different bottles were available from which 16 different glasses of vodka were poured. The number of permutations of 16 glasses of vodka is 16! (= 1 x 2 x 3 x ... x 15 x 16), which is a very large number. Even if one thousand of the permutations could be generated and evaluated each second using a high-speed computer, it would take more than six centuries to exhaust the list of possible permutations!

Fortunately, it is not necessary to exhaust all possible permutations to arrive at a reasonably accurate significance level for a test statistic. Ideally, one would

like to assess the significance of the test statistic relative to the probability distribution of the test statistic, which is generated by the exact randomization method. However, this probability distribution can be approximated to any desired level of precision by sampling. Each shuffle in an approximate randomization test generates one permutation of the variables. A thousand shuffles can be viewed as a sample of size 1000 from the population of all possible permutations. Thus the distribution of the test statistic in 1000 (or however many shuffles) can be used to approximate the exact randomization distribution of the test statistic. Of course, as the number of shuffles increases, the approximation becomes better. The question of how many shuffles is enough is deferred to the next chapter.

Since exact randomization tests are seldom feasible, this book will henceforth be concerned only with approximate randomization tests. Exact and approximate randomization tests differ only in how permutations are generated. Edgington [1980] extensively discusses exact randomization tests and provides a subroutine that can be used to exhaustively list all permutations.

To return to the example of testing the difference in grades between transfer and nontransfer students, it would obviously be a tedious and an error-prone process to actually shuffle a deck of 47 cards 1000 times and compute the absolute value of the difference in means between the first 13 and last 34 cards after each one of those shuffles. Fortunately, a computer can be used to simulate the process of shuffling the cards and computing the difference in means. With a computer, the deck can be shuffled and the test statistic computed hundreds or thousands of times quickly, accurately, and inexpensively. Moreover, for most problems personal computers provide sufficient computing power. Indeed, all of the examples in this book were run on a standard Apple MacIntosh personal computer.

Figure 2.2 illustrates the general approach used in testing hypotheses with the approximate randomization method. The first and perhaps most important step is to select a test statistic that is sensitive to the veracity of the substantive theory. Then, after the data are read, the test statistic is computed. The desired number of shuffles, NS, is set and the various counters are initialized to zero. The algorithm then loops through the randomization procedure, which consists of shuffling the data, computing the test statistic for the shuffled data, and then comparing the value of the test statistic for the shuffled data to the test statistic for the original, unshuffled data. If the pseudostatistic for the shuffled data is greater than or equal to the actual statistic for the original unshuffled data, then one is added to the "nge" counter.[5]

[5] "nge" is an acronym for "number greater than or equal to."

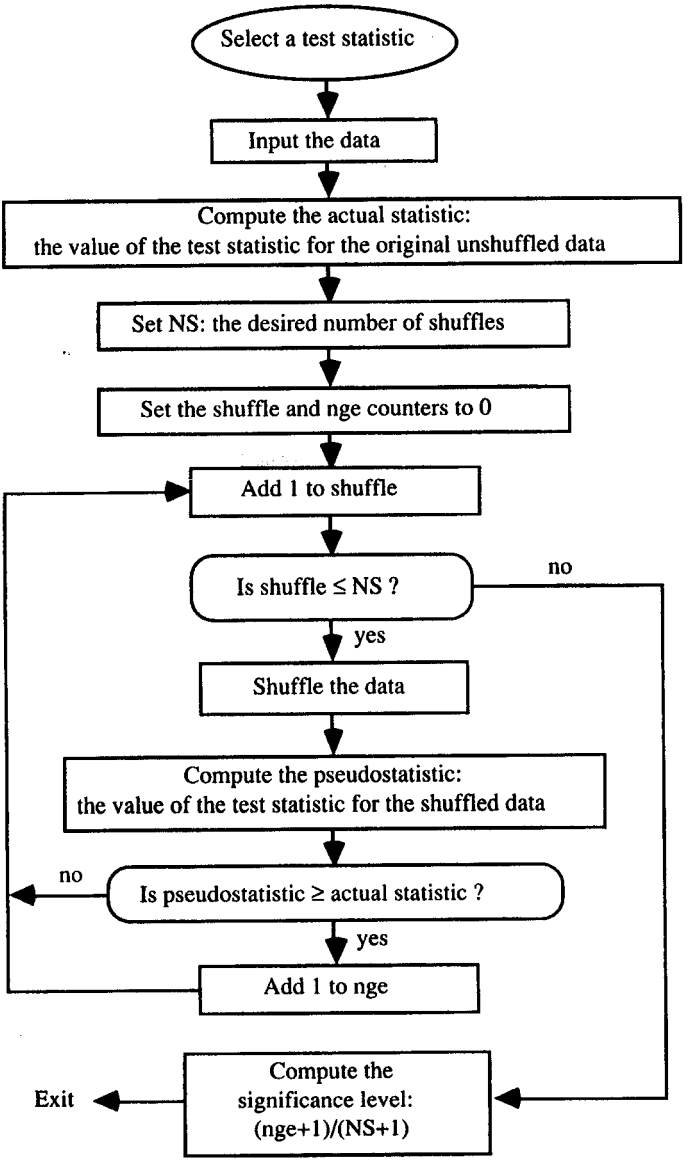

Figure 2.2 Flowchart for an approximate randomization test

Depending on the hypothesis and substantive theory, the data can be shuffled in a number of ways. Most commonly, there is one dependent variable and one or more explanatory variables, and the null hypothesis is that the dependent variable is in fact unrelated to the supposed explanatory variables. In this situation, the dependent variable is shuffled relative to the explanatory variable(s).[6] This procedure ensures that the variables are unrelated to each other.[7]

The final step is to compute the ratio (nge+1)/(NS+1), which is the significance level of the test. The null hypothesis is rejected if the significance level (nge+1)/(NS+1) is less than or equal to the specified rejection level for the test. This ratio requires some explanation. The ratio nge/NS is the frequency with which the pseudostatistic for the shuffled data was greater than or equal to the actual statistic for the unshuffled data. The significance level of the test, however, is the ratio (nge+1)/(NS+1). Why is 1 added to both nge and NS when the significance level is computed? Without going into technical details at this point, this minor adjustment ensures that the test is valid, that is, the probability of rejecting the null when it is true is no greater than the rejection level specified for the test.

A general template for testing the generic hypothesis that two variables are unrelated is listed in the Program Appendices at the end of the book. There are three Program Appendices – one each for BASIC, FORTRAN, and PASCAL. Refer to the appendix for whichever language you are most familiar with. The sections of the template that are in boldface type need to be modified for whatever data and test statistic are at hand. The most significant of these modifications is that code would have to be written to compute the test statistic.

As an example, this template was used to test the difference in mean grades between transfer and nontransfer students in introductory managerial accounting. (See Program 2.1 in one of the Program Appendices.) In this program, the dependent variable is the grade of a student and the explanatory variable is the student's transfer/nontransfer status. If the student is a transfer student, the explanatory variable is coded 1; if the student is not a transfer student, it is coded 0. The data are stored in a file called "transfer data," which is listed at the end of the Program 2.1.

The program begins by opening the transfer data file. The arrays y and x are dimensioned to allow storage of 47 observations, one for each student. The program reads the grades and status of each student and computes the difference in the mean grades between the transfer students (x = 1) and the nontransfer students (x = 0). The program requests as input the number of shuffles, NS. (When debugging a program, a small number of shuffles should be specified.)

[6] There are situations, however, in which it is desirable to exert greater control over the shuffling. In Section 2.6, stratified shuffling is described.

[7] More than one dependent variable can be easily accommodated by shuffling an index vector and then using the shuffled index vector as the index for the dependent variables.

The program then begins the shuffling procedure. The grades (stored in y) are shuffled. Shuffling y alone results in shuffling y relative to x. After shuffling, each of the grades actually given in the course will have been randomly assigned to a student. The program then computes the difference in the mean of the randomly assigned grades between the transfer students (x = 1) and the nontransfer students (x = 0). If this pseudodifference is at least as large as the actual difference in the means, one is added to the nge counter.[8]

The process of shuffling the grades before computing the pseudodifference in the means ensures that the grades are unrelated to transfer/nontransfer status. Hence, the probability distribution of the pseudodifferences for the shuffled data is the distribution of the test statistic under the null hypothesis that grades are unrelated to transfer status.

The final step in the program is to compute and print the significance level of the test. The significance levels differ slightly between the BASIC, FORTRAN, and PASCAL versions of the program due largely to differences in the random number generators that are used in the shuffling algorithms. Rather than referring to all three versions, I will make a practice in the text of referring to only the BASIC program results. In this example, the significance level from the BASIC program is 0.319. This means that in 318 of the 999 shuffles, the difference in the pseudo means was at least as large as 0.283, the actual value of the test statistic for the unshuffled data.

Also note the three lines in the program listing that follow the significance level. These lines will be more fully explained in Appendix 3A of the next chapter. Briefly, phi is the (unknown) significance level for an exact randomization test run on the same data. This exact significance level is estimated using an approximate randomization test. The printout from running the program indicates that, given the estimated significance level of 0.319 after 999 shuffles, the probability that the exact significance level is less than or equal to any of the conventional rejection levels is essentially zero. Thus, even if the shuffling were to continue and the approximation to the exact significance level were made more precise, it is extremely unlikely that the basic conclusion (i.e., the null hypothesis is not rejected) would be overturned.

Many (perhaps most) hypotheses in which researchers are interested can be tested using this simple template. The template can be used, in conjunction with any test statistic, to test the null hypothesis that two variables are unrelated. With only minor adjustments, this template can be used to test the hypothesis that one set of variables is unrelated to another set of variables. In much research, this is precisely the hypothesis that the researcher would like to test.[9]

[8] Each time the data are shuffled, the program counts the number of transfer and nontransfer students. This is not really necessary since the number of transfer and nontransfer students never changes. The counting slows down execution speed. Usually, however, execution speed is not much of an issue and simpler programs are to be preferred to faster but more complex programs.

Several examples of approximate randomization tests follow.[10] The intent is to illustrate how approximate randomization tests can be used to test hypotheses in a variety of situations. An important advantage of the randomization method over conventional techniques is its generality. Once the method is understood, it can be used to test an almost unbelievable variety of research hypotheses.[11]

2.4 EXAMPLE: VOTER TURNOUT IN THE 1844 PRESIDENTIAL ELECTION

It has been suggested that citizens will be most inclined to vote in close elections. The 1844 U.S. presidential election was the closest that had been held up to that time, with the exception of the 1824 election which had been decided in the House of Representatives. In the 1844 campaign, the Democratic candidate James Polk was pitted against the Whig candidate Henry Clay. While the popular vote was very close (1,338,464 for Polk versus 1,300,097 for Clay), the vote in the electoral college was 170 for Polk versus 105 for Clay.

The US presidential election is decided in the electoral college rather than by the popular vote. Within each state, there is a winner-take-all rule; whoever wins the popular vote in the state gets all of the state's electoral votes. Thus, the incentives to vote may well differ from state to state, depending on how close the election is in each state. In states where the election is expected to be close, voters should be more motivated to vote than in states where the election is not expected to be close.

Data concerning the voter turnout (or participation rate) and the spread between the percentages of the popular vote obtained by Polk and Clay in each state are displayed in Table 2.4. The smaller the spread, the closer the election was in the state.

Assuming voters had some ability to forecast how close the election was going to be in their own states, there should be a negative relationship between participation rates and the actual vote spread. The data, which are plotted in Figure 2.3, exhibit such a negative relationship. Roughly speaking, the participation rate does appear to decline as the spread between the votes for the two presidential candidates increases. How likely is it that such an apparent relationship would have occurred by chance?

[9] Even more generally, similar procedures can be used to assess the significance of any test statistic under the null hypothesis that one set of variables is stochastically related in a specified way to another set of variables. An empirical distribution for the test statistic can be generated by ensuring that the stochastic relationship between the sets of variables is as specified in the null hypothesis.

[10] Edgington [1980] provides additional examples.

[11] Randomization tests are not appropriate, however, when the researcher is concerned with drawing an inference about a population parameter based on a random sample. In those cases, conventional parametric or Monte Carlo sampling techniques must be used.

Table 2.4
Voter participation in the 1844 presidential election

State	Participation[a]	Spread[b]
Maine	67.5	13
New Hampshire	65.6	19
Vermont	65.7	18
Massachusetts	59.3	12
Rhode Island	39.8	20
Connecticut	76.1	5
New York	73.6	1
New Jersey	81.6	1
Pennsylvania	75.5	2
Delaware	85.0	3
Maryland	80.3	5
Virginia	54.5	6
North Carolina	79.1	5
Georgia	94.0	4
Kentucky	80.3	8
Tennessee	89.6	1
Louisiana	44.7	3
Alabama	82.7	18
Mississippi	89.7	13
Ohio	83.6	2
Indiana	84.9	2
Illinois	76.3	12
Missouri	74.7	17
Arkansas	68.8	26
Michigan	79.3	6
National average	74.9	9

[a]The percentage of eligible voters who voted in the presidential election.

[b]The absolute value of the difference in the percentage of the total vote obtained by Polk and Clay in the state.

To be precise about this question, it is necessary to define a test statistic which encapsulates the notion that participation rates should be negatively related to the actual vote spread in the election. The correlation coefficient is a natural choice in this case. The correlation between participation rates and vote spreads is –0.374. By convention, a large value of the test statistic should be viewed as evidence that is consistent with the alternative hypothesis. Since a negative correlation is expected, the test statistic will be the negative of the correlation, or just 0.374. Defining the test statistic as the negative of the

correlation allows us to use the standard templates without modification. Positive values of the test statistic (i.e., negative correlations) are consistent with the alternative hypothesis, while negative values of the test statistic (i.e., positive correlations) are not consistent with the alternative hypothesis.

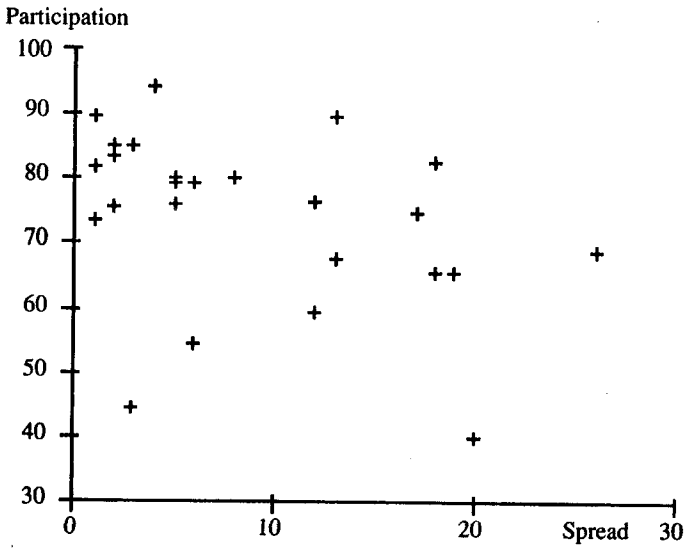

Figure 2.3 Participation rates versus vote spread in the 1844 presidential election

The null hypothesis is that the participation rate is unrelated to how close the election is (i.e., each permutation of participation rates relative to vote spreads was equally likely). The distribution of the correlation coefficient under this null hypothesis can be approximated by shuffling the participation rate relative the vote spread many times and, after each shuffle, computing the correlation coefficient for the shuffled data. See Program 2.2 in the Programs Appendix of your choice for a listing of a program to carry out this process. In the case of the BASIC program, the significance level of the test was 0.036, i.e., on only 35 of the 999 shuffles was the correlation negative and as large as 0.374. The three lines following the significance level in the printout indicate the probabilities that the exact significance level is less than or equal to 0.01, 0.05, and 0.10. In this case, there is a great deal of confidence that the exact significance level (i.e., the significance level that would be obtained from an exact randomization test on the same data) would be less than or equal to 0.05 or 0.10. In contrast, there

is very little confidence that the exact significance level is less than 0.01. This is, of course, to be expected since the estimated significance level of .0.036 is greater than 0.01. Therefore, the null hypothesis that participation rates and vote spread are unrelated can be confidently rejected at the 0.05 or 0.10 level, but not at the 0.01 level.

2.5 EXAMPLE: SLAVEHOLDINGS AND THE VOTE FOR SECESSION

Lipset [1960] recounts the events that led to the secession of the Confederate states from the Union in 1861. Three to six months after the election of Lincoln as President in the autumn of 1860, seven southern states held referenda in which voters elected county delegates to state conventions which were to consider seceding from the Union. Lipset reports that

> *These convention-delegate elections were hotly contested in most Southern states, and the results were closer than many realize, with the Union forces getting over 40 per cent of the vote in many states* [p. 642].

Lipset classified the vote for secessionist delegates by the relative slave holdings in the counties. Slavery was an important point of friction between the North and the South. And, as Lipset notes, "in all the southern states... the proportion of slaves in the population served to differentiate the wealthier from the poorer counties...." Therefore, to the extent that the Civil War grew out of economic conflicts or disputes over slavery, the relative proportion of slaves in a county may serve to predict the vote on secession. Indeed, one would expect that the higher the slave holdings, the more likely it is that a county would have voted for secession. Indeed, this is what happened, as is evident in Table 2.5.

Table 2.5
Actual vote by county in the 1861 vote on secession

		Secession	Union	Total
Relative slave holdings	High	130 (72%)	51 (28%)	181
	Medium	92 (60%)	61 (40%)	153
	Low	75 (37%)	128 (63%)	203
	Total	297 (55%)	240 (45%)	537

How would you go about testing the conjecture that the higher the relative slave holdings, the more likely it is that a county would have voted for secession and the lower the slave holdings, the more likely it is that a county would have

voted for the Union? What test statistic would you use? How would you assess the significance of that statistic?

One's first impulse in this situation might be to perform a chi-squared test of the 3 x 2 contingency table. The chi-squared test is based on comparing the actual counts in the cells to the counts that would be expected if the vote by county were independent of the relative slave holdings. However, the chi-squared test is unable to distinguish between departures from expectations that are and are not in the directions expected.[12] Hence, a test based on the chi-squared statistic is not as powerful as it could be.

Consider the high relative slave holding counties. If there were no relationship between relative slave holdings and the vote, we would expect to see 55% of the counties voting for secession and 45% voting for the Union. These are the relative proportions of all counties voting for secession and the Union. Thus if the vote is independent of slave holdings, then the expected counts in the cells (rounded to the nearest whole numbers) would be as indicated in Table 2.6.

Table 2.6
Expected vote by county in the 1861 vote on secession if the vote for secession or Union was independent of relative slave holdings:

		Secession	Union	Total
Relative	High	100 (55%)	81 (45%)	181
slave	Medium	85 (55%)	68 (45%)	153
holdings	Low	112 (55%)	91 (45%)	203
	Total	297 (55%)	240 (45%)	537

However, we would suspect that the high relative slave holding counties would be more inclined to vote for secession than other counties. Thus, for the high relative slave holding counties, there should be more than 100 counties in the secession column and less than 81 counties in the Union column. An index of how well the data agree with a priori reasoning for the high relative slave holding counties could be constructed as follows:

$$\text{High agreement} = (\text{actual voting for secession} - 100) + (81 - \text{actual voting for the Union})$$

Similarly, an index could be constructed for the low relative slave holding counties as follows:

[12] Moreover, the chi-squared test assumes that departures from expectations are Normally distributed.

Low agreement = (112 – actual voting for secession)
+ (actual voting for the Union – 91)

Both of these indices will be positive and large if the reasoning is correct.

Suppose there is a dividing line for relative slave holdings and above the dividing line the counties tend to vote for secession and below the line they tend to vote for the Union. The medium counties present a problem since the dividing line could be above, below, or in the midst of those counties. One approach would be to simply throw out the counties. However, if the dividing line were really above or below the medium counties, throwing out the medium counties could lead to failure to reject a false null hypothesis. That is, throwing out the medium counties may lead to a loss of power. A better approach would be to construct an index for the medium counties that would be sensitive to deviations in either direction. A natural choice of index would be:

Medium deviations = ABS(actual voting for secession – 85)
+ ABS(actual voting for the Union – 68)

Finally, combining the indices yields an interpretable test statistic:

Deviations as expected = high agreement + medium deviations
+ low agreement

For the actual data, the value of this test statistic is:

Deviations as expected = 60+14+74 = 148

That is, 148 out of 537 counties deviated from expectations under the null hypothesis *in the directions expected under the alternative hypothesis.*

This is the test statistic; now how can its significance be assessed? The first step is to recognize what the underlying data really look like. There are 537 counties and two variables – a county's relative slave holdings and its vote. This data can be organized as illustrated in Table 2.7.

There are actually 537 rows in this data set, one for each county. In the data file listed below, the first column is the county's relative slave holdings: 1 represents high, 2 represents medium, and 3 represents low. The second column is the county's vote: 1 represents a vote for secession and 2 represents a vote for the Union. For example, there are 130 identical entries coded 1,1 to represent the 130 high relative slave holding counties that voted for secession.

The procedure for assessing the significance of the test statistic is the same as before. One of the variables (either slave holdings or vote) is shuffled. After

each shuffle, the contingency table is reconstructed and the value of the test statistic is recomputed. As it turns out, in 999 shuffles, there was no shuffle on which the test statistic was as large as it was for the original unshuffled data. Therefore, the null hypothesis that the vote on secession was unrelated to relative slave holdings is rejected. BASIC, FORTRAN, and PASCAL programs to accomplish this process are reproduced in the Programs Appendices. These programs took much longer to execute than the other programs illustrated in this book because there are 537 observations in the data file that must be shuffled.

Table 2.7
Listing of relative slave holdings data

1,1	*there are 130 of these entries*
1,2	*there are 51 of these entries*
2,1	*there are 92 of these entries*
2,2	*there are 61 of these entries*
3,1	*there are 75 of these entries*
3,2	*there are 128 of these entries*

2.6 EXAMPLE: DO YOU GET WHAT YOU PAY FOR?

A report on skin moisturizers appeared in the November 1986 issue of CONSUMER REPORTS. The method used to compile the ratings of the skin moisturizers was described as follows:

> *We didn't test the moisturizers in our labs, since reading labels and analyzing ingredients couldn't tell us how the products would perform on a variety of skin types or which products people would really prefer. We turned instead to a panel of 600 female readers.... We sent each panelist two products packed in plastic bottles marked only with a red or white dot. We told them to use one product for a week, then switch to the other... In a questionnaire we asked each panelist to rate how well the products performed.... Our statisticians averaged the scores, then ranked the products according to the overall judgments.*[13]

The results of this survey are displayed in Table 2.8.

Hopefully, it is generally true that if a consumer buys a higher priced brand of a particular product, the quality of the product is higher as well. If the retail

[13] Copyright 1986 by Consumers Union of United States, Inc., Mount Vernon, NY 10553. Excerpted by permission from CONSUMER REPORTS, November 1986, p. 734.

market is functioning efficiently, high-priced inferior brands should be driven from the marketplace. An interesting question is whether the retail market is indeed efficient or whether price and quality are unrelated. The above data clearly indicate that the market is not completely efficient. For example, the highest priced brand was rated third from the bottom in terms of quality by users. Nevertheless, it is not immediately obvious that the market is completely inefficient either. There might be some positive relationship between price and quality. How would one approach the problem of testing whether the market is inefficient (i.e., there is no relationship between price and quality)?

Table 2.8
Price per ounce of skin moisturizers in order of descending estimated quality*

Rank	Price per oz.	Rank	Price per oz.
1	$0.83	25	$1.65
2	0.23	26	3.43
3	1.52	27	0.59
4	1.91	28	0.42
5	0.25	29	0.40
6	0.10	30	1.56
7	0.12	31	0.24
8	0.24	32	0.26
9	0.33	33	1.69
10	0.19	34	0.10
11	0.26	35	0.62
12	0.26	36	0.25
13	0.28	37	3.89
14	0.11	38	0.17
15	0.12	39	1.65
16	0.12	40	0.38
17	0.30	41	0.45
18	0.45	42	1.30
19	0.24	43	3.07
20	0.22	44	1.42
21	0.11	45	2.11
22	0.25	46	6.10
23	3.33	47	4.29
24	1.31	48	0.25

Those who are used to classical statistical methods might approach this problem using a variety of methods. Some would conduct a test of the correlation between quality and price. Either a product-moment (Pearson) or rank (Spearman) correlation could be computed. However, there are drawbacks to both of these approaches. If the product-moment correlation is used, it is implicitly assumed that price is a linear function of quality rank. For this to be true, it would have to be the case that the difference in quality is about the same between any two brands that are adjacent in the table. For example, the difference in quality between the two highest rated brands and between the two lowest rated brands may be quite different. The rank correlation, on the other hand, converts prices into ranks and thereby throws away valid information about differences in prices. There is only a penny difference between the prices of the two cheapest brands, but there is a $1.81 difference between the two most expensive brands. When a rank correlation coefficient is used, this information is ignored.

What test statistic would be more appropriate? If a customer has selected a brand and the market is efficient, there should not be much gain expected from searching for a brand that is better and less expensive. It may be possible to find a better brand, but it would be more expensive. And it may be possible to find a cheaper brand, but it would not be as good. To simplify the discussion, suppose there are only five brands and, in order of decreasing quality, they cost $2, $5, $3, $4, and $1 per ounce [see Table 2.9]. The $2 brand dominates the $5, $3, and $4 dollar brands; it is better and cheaper. The $3 brand dominates the $4 brand. If a consumer had selected the $2 brand, the expected monetary gain to searching for a dominating brand would be zero since no brand dominates the $2 brand. If, on the other hand, the $5 brand had been chosen, the expected monetary gain from searching for a dominating brand would be $0.75, since there would be a one in four chance of saving $3 by buying the better $2 brand. See Table 2.9 for a listing of the expected gains from searching for a dominating brand.

Table 2.9
Illustration of the test statistic designed to gauge market efficiency

Brands listed in order of decreasing quality	Expected gain from further searching
$2	$0
$5	($5–$2)x0.25 = $0.75
$3	($3–$2)x0.25 = $0.25
$4	($4–$2)x0.25+($4–$3)x0.25 = $0.75
$1	$0

The average expected gain across all brands in Table 2.9 from further searching is \$0.35 [= (\$0+\$0.75+\$0.25+\$0.75+\$0)/5]. If the market is efficient, the average expected gain should be small (and less than the cost of further searching). If the market is inefficient, this number will be relatively large.

This test statistic appropriately uses both the rank and price information in a way that has a convenient economic interpretation. However, it would be very difficult to analytically derive a conventional small sample distribution for this test statistic. There is no difficulty, however, in assessing the significance of the test statistic under the null hypothesis that price and quality rankings are unrelated. A program to accomplish this is listed as Program 2.4 in the Program Appendices. The average expected gain for the actual brands of moisturizers in Table 2.8 is about \$0.477 per ounce. Remarkably, in 999 trials, there was no trial in which the expected gain was this large. This means that the expected gain from switching products using the real data is greater than if price and quality were completely unrelated! This evidence is consistent with a perverse market – one in which quality tends to go down as price goes up.

2.7 STRATIFIED SHUFFLING

In the example earlier in this chapter in which the grades of transfer students were compared to the grades of nontransfer students, one instructor assigned all of the grades. Suppose, however, that grades had been collected from a number of different instructors. It is conceivable that the nontransfer students have superior information concerning instructors' grading practices and will tend to enroll in classes taught by instructors with the most liberal grading policies. In that case, it would be desirable for the researcher to control for possible differences in the grading practices of the instructors. There might appear to be a significant difference in the performance between transfer and nontransfer students because some nontransfer students purposefully select instructors who assign higher grades.

Table 2.10 contains data from five instructors concerning the grades of juniors (18 transfer students and 39 nontransfer students) who completed the second quarter introductory financial accounting course at the University of Washington. The mean grade earned by the transfer students was 2.29 and the mean grade of the nontransfer students was 2.37 (a difference of 0.08).

Since there is reason to believe that grades may not be independent of the instructor, we would not want to indiscriminately shuffle grades across instructors relative to transfer status. To control for the effects of differing grading policies across instructors, grades could be shuffled within each instructor's class. In this way, the distribution of the test statistic could be generated under the null hypothesis that within classes, grades are unrelated to whether an individual is a transfer student.

Table 2.10
Grades attained in financial accounting by juniors

	Instructor					
	A	B	C	D	E	
Transfer students	2.0, 3.0, 2.2, 2.1, 2.2	2.3, 2.8	2.8	2.2, 2.0, 1.1, 2.5, 2.6	0.7, 3.5, 2.4, 2.3, 2.5	n = 18 mean = 2.29
Non-transfer students	3.2, 2.9, 2.0, 2.2, 2.1, 1.4	3.3, 2.6, 1.9, 2.2, 1.4	2.9, 3.3, 2.5, 2.4, 2.3, 2.8, 1.3	3.6, 0.7, 3.5, 2.6, 1.6, 3.2, 1.6, 0.9, 1.9, 1.8, 1.8, 3.6, 3.1	1.5, 3.0, 2.2, 3.0, 2.1, 4.0, 1.9, 2.1	n = 39 mean = 2.37
	n = 11 mean = 2.30	n = 7 mean = 2.36	n = 8 mean = 2.54	n = 18 mean = 2.24	n = 13 mean = 2.40	n = 57 mean = 2.34

To test the statistical significance of the difference in mean grades of 0.08, the grades were shuffled relative to the transfer/nontransfer status variable within each instructor's class. This stratified shuffling is carried out in Program 2.5, which is listed in the Program Appendices. In the BASIC version of the program, in 717 out of 999 trials the difference between the grades of the transfer and nontransfer students was at least as large as 0.08 grade point, so the actual difference is not statistically significant (i.e., the significance level is 0.718). That is, even if grades have nothing to do with whether a student has transferred from another institution, the mean grades would differ by as much as 0.08 about 72% of the time.

Stratified shuffling is appropriate whenever there is reason to believe that the value of the dependent variable depends on the value of a categorical variable that is not of primary interest in the hypothesis test. In the above example, there was concern that grades might differ between instructors, so the observations were stratified by instructor. This effectively controls for the effects on students' grades of this nuisance explanatory variable.

Several nuisance categorical explanatory variables can be controlled for simultaneously. Suppose, for example, that a student's gender may influence the grade the student receives. If the effect of transfer/nontransfer status on grades is of interest, but the effects of gender or instructors on grades are not of interest, it is desirable to control for those nuisance explanatory variables. Gender and instructor could be simultaneously controlled for by shuffling grades relative to transfer/nontransfer status within classes and gender. To illustrate, the grades of all female students who have a specific instructor would be shuffled relative to their transfer/nontransfer student status. Then the grades of all male students with the same instructor would be shuffled relative to their transfer/nontransfer student status. The process of shuffling within gender would be repeated for each instructor.

This method of controlling for nuisance categorical explanatory variables by stratified shuffling effectively controls for any relationship that might exist between the dependent variable and nuisance categorical variables.[14]

2.8 REGRESSION AND ANOVA

Two of the most common statistical procedures are regression and ANOVA. Least-squares regression is ordinarily used to estimate a model in which the dependent variable is a linear function of the explanatory variables. ANOVA can be viewed as a version of least-squares regression in which all of the explanatory variables are categorical. The significance of test statistics produced by regression or ANOVA can be assessed using approximate randomization methods. In this section this application of the approximate randomization method is briefly discussed. The discussion is brief because the applications should be straightforward. I assume familiarity with regression and ANOVA.

Each of the estimated coefficients of a regression model, as well as the overall fit of the model (the r^2), is a test statistic whose significance can be assessed using the approximate randomization method.[15] A variety of different hypotheses can be tested, depending upon what is shuffled relative to what. The simplest procedure is to shuffle the dependent variable relative to the fixed matrix of explanatory variables. This provides significance levels for each of the test statistics under the null hypothesis that the dependent variable is unrelated to the explanatory variables.[16] When the approximate randomization method is used

[14] In contrast, the most frequently used conventional methods assume that the effect of the categorical variable is confined to a shift in the mean of the dependent variable.

[15] When the matrix of explanatory variables is fixed it doesn't make any difference whether the estimated coefficients or their t statistics are used as the test statistics in an approximate randomization test. This is because the t statistics are directly proportional to the coefficients when the covariance matrix is fixed.

[16] Ordinarily, the null hypothesis in a conventional regression test is that the coefficients are zero. This is different from the null hypothesis that the dependent variable is independent of the explanatory variables. Except for the test of the significance of the

to assess significance, there is no need to be concerned with whether the residuals are Normally distributed. However, other econometric problems remain. Heteroscedasticity, a nuisance form of dependence between the dependent and explanatory variables, can result in inappropriately rejecting the null hypothesis.[17] Multicollinearity among the explanatory variables still may make it difficult to unambiguously determine their relative importance. Nevertheless, approximate randomization tests are a useful alternative to the usual significance tests – particularly for small samples and where the assumption of Normal residuals is questionable.

Randomization can also be used to assess the significance of test statistics in an analysis of variance (ANOVA) table. For example, a two-way ANOVA would partition the variance in the dependent variable into two main effects: an interaction effect and a portion that is apparently not due to any of the explanatory variables.[18] The test statistics could be the usual F statistics. The significance of the F statistics, under the null hypothesis that the dependent variable is unrelated to the explanatory variables, can be assessed by shuffling the dependent variable relative to the explanatory variables. In addition, a great deal of control is possible by shuffling within categories of the explanatory variables. For example, suppose the effectiveness of a drug is in question, and an experiment was conducted in which the drug was administered to one set of patients and a placebo to another set of patients in five different hospitals. The possible nuisance effect of the hospitals themselves can be effectively controlled by shuffling the measure of effectiveness relative to the real/placebo status of the treatment by patients *within* each hospital.

2.9 SUMMARY

In this chapter the approximate randomization method of assessing the significance of a test statistic was introduced. This method can be used to test the hypothesis that a dependent variable is unrelated to the explanatory variable(s).

intercept, however, the results will often be essentially the same for conventional and approximate randomization tests of the regression model. To see the difference in a test of the intercept, suppose a regression is run in which there is only an intercept term (i.e., there is only one explanatory variable and its value is always 1). The intercept will then be simply the mean of the dependent variable. No matter how the dependent variable is shuffled, its mean will always be the same and hence the intercept will always be the same. Hence, the null hypothesis of independence will never be rejected if a randomization test is used. On the other hand, if the mean of the dependent variable is not zero, the conventional significance test may well result in the rejection of the null hypothesis that the intercept term is zero.

[17] I am indebted to Vic Bernard for pointing this out.

[18] Alternatively, the absolute value of the deviation between values of the dependent variable and its mean could be used as the basis of the analysis. Such an analysis might be easier to interpret than ANOVA, which is based on squaring the deviations.

Such a test is accomplished by shuffling the dependent variable relative to the explanatory variable(s) and recomputing the test statistic for the shuffled data. This shuffling ensures that the dependent variable is unrelated to the explanatory variable(s) and hence that the null hypothesis is true.

The distribution of the test statistic under this hypothesis is approximated by shuffling the data and recomputing the test statistic many times. The significance of the actual test statistic for the original unshuffled data is assessed relative to this empirically generated distribution. The null hypothesis is rejected if the actual value of the test statistic for the original unshuffled data is unusually large relative to the values of the test statistic that would have been expected if the dependent variable is in fact unrelated to the explanatory variable(s).

These ideas were illustrated by testing hypotheses in a variety of situations.

In Appendix 3A, it is demonstrated that approximate randomization tests are valid, that is, the probability of falsely rejecting the null hypothesis when it is in fact true is no greater than the prespecified nominal rejection level for the test.

HISTORICAL PERSPECTIVE

Fisher [1966] originated the notion of a randomization test in his 1935 book The Design of Experiments. In the two decades that followed publication of that book, randomization tests attracted the attention of theoretical statisticians such as Pitman [1937,1938], Pearson [1937], Scheffe [1943], Noether [1949], Lehmann and Stein [1949], and Hoeffding [1951, 1952]. Application of randomization tests to testing real hypotheses was impeded, however, by the high cost of manually recomputing the test statistic many times. Statisticians finessed this practical difficulty by constructing probability tables for generic data (i.e., ranks and nominal categories). The end result of this effort is the broad range of nonparametric tests found in such standard references as Siegel [1956] and Hollander and Wolfe [1973]. All truly nonparametric tests are special cases of exact randomization tests in which observations are, or have been replaced by, ranks or nominal categories.

APPENDIX 2A

THE POWER OF APPROXIMATE RANDOMIZATION TESTS

2A.1 INTRODUCTION

In this appendix, the performance of approximate randomization tests is compared to the performance of conventional parametric tests using artificial data. Both a conventional parametric test and an approximate randomization test are applied to the same data and the frequencies with which the null hypothesis is rejected are compared. The test that rejects the null hypothesis the most frequently when it is in fact false is the more powerful test. In a nutshell, for the data considered here, there is virtually no loss in power when an approximate randomization test is used instead of a conventional parametric test. This is true even when the data are generated to conform to the assumptions of the parametric test. Hoeffding [1952] demonstrates in greater generality that randomization tests are asymptotically (i.e., as sample size becomes very large) as powerful as related conventional parametric tests when the assumptions underlying the conventional parametric tests are true.

Two common conventional parametric tests are considered: a t test of the difference in means between two groups and a t test of the correlation between two variables. These two tests are special cases of common multivariate tests. The t test of the difference in means is the univariate case of one-way analysis of variance (ANOVA). The t test of a correlation is, loosely speaking, the univariate case of multiple regression.

At the outset it should be reiterated that the null hypotheses are different for the two methods of assessing the significance of a test statistic. When randomization is used, the null hypothesis is that the dependent variable is unrelated to the explanatory variable(s); or, more precisely, all permutations of the dependent variable relative to the explanatory variables were equally likely. When a conventional parametric method is used, the null hypothesis is that the data are a random sample from a population with certain specified characteristics.

2A.2 TESTS OF THE DIFFERENCE IN THE MEANS OF TWO GROUPS

A conventional t test is often used when a researcher is interested in the difference in the means between two groups. The conventional parametric pooled variance t test of the difference between the means of two groups is valid if, in effect, the two groups are random samples from the same Normal population. In addition, the Central Limit Theorem ensures that the t test is asymptotically valid if the two groups are random samples from any distribution with finite mean and variance.

Artificial data sets consisting of M observations each were constructed by generating random standard Normal scores and adding a constant d to the scores for the first half of the M observations. The test statistic is the difference in the means between the first half and the last half of the observations. By construction, the expected value of this test statistic is d.

Two values of the constant, d = 0.0 and d = 0.5, and two sample sizes, M = 10 and M = 100, were used. For each of the two sample sizes, 1000 basic data sets were independently generated. Each basic data set was used to generate two derived data sets: one for the case where d = 0.0 and one for the case where d = 0.5. The only difference in the derived data sets is the constant amount d added to the first half of the observations. For each derived data set, a pooled variance t test of the difference in means was conducted. The frequencies with which the null hypothesis was rejected at the 0.10 level are reported in Table 2A.1. For example, when the derived data sets consisted of 10 observations each with the constant d set at 0.5, the null hypothesis was rejected 31.3% of the time at the 0.10 level.

In addition, the approximate randomization method was used to assess the significance of the difference in the means for each derived data set. This was accomplished by shuffling the observations and computing the difference in means between the first and last half of the shuffled observations. If the difference in means for the shuffled data was at least as large as the difference in means for the unshuffled data, one was added to the nge counter. This process was repeated 99 times for each data set. The null hypothesis was rejected for a data set if, at the end of 99 shuffles, the ratio (nge+1)/100 was less than or equal to 0.10.

The approximate randomization method is used to test the null hypothesis that the score is independent of the group (i.e., first half or last half) to which the observation belongs. The frequencies with which this null hypothesis was rejected are also reported in Table 2A.1. For example, when the derived data sets consisted of 10 observations each with the constant d set at 0.5, the null hypothesis was rejected 29.7% of the time at the 0.10 level.

By construction, the parametric t test is valid for this data. And, as shown in Appendix 3A, an approximate randomization test is valid for any data and any test statistic. Therefore, it should not be surprising that the null hypothesis is rejected about 10% of the time at the 0.10 level when the null hypothesis is true (i.e., d = 0.0). All of the rejection frequencies are within a 90% confidence interval surrounding 0.10.[19]

When the null hypothesis is false (i.e., d = 0.50), the parametric t test rejects the null hypothesis slightly more frequently than does the approximate randomization test. However, it should be noted that the differences in the rejection rates

[19]Rejection of the null hypothesis is a binomial event. Using the Normal approximation to the binomial distribution, the approximate 90% confidence interval surrounding the rejection frequency 0.100 is $0.100 \pm 1.645[(.10)(.90)/1000]^{1/2}$ or {.084,.116}.

are not statistically significant.[20] Thus, there is no evidence that the parametric test is practically more powerful, even when its assumptions are satisfied.

Table 2A.1
Frequency with which the null is rejected at the 0.10 level: Tests of the difference in two means

	Sample Size	
d = 0.50	M = 10	M = 100
Pooled variance t test	0.313	0.891
Randomization	0.297	0.886
d = 0.00		
Pooled variance t test	0.107	0.089
Randomization	0.106	0.086

Note: The tests were conducted on 1000 independently generated data sets, each of which consisted of M standard Normal scores. The constant d was added to the first M/2 observations. The test statistic was the difference in the means between the first M/2 and last M/2 observations. In the randomization tests, the observations were shuffled 99 times.

Basically, for the data considered here, the performances of the two tests cannot be distinguished when the data sets are constructed to be appropriate for the conventional parametric t test. If the observations are a random sample from a Normal population that is independent of the group to which the observation belongs, then every possible configuration of the observed scores across groups was equally likely. Thus, the randomization test null hypothesis is implied by the conventional parametric t test null hypothesis. The converse is not true, however. That is, the conventional parametric t test null hypothesis is not implied by the randomization test null hypothesis.

Suppose an experiment is conducted in which there are two groups of two subjects each and the scores that are observed for the subjects are –1.01, –0.99, 0.99, and 1.01. If the randomization test null hypothesis is true, all possible assignments of these scores to the subjects are equally likely.

[20] Let ft and fr be the rejection frequencies for the t test and randomization test, respectively, and N be the number of data sets on which each of the tests were run. Using the Normal approximation to the binomial, the difference in the rejection frequencies can be tested with the ratio $[(N-1)^{1/2}(ft-fr)]/[ft(1-ft)+fr(1-fr)]^{1/2}$ which is distributed approximately as Student's t with 2N–2 degrees of freedom. The largest value of this t ratio for the data in the table is less than 0.80.

The test statistic for the randomization test is the difference between the means of the two groups. The 24 (= 4!) possible permutations of the four observations and the difference in the means for each of these permutations are listed in Table 2A.2.

Table 2A.2
An example of a test of the difference in means

Permutation	Group 1		Group 2		Difference in means	t statistic
1	–1.01	–0.99	0.99	1.01	2.00	141.42
2	–1.01	–0.99	1.01	0.99	2.00	141.42
3	–1.01	0.99	–0.99	1.01	0.04	0.03
4	–1.01	0.99	1.01	–0.99	0.04	0.03
5	–1.01	1.01	–0.99	0.99	0.00	0.00
6	–1.01	1.01	0.99	–0.99	0.00	0.00
7	–0.99	–1.01	0.99	1.01	2.00	141.42
8	–0.99	–1.01	1.01	0.99	2.00	141.42
9	–0.99	0.99	–1.01	1.01	0.00	0.00
10	–0.99	0.99	1.01	–1.01	0.00	0.00
11	–0.99	1.01	0.99	–1.01	–0.04	–0.03
12	–0.99	1.01	–1.01	0.99	–0.04	–0.03
13	0.99	1.01	–1.01	–0.99	–2.00	–141.42
14	0.99	1.01	–0.99	–1.01	–2.00	–141.42
15	0.99	–0.99	1.01	–1.01	0.00	0.00
16	0.99	–0.99	–1.01	1.01	0.00	0.00
17	0.99	–1.01	1.01	–0.99	0.04	0.03
18	0.99	–1.01	–0.99	1.01	0.04	0.03
19	1.01	0.99	–0.99	–1.01	–2.00	–141.42
20	1.01	0.99	–1.01	–0.99	–2.00	–141.42
21	1.01	–0.99	–1.01	0.99	–0.04	–0.03
22	1.01	–0.99	0.99	–1.01	–0.04	–0.03
23	1.01	–1.01	0.99	–0.99	0.00	0.00
24	1.01	–1.01	–0.99	0.99	0.00	0.00

The relative frequencies of the five possible values of the test statistic are listed in Table 2A.3.

Using an exact randomization test, the null hypothesis would never be rejected at the 0.10 level. The largest possible value of the test statistic is 2.00, which occurs too frequently (16.7% of the time) to reject the null hypothesis at the 0.10 level.[21]

[21] If an approximate randomization test were used, the null hypothesis may be rejected due to errors in approximating the exact randomization distribution. This source of error decreases as the number of shuffles increases. To take an extreme case, if there were only nine shuffles, the null hypothesis would be incorrectly rejected 3.7% of the time.

Table 2A.3
Relative frequencies of the differences in means

Difference in means	Relative frequency
2.00	4/24
0.04	4/24
0.00	8/24
–0.04	4/24
–2.00	4/24

In contrast, a conventional parametric t test would reject the null hypothesis far too frequently. The pooled variance t statistics for the data permutations are listed in the last column of Table 2A.2. A t statistic of 141 with 2 degrees of freedom is significant at the 0.001 level. Therefore, if it is true that all permutations of the data are equally likely, then the null hypothesis would be rejected at the 0.001 level 16.7% of the time! This is because 16.7% of the possible permutations result in a t statistic of 141. The difficulty is that the randomization and conventional parametric t test are tests of different null hypotheses. While it is true that the value of an observation is independent of the group to which the observation belongs (which is the randomization test null hypothesis), it is not true that the observations are a random sample from a Normal population (which is the conventional t test null hypothesis). A randomization test is a valid test of the hypothesis that the score is independent of the group; a conventional parametric t test is not necessarily a valid test of that hypothesis.

2A.2 TESTS OF THE CORRELATION BETWEEN TWO VARIABLES

A conventional t test is often used when a researcher is interested in the correlation between two variables. The conventional parametric t test of the correlation between two variables is valid when applied to a random sample from a bivariate Normal population with zero correlation.

Artificial data sets consisting of M observations each were constructed by generating two standard Normal scores for each observation. The first standard Normal score was used as the first variable. The second variable was obtained by multiplying the first variable by r and then adding the second Normal score. The test statistic is the correlation between the two variables.

Two values of the constant, r = 0.0 and r = 0.25, and two sample sizes, M = 10 and M = 100, were used. For each of the two sample sizes, 1000 basic data sets were independently generated. Each basic data set was used to generate two derived data sets: one for the case where r = 0.0 and one for the case where r =

0.25. The only difference in the derived data sets is the constant amount r which is multiplied by the first variable when constructing the second variable. For each derived data set, a conventional t test of the correlation was conducted. The frequencies with which the null hypothesis was rejected at the 0.10 level are reported in Table 2A.4. For example, when the derived data sets consisted of 100 observations each with the constant r set at 0.25, the null hypothesis was rejected 87.7% of the time at the 0.10 level.

Table 2A.4
Frequency with which the null is rejected at the 0.10 level: Tests of the correlation

	Sample Size	
r = 0.25	M = 10	M = 100
t test	0.297	0.877
randomization	0.300	0.870
r = 0.00		
t test	0.106	0.092
randomization	0.101	0.091

Note: The tests were conducted on 1000 independently generated data sets. Each data set consisted of M observations on two variables, the first of which was M standard Normal scores. The values of the second variable were generated by multiplying the values of the first variable by r and adding another standard Normal score. The test statistic was the correlation between the two variables. Randomization tests involved shuffling the two variables relative to each other 99 times.

In addition, the approximate randomization method was used to assess the significance of the correlation for each derived data set. This was accomplished by shuffling the second variable relative to the first and then computing the correlation. If the correlation for the shuffled data was at least as large as the correlation for the unshuffled data, 1 was added to the nge counter. This process was repeated 99 times for each data set. The null hypothesis was rejected for a data set if, at the end of 99 shuffles, the ratio (nge+1)/100 was less than or equal to 0.10.

The approximate randomization method is used to test the null hypothesis that the two variables are independent. The frequencies with which this null hypothesis was rejected are also reported in Table 2A.4. For example, when the

derived data sets consisted of 100 observations each with the constant r set at 0.25, the null hypothesis was rejected 87.0% of the time at the 0.10 level.

By construction, the parametric t test is valid for these data. And, as shown in Appendix 3A, an approximate randomization test is valid for any data and any test statistic. Therefore, it should not be surprising that the null hypothesis is rejected about 10% of the time at the 0.10 level when the null hypothesis is true (i.e., r = 0.00). All of the rejection frequencies are within a 90% confidence interval surrounding 0.10. And, as with the t test of the difference in the means of two groups, when the null hypothesis is false (i.e., r = 0.25), the differences in the rejection rates are not statistically significant.

As before, the performances of the two tests cannot be distinguished when the data sets are constructed to be appropriate for the conventional parametric t test. If the observations are a random sample from a bivariate Normal population with zero covariance, then every permutation of the variables relative to each other was equally likely. Thus, the conventional parametric t test null hypothesis implies the randomization test null hypothesis. The converse is not true, however. That is, the conventional parametric t test null hypothesis is not implied by the randomization test null hypothesis.

Suppose an experiment is conducted in which observations are taken on two variables, x and y, for four subjects. Further suppose that the observed values of x are {2, 4, 6, 8} and the observed values of y are {1, 2, 3, 4}. If the randomization test null hypothesis is correct, all possible permutations of y relative to x are equally likely. The 24 (= 4!) possible permutations are listed in Table 2A.5 along with the correlation associated with each permutation.

Table 2A.5
An example of a test of the correlation

1	2	3	4		
Permutations of y				Correlation	t Statistic
1	2	3	4	1.0	$+\infty$
1	2	4	3	0.8	1.89
1	3	2	4	0.8	1.89
1	3	4	2	0.4	0.62
1	4	2	3	0.4	0.62
1	4	3	2	0.2	0.29
2	1	3	4	0.8	1.89
2	1	4	3	0.6	1.06
2	3	1	4	0.4	0.62
2	3	4	1	–0.2	–0.29
2	4	3	1	–0.4	–0.62
2	4	1	3	0.0	0.00
3	4	1	2	–0.6	–1.06
3	4	2	1	–0.8	–1.89

Permutations of y				Correlation	t Statistic
3	2	4	1	–0.4	–0.62
3	2	1	4	0.2	0.29
3	1	4	2	0.0	0.00
3	1	2	4	0.4	0.62
4	3	2	1	–1.0	$-\infty$
4	3	1	2	–0.8	–1.89
4	2	1	3	–0.4	–0.62
4	2	3	1	–0.8	–1.89
4	1	3	2	–0.4	–0.62
4	1	2	3	–0.2	–0.29

The 11 possible values of the correlation and their relative frequencies are listed in Table 2A.6. Using an exact randomization test, the null hypothesis would be rejected with probability 1/24 at the 0.10 level.[22]

Table 2A.6
Relative frequencies of the correlation

Correlations	Relative frequency
1.0	1/24
0.8	3/24
0.6	1/24
0.4	4/24
0.2	2/24
0.0	2/24
–0.2	2/24
–0.4	4/24
–0.6	1/24
–0.8	3/24
–1.0	1/24

The t statistics for the conventional parametric t test of the correlation are listed in the last column of Table 2A.5. Since the t statistic for a perfect correlation is infinite, the null hypothesis would be falsely rejected at any level at least 4% of the time – at the .10 level, the false rejection rate would be 16.67%.

[22] If an approximate randomization test were run, the null hypothesis may be rejected more frequently. To take an extreme example, with only nine shuffles, the null hypothesis would be rejected about 6% of the time at the 0.10 rejection level.

$0.06 \approx (1/24)(23/24)^9+(3/24)(20/24)^9+(1/24)(19/24)^9+(4/24)(15/24)^9+0...+(3/24)(1/24)^9$

Once again, it must be emphasized that the randomization and conventional parametric t test are tests of different null hypotheses. It has been assumed in this example that the randomization null hypothesis is true; the values of the two variables are independent (and, as a consequence, all permutations of one variable relative to another were equally likely). This does not imply that the observations are a random sample from a bivariate Normal population with zero covariance, which is the null hypothesis for the conventional parametric t test of the correlation. A randomization test is a valid test of the hypothesis that the variables are independent; a conventional parametric t test is not necessarily a valid test of that hypothesis.

2A.3 SUMMARY

The performance of approximate randomization and conventional parametric tests was compared for two common situations: a test of the difference in the means between two groups and a test of the correlation between two variables. Consistent with theory, the randomization test is valid for situations in which the conventional parametric tests are valid and, furthermore, there appears to be essentially no loss in power when a randomization test is used instead of the conventional parametric test. On the other hand, the conventional parametric tests are not always valid in situations in which the randomization test is valid.

CHAPTER THREE

Monte Carlo Sampling

Monte Carlo sampling is used to test the null hypothesis that a sample was randomly drawn from a specified population. The test is conducted by simulating the process of drawing random samples from the population. The values of the test statistic for the simulated random samples are compared to the value of the test statistic for the real sample. If the value of the test statistic for the real sample is unusual relative to the values for the simulated random samples, then the null hypothesis is rejected.

3.1 INTRODUCTION

The Monte Carlo method of assessing the significance of a test statistic, which was introduced by Barnard [1963], is used to test the hypothesis that the data are a random sample from a specified population. This is accomplished by drawing simulated samples from the specified population and comparing the values of the test statistic for the simulated samples to the value of the test statistic for the real sample. The Monte Carlo method is particularly valuable in situations where the population distribution is known, but the sampling distribution of the test statistic has not been analytically derived. The procedure is illustrated in Figure 3.1.

When the Monte Carlo method is used, the population from which simulated samples are to be drawn must be defined. The test is conducted by drawing simulated random samples from the specified population. The test statistic is computed for each simulated random sample and the null hypothesis is rejected if the value of the test statistic for the actual sample is unusually large relative to the values of the test statistic for the simulated samples. To be specific, the significance level of the test is (nge+1)/(NS+1), where nge is the number of simulated samples for which the value of the test statistic is at least as large as the test statistic for the real sample, and NS is the number of simulated samples generated from the specified population.[1]

As will be demonstrated in Appendix 3A, a hypothesis test based on the significance level (nge+1)/(NS+1) is valid. That is, the probability of rejecting the null hypothesis when it is true is no greater than the rejection level selected for the test. The null hypothesis in this case is that the data are a random sample drawn from the specified population.

Most commonly, there are N discrete elements in the population and a sample of M of the elements is drawn at random (and without replacement) from the population. The object of this sampling is to make an inference concerning the population.

Usually, in the sampling process each element of the population has an equal chance of being included in the sample. In sampling without replacement, the observation is removed from the population once it has been selected for the sample and cannot be selected again. Mechanically, sampling without replacement can be carried out by assigning an identification number to each member in the population, shuffling the identification numbers, and then designating the first M as the sample.

---

[1] Approximate randomization tests, discussed in Chapter 2, are an important special case of Monte Carlo sampling, so it is no accident that the structure of a Monte Carlo test appears to be so similar to that of an approximate randomization test. The population in an approximate randomization test is the set of permutations of one variable relative to the others and the data are a sample of size one from this population.

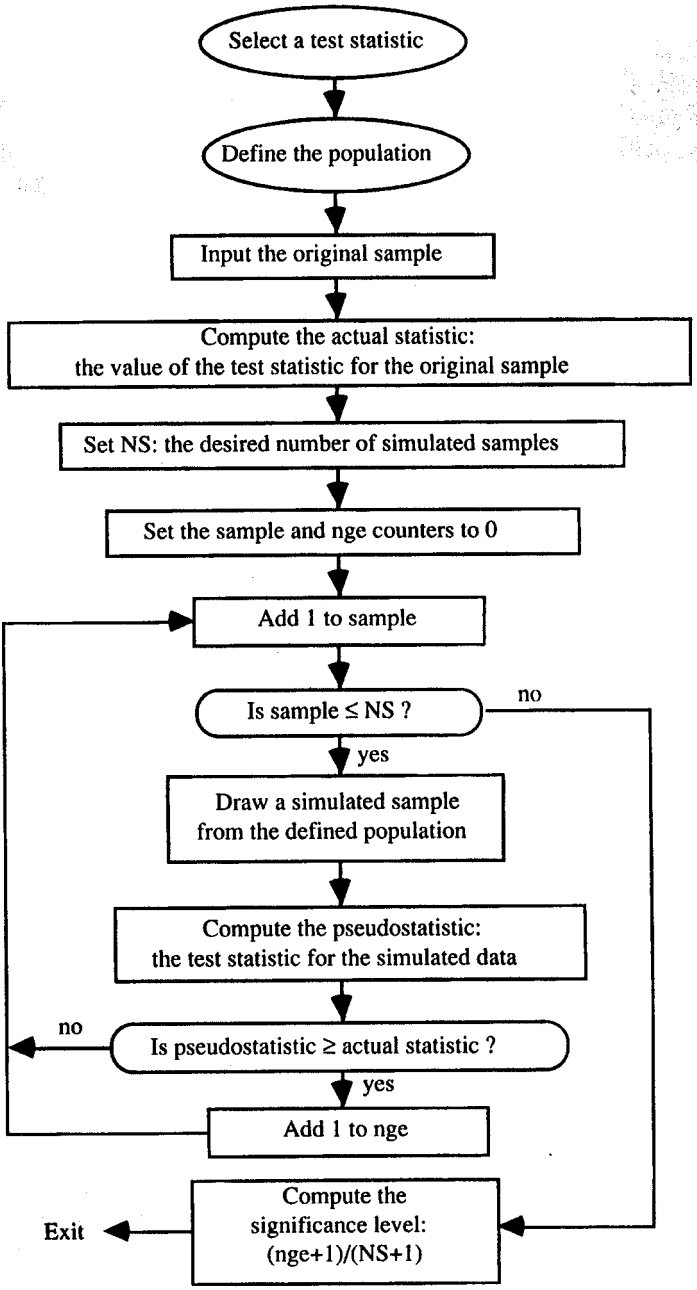

Figure 3.1 Flowchart for Monte Carlo sampling from a defined population

3.2 EXAMPLE: UNIFORM SAMPLING WITHOUT REPLACEMENT

A firm which manufactures high-fidelity speakers has just received its first shipment of 1000 units of a component from a new supplier. If this component is defective and is used in final assembly, routine diagnostic tests will uncover the trouble and the speaker unit can be reworked. A less expensive, but still costly, alternative is to test the components before assembly. In order to minimize rework and testing costs, the supplier has agreed to the stipulation that at least 98% of the components delivered to the factory must be free of defects.

In order to audit the supplier's compliance with this stipulation, a random sample of 100 components was drawn (without replacement) from the shipment. Each of the 100 units in the random sample was tested and 4 were found to be defective, a defect rate of 4%. When confronted with this evidence, the supplier claimed that the sample was not representative of the entire shipment – the sample just happened to contain more than 2% defects. Indeed, there are two possibilities: either the defect rate in the shipment is actually greater than 2% or, as the supplier claims, the sample is misrepresentative. The likelihood that the supplier is correct can be assessed using the Monte Carlo method.[2]

The test statistic in this case is the number of defective units in a sample of 100. The value of the test statistic is 4 for the original sample. The population of 1000 units is alleged to contain no more than 2% defective units, or at most 20 defective units altogether. The test proceeds by constructing a mathematical representation of the population, which in this case would be a column of 20 ones (i.e., defective components) and 980 zeros (i.e., nondefective components). Then an artificial sample of size 100 is drawn from this model of the population, taking care to exactly mimic the way in which the real sample was selected from the real population. The test statistic is computed for the artificial sample and

[2] The exact probability of obtaining 4 or more defects in a sample of 100 taken without replacement from a lot of 1000 in which 20 are defective is given by the hypergeometric distribution and is 0.131. One might ask why the Monte Carlo method is used when the exact probability can be obtained. First, plainly speaking, most researchers would not realize that the hypergeometric distribution is appropriate in this instance. Second, it would be difficult for most researchers to obtain the value of the hypergeometric distribution for these parameter values ; ordinary tables don't go up to lot sizes of 1000. Well, why not fall back on the binomial approximation to the hypergeometric distribution? In other words, why not assume that the sampling is done with replacement (the binomial assumption) and that the probability of obtaining a defective unit on any one draw is just 0.02? Indeed, use of the binomial approximation would yield a probability of 0.141, which is quite close to the actual probability of 0.131. However, most researchers would have practical difficulties in finding this probability since ordinary tables for the binomial don't extend up to sample sizes of 100. Well, why not use the Normal approximation to the binomial distribution? If this approximation is done correctly, the probability level would be computed as 0.142. I would speculate, however, that most researchers, if they got this far, would neglect to make the continuity correction (i.e., integrating from 3.5 rather than from 4) and compute a probability level of 0.077, which is quite a bit different from the actual probability of 0.131. The point here is that even in this simple example, using conventional methods properly requires considerable statistical expertise.

compared to the value of the test statistic for the real sample. This simulation of the sampling procedure is repeated many times and, as with approximate randomization tests, the significance of the test statistic is given by the ratio (nge+1)/(NS+1), where NS is the number of artificial samples drawn from the hypothesized population, and nge is the number of artificial samples for which the value of the test statistic is equal to or greater than the value of the test statistic for the real sample.

See Program 3.1 in the Programs Appendix of your choice for a listing of a computer program that carries out this test. Referring to the results of running the BASIC program, the probability of obtaining 4 or more defects in a sample size of 100 is 0.119. That is, there were 4 or more defects in 118 of the 999 artificially constructed samples.

3.3 AN EXAMPLE OF A TEST OF WHETHER A SAMPLE IS RANDOM

The last example illustrates a situation in which it is known that the data are a random sample. The question in such a situation is whether the random sample was drawn from the population that is specified in the null hypothesis. Less commonly, the population from which a sample is drawn is known and the question is whether the sample is a random sample.

Suppose you are trying to decide whether to follow an investment analyst's advice and you have in front of you the analyst's "Ten Best Picks for 1986" list which he had mailed to clients at the end of 1985. The ten stocks the analyst selected and their prices and dividends are listed in Table 3.1.

The total investment in the portfolio would have been $438 and the value of the portfolio at the end of the year, including dividends, was $468.873. Thus the rate of return on the portfolio was 7.05% [=($468.873 –$438.00)/$438.00]. While positive, this rate of return may or may not have been better than could have been earned by selecting a portfolio at random.

A natural null hypothesis to consider is that the method used by the analyst to select stocks is substantively equivalent to selecting stocks at random. That is, the null hypothesis is that an investor would be just as well off selecting stocks by throwing darts at the financial page as relying on the analyst's advice. The alternative one-tailed hypothesis would be that the portfolio selected by the analyst is not random and that funds invested as suggested by the analyst would earn returns exceeding the returns on funds invested in randomly selected portfolios.[3]

The test statistic is the rate of return on the portfolio. The significance of this test statistic can be assessed by randomly selecting portfolios of ten stocks

[3] Over the long haul, an analyst can do better than average by selecting riskier stocks since higher risks generally earn higher returns. Thus, a more revealing test would control for the riskiness of the analyst's portfolio.

and comparing the returns on those random portfolios to the returns on the analyst's portfolio. The null hypothesis is that the analyst's portfolio is substantively a random sample of stocks listed on the New York Stock Exchange. Since the population from which the analyst selected the portfolio is known, the only question is whether the analyst's choices were essentially random.

Table 3.1
The stocks picked by the analyst

	Price at 12/31/85	Price at 12/31/86	Dividend for 1986
PPG INDUSTRIES INC	51.000	72.875	1.880
HILTON HOTELS CORP	64.875	67.250	1.800
GENERAL ELECTRIC CO	72.750	86.000	2.370
MCDERMOTT INTL INC	18.250	21.750	1.800
LOUISIANA LAND & EXPLORATION	30.250	27.250	1.000
DRESSER INDUSTRIES INC	18.125	19.375	0.700
MCDONALD'S CORP	80.875	60.875	0.645
WEST POINT-PEPPERELL	43.375	52.125	2.385
SERVICE CORP INTERNATIONAL	31.250	24.750	0.293
AMERADA HESS CORP	27.250	23.750	0.000
Total	438.00	456.00	12.873

A program to carry out this test is listed as Program 3.2 in the Programs Appendix of your choice. Out of 999 randomly formed portfolios, about 26% had a return as high as 7.05%. Following the usual academic formulation, the null hypothesis that the analyst's selections were random cannot be rejected at conventional levels of significance.[4]

3.4 MONTE CARLO AND APPROXIMATE RANDOMIZATION TESTS

Sampling without replacement can be structured as an approximate randomization test if each element in the population has an equal chance of being included in the sample. Mechanically this would be done by shuffling a dummy variable that is a column of M ones and N –M zeros. After shuffling, a value of 1 indicates that the observation is included in the sample and a value of 0 indicates that the observation is not included in the sample. In the routine that computes the test statistic, an observation is ignored if the value of the dummy variable for that observation is 0.

[4] On the other hand, the analyst can point to his success in beating 72% of the randomly formed portfolios. Conventional academic rejection levels such as 10% and 5% may be unnecessarily rigid – particularly when some decision needs to be made based on the evidence at hand.

There is symmetry. Sampling without replacement can be viewed as a special case of an approximate randomization test, but an approximate randomization test can also be viewed as a special case of a Monte Carlo test. In the case of a randomization test, the population consists of all possible permutations of one variable relative to another. The actual data consist of a sample of size one from that population, i.e., in the actual data, there is one and only one of the possible permutations. The approximate randomization test is conducted by sampling (with replacement) from the population of possible permutations. Since Monte Carlo tests are valid (see Appendix 3A to this chapter) and approximate randomization tests are Monte Carlo tests, approximate randomization tests are also valid.

3.5 CONCLUSION

The key element in the Monte Carlo sampling method, as in conventional parametric methods, is the population that is specified in the null hypothesis. In each of the above examples, the population that should be used in the test was clear. However, that will not always be the case. Bootstrap resampling, which is discussed in the next chapter, takes one approach to specifying the population to be used in generating the artificial data. In addition, a researcher can always fall back on the models of populations used in conventional parametric tests.[5] For example, sampling from a standard Normal population can be easily simulated. There would be no apparent advantage to doing this when the exact sampling distribution of the test statistic is known. However, the exact sampling distributions are not known for all possible test statistics (e.g., the median) even when sampling from the standard Normal population.

In general, a valid Monte Carlo significance level can be computed for any test statistic that is a function of data drawn from any specified population. The population does not have to have a familiar, well-behaved distribution studied by statisticians; the population can be entirely arbitrary. The only requirement is that relative frequencies of any two distinct elements in the population must be specified in some manner.[6]

---

[5] Rubinstein [1981] and Cooke, Craven, and Clarke [1982] describe how to generate random samples from a variety of populations including the binomial, Poisson, geometric, exponential, Normal, Student's t, chi-squared, F, gamma, beta, Weibel, and Cauchy. Briefly, random sampling from such populations can be simulated by generating a uniformly distributed random number between 0 and 1 and then inverting the cumulative distribution function form of the model. The value of a test statistic can be computed for the simulated random sample and the significance of the actual value of the test statistic can then be assessed relative to the values for the simulated samples in the usual way. Software packages such as IMSL, NAG, and NUMERICAL RECIPES contain subroutines that generate random variables from many of these distributions.

[6] Two elements in a population are distinct if they result in different values of the test statistic when substituted for each other in some sample drawn from the population.

APPENDIX 3A

VALIDITY OF COMPUTER-INTENSIVE HYPOTHESIS TESTS

3A.1 THE NOTION OF THE VALIDITY OF A TEST

The probability of falsely rejecting the null hypothesis should be no greater than the rejection level of the test. For example, if the rejection level is 0.10, the probability of rejecting the null hypothesis when it is true should be no greater than 0.10. Ideally, the probability of falsely rejecting the null hypothesis would be exactly 0.10.

A test is valid if the probability of rejecting the null hypothesis, when it is true, is less than or equal to the rejection level of the test. A test is exactly valid if the probability of falsely rejecting the null hypothesis is equal to the rejection level of the test. As will be demonstrated in this appendix, a Monte Carlo hypothesis test is valid for any preset rejection level and is, under many circumstances, for practical purposes exactly valid as well. This conclusion holds for approximate randomization tests and bootstrap tests (to be discussed in the next chapter) since they are both special cases of the Monte Carlo method of assessing significance.

Hoeffding [1952] and Box and Andersen [1955] prove that exact randomization tests are valid but do not address the validity of approximate randomization tests. Dwass [1957] proves the validity of approximate randomization tests under the assumption that there are an infinite number of possible permutations of the data and that the value of the test statistic is different for each permutation. Edgington [1969] provides an intuitive explanation for the validity of approximate randomization tests under the same assumption. Noreen [1986] proves that approximate randomization tests are, in general, valid without invoking the Dwass assumptions.

Foutz [1980] proves that exactly valid Monte Carlo tests can be constructed by employing an auxiliary random variable to break ties. The discussion below is built on the Foutz argument and extends it to consider the case where the researcher chooses not to use an auxiliary random variable.

3A.2 VALIDITY USING AN AUXILIARY RANDOM VARIABLE

Let the test statistic t be a function of the data matrix X, which has a known distribution under the null hypothesis. The Monte Carlo method consists of generating NS independent samples $X_1, X_2, ..., X_{NS}$ from the null distribution. In the case of an approximate randomization test, these independent samples are generated by permuting at least one of the elements of X relative to the others. For notational convenience, assume the NS samples are ordered so that:

$$t(X_1) \geq t(X_2) \geq \cdots \geq t(X_{NS})$$

Define: $nge = \max\{\text{integer } k \mid t(X_k) \geq t(X_0) \text{ for } k = 0, ..., NS\}$

X_0 is the original data and $t(X_0)$ is the value of the test statistic for the original data. The random variable nge is the number of times the value of the test statistic for the simulated data is greater than or equal to the the value of the test statistic for the original data. The null hypothesis is rejected at the α level of rejection if $(nge+1)/(NS+1) \leq \alpha$.

The claim is that this test is valid; that is, $prob\{(nge+1)/(NS+1) \leq \alpha\} \leq \alpha$. The potential for tied test statistics creates some difficulty in the proof. To eliminate ties, generate NS+1 auxiliary random variables $\varepsilon_0, ..., \varepsilon_{NS}$ that are uniformly distributed on an arbitrary small real interval $(-\delta, +\delta)$.[7] Transform the test statistics by adding to them these random disturbances.

$$t'(X_i) = t(X_i) + \varepsilon_i \quad \text{for } i = 0, ..., NS$$

Adding this auxiliary random variable eliminates tied test statistics, so that

$$t'(X_1) > t'(X_2) > \cdots > t'(X_{NS})$$

The transformed value of the test statistic for the original data $t'(X_0)$ must fall in one of the NS+1 intervals:

$$(-\infty, t'(X_{NS})], ..., (t'(X_2), t'(X_1)], (t'(X_1), \infty)$$

Under the null hypothesis, the observed value of the test statistic $t'(X_0)$ for the original data is a sample of size one from a distribution that is independent of, and identical to, the distribution of $t'(X_i)$ for all i. Under this condition, the probability that $t'(X_0)$ falls into any specific interval above is 1/(NS+1). The random variable nge' is defined by

$$nge' = \max\{\text{integer } k \mid t'(X_k) \geq t'(X_0) \text{ for } k = 0, ..., NS\}$$

Or, nge' is the number of intervals above that are contained in the half-closed interval $[t'(X_0), \infty)$. Hence, nge' can take on any integer value from 0 to NS and each of these integer values is equally likely under the null hypothesis. Therefore, given the rejection level α, if NS is selected so that $\alpha(NS+1)$ is an integer, then

[7] Assume that the interval is chosen to be so small that it has the effect of only breaking ties between test statistics. In other words, if $t(X_i) < t(X_j)$, then $t'(X_i) < t'(X_j)$. Any δ that satisfies the following condition will suffice: $\delta < \min\{\, |t(X_i) - t(X_j)| > 0 \,\}$.

$$\text{prob}\{(\text{nge}+1) \le \alpha(\text{NS}+1)\} = \alpha(\text{NS}+1)/(\text{NS}+1) = \alpha$$

or

$$\text{prob}\left\{\frac{\text{nge}+1}{\text{NS}+1} \le \alpha\right\} = \alpha$$

This demonstrates that if NS is selected so that $\alpha(NS+1)$ is an integer, and if the auxiliary randomization is carried out to ensure that there are no ties, then a Monte Carlo test is exactly valid. That is, the probability of rejecting the null hypothesis when it is true is exactly equal to the rejection level for the test.

Common rejection levels are 0.01, 0.05, and 0.10. If α is restricted to this set, then setting NS = 100k – 1, where k is any positive integer, will satisfy the condition that $\alpha(NS+1)$ is an integer. For example, NS = 99 or 199 or 999 will satisfy the condition for $\alpha = 0.01, 0.05$, and 0.10.

3A.3 VALIDITY WITHOUT AN AUXILIARY RANDOM VARIABLE

The use of an auxiliary random variable to break ties (i.e., convert the distribution of the test statistic into a continuous variable measured on the real line) would be bothersome to most empirical researchers. It is unsettling to explicitly allow a conclusion drawn from an experiment to depend on a gratuitous flip of the coin.

If auxiliary randomization is omitted, the test will not necessarily be exact, but it will still be valid in that the probability of rejecting the null hypothesis when it is true will be no more than the nominal rejection level of the test.

To see this, suppose that the interval $(-\delta, \delta)$ is chosen to be so small so that if $t(X_i) < t(X_j)$, then $t'(X_i) < t'(X_j)$. If δ is chosen to satisfy this condition, then the auxiliary randomization will have no effect on the hypothesis test unless $t(X_0) = t(X_i)$ for some i. The effect of auxiliary randomization in that case is to randomly partition the samples X_i for which $t(X_0) = t(X_i)$ into two groups. The samples in the first group are counted as if $t(X_0) < t(X_i)$ and the samples in the second group are counted as if $t(X_0) > t(X_i)$. Recall that nge' is the number of samples for which the value of the test statistic is greater than or equal to the value of the test statistic for the original data, taking into account this auxiliary randomization. And recall that nge is the number of samples for which the value of the test statistic is greater than or equal to the value of the test statistic for the original data without resort to auxiliary randomization. The difference between nge and nge' is just the number of samples for which $t(X_0) = t(X_i)$ and, as a consequence of auxiliary randomization, are counted as if $t(X_0) > t(X_i)$. In other words, nge cannot be less than nge'. Therefore,

$$\text{prob}\left\{\frac{nge+1}{NS+1} \le \alpha\right\} \le \text{prob}\left\{\frac{nge'+1}{NS+1} \le \alpha\right\} = \alpha$$

The implication is that a Monte Carlo hypothesis test based on the ratio (nge+1)/(NS+1) is valid. Note that even in the absence of auxiliary randomization, the test is exactly valid unless $t(X_0) = t(X_i)$ for some sample i.

3A.5 DISCUSSION

Despite this reassurance, researchers should consider using auxiliary randomization whenever $t(X_0) = t(X_i)$ for some sample i. First, auxiliary randomization ensures that the test is exactly valid. Second, because auxiliary randomization increases the probability of rejecting the null hypothesis when there are tied test statistics, auxiliary randomization increases the power of the test.

On the other hand, how serious is the bias of the Monte Carlo method against rejecting the null hypothesis when auxiliary randomization is not used? The bias is likely to be most pronounced in randomization tests conducted on small data sets. For example, if a randomization test is conducted on just four observations of two variables, there are only 24 (= 4!) possible permutations of one of the variables relative to the other. If the data are shuffled 99 times, the likelihood of $t(X_0) = t(X_i)$ for some i is about 98.5%. Table 3A.1 indicates how seriously an approximate randomization test, without auxiliary randomization, is biased against rejecting the null hypothesis when it is true. With as few as seven observations to be shuffled, the bias is negligible.

Table 3A.1
The probability of falsely rejecting the null hypothesis at the α rejection level using an approximate randomization test: (NS=99)

	The exact probability of falsely rejecting the null hypothesis				
α	K = 4!	K = 5!	K = 6!	K = 7!	K = 8!
0.01	0.0006	0.0064	0.0093	0.0099	0.0100
0.05	0.0285	0.0458	0.0493	0.0499	0.0500
0.10	0.0791	0.0958	0.0993	0.0999	0.1000

APPENDIX 3B

CONFIDENCE LEVELS FOR COMPUTER-INTENSIVE TESTS

Computer-intensive hypothesis tests are conducted by sampling from a specified distribution using Monte Carlo simulation. (In the case of approximate randomization tests, the assumed population consists of all permutations of a variable. The original data are a sample of size one from this population.) In the last appendix, it was shown that the Monte Carlo significance level, (nge+1)/(NS+1), for a test statistic is valid for any number of Monte Carlo trials NS. That is, the probability of rejecting the null hypothesis that the data are a random sample from the assumed population is no higher than the rejection level of the test. Despite the fact that the test is valid for any NS, no matter how small, confidence in the inferences drawn from the test increases as NS is increased. In this appendix, the relationship between NS and the confidence level associated with the test is explored.

$$\text{Let } \phi = \lim_{NS \to \infty} \frac{nge + 1}{NS + 1}$$

If ϕ were known, then the null hypothesis should be rejected if and only if ϕ is less than or equal to α, the rejection level of the test. But ϕ is not known; it can only be approximated with any finite number of Monte Carlo trials. For any given NS, the null hypothesis is rejected if (nge+1)/(NS+1) is less than or equal to the rejection level of the test. But, even if (nge+1)/(NS+1) $\leq \alpha$, it may not be true that $\phi \leq \alpha$. If the null hypothesis is rejected, but ϕ is actually greater than α, an incorrect inference has been drawn. Or, if the null hypothesis is not rejected, but ϕ is actually less than or equal to α, then an incorrect inference has also been drawn. Fortunately, it is possible to compute, for given nge and NS, the posterior probability β that $\phi \leq \alpha$. This probability β can be interpreted as the confidence level of the test.

By Bayes' Theorem,

$$\beta = \text{prob}(\phi \leq \alpha \mid nge, NS) = \frac{\sum_{\phi \leq \alpha} \text{prob}(nge \mid \phi, NS)\,\text{prob}(\phi \mid NS)}{\sum_{\phi} \text{prob}(nge \mid \phi, NS)\,\text{prob}(\phi \mid NS)}$$

But prob(ϕ | NS) = prob(ϕ), so

$$\beta = \text{prob}(\phi \le \alpha | \text{nge}, \text{NS}) = \frac{\sum_{\phi \le \alpha} \text{prob}(\text{nge} | \phi, \text{NS}) \text{prob}(\phi)}{\sum_{\phi} \text{prob}(\text{nge} | \phi, \text{NS}) \text{prob}(\phi)}$$

The posterior probability β depends on the prior beliefs, prob(ϕ), concerning the distribution of ϕ. For simplicity, we will assume that ϕ is a uniformly distributed random variable on the closed interval [0, 1]; that is, the probability density function of ϕ is $f(\phi) = 1$ for $0 \le \phi \le 1$. To be honest, in some settings ϕ cannot be uniformly distributed. For example, in approximate randomization tests there are only a finite number of permutations of the data and therefore only a finite number of different values of the test statistic. However, the uniform distribution is the limiting case as the number of equally likely possible values of the test statistic becomes large. Practically, there is no difference between the confidence levels computed using the uniform distribution and the confidence levels using a discrete distribution if the data are shuffled 99 times or more and there are more than about 500 possible different test statistics.

To distinguish the uniform prior from the more general case, βu will denote the posterior probability when the prior is uniform.

$$\beta u = \text{prob}(\phi \le \alpha \mid \text{nge, NS, uniform prior}) = \frac{\int_0^{\alpha} \text{prob}(\text{nge} | \phi, \text{NS}) d\phi}{\int_0^{1} \text{prob}(\text{nge} | \phi, \text{NS}) d\phi}$$

ϕ is the probability of obtaining a value of the test statistic greater than or equal to the actual value of the test statistic on any one Monte Carlo trial. Hence, prob(nge | NS, ϕ) is the binomial probability:

$$\text{prob}(\text{nge} \mid \text{NS}, \phi) = \binom{\text{NS}}{\text{nge}} \phi^{\text{nge}} (1-\phi)^{\text{NS}-\text{nge}}$$

Therefore,

$$\beta u = \text{prob}(\phi \le \alpha \mid \text{nge, NS, uniform prior}) = \frac{\int_0^{\alpha} \binom{NS}{nge} \phi^{nge}(1-\phi)^{NS-nge}\, d\phi}{\int_0^{1} \binom{NS}{nge} \phi^{nge}(1-\phi)^{NS-nge}\, d\phi}$$

or

$$\beta u = \text{prob}(\phi \le \alpha \mid \text{nge, NS, uniform prior}) = \frac{\int_0^{\alpha} \phi^{nge}(1-\phi)^{NS-nge}\, d\phi}{\int_0^{1} \phi^{nge}(1-\phi)^{NS-nge}\, d\phi}$$

The expression on the right-hand side of the above equation is the cumulative incomplete beta function with parameters nge+1 and NS–nge+1. βu is tabulated in the Probability Tables Appendix for various rejection levels, significance levels, and values of NS. For example, suppose that the significance level of the test, (nge+1)/(NS+1), is 0.06 with NS = 99. Consulting Table A, the probability that the actual value of ϕ is less than or equal to 0.10 is 0.942. Therefore, one can be reasonably confident that the null hypothesis should be rejected. If, on the other hand, the significance level is 0.10 with NS = 99, then the probability that the actual value of ϕ is less than or equal to 0.10 is only 0.549 and so the null hypothesis would be rejected with relatively little confidence at the 0.10 level.

Suppose that the significance level of the test is 0.14 with NS = 99. How confident can we be that the actual value of ϕ is actually greater than 0.10 and the null hypothesis should not be rejected at the 0.10 level? Again, referring to Table A, the probability that the actual value of ϕ is less than or equal to 0.10 is 0.124. Therefore, the probability that the actual value of ϕ is greater than 0.10 is 0.876 (= 1–0.124).

Another way to look at this problem is to estimate a confidence interval for ϕ. In this approach, the confidence level βu is preselected and a confidence interval is estimated such that the probability that the actual value of ϕ lies within that interval is βu. Let $\alpha_{\beta u}$ be implicitly defined by

$\alpha_{\beta u}$: $\beta u = \text{prob}(0 \le \phi \le \alpha_{\beta u} \mid \text{nge, NS, uniform prior})$ if the null is rejected.

βu = prob($\alpha_\beta < \phi \leq 1$ | nge, NS, uniform prior) if the null is not rejected.

The parameter $\alpha_{\beta u}$ can be interpreted as the end point of a one-sided confidence interval for ϕ. To be precise, when the null hypothesis is rejected, the confidence interval is $0 \leq \phi \leq \alpha_{\beta u}$ and βu is the probability that the actual value of ϕ lies within that interval. When the null hypothesis is not rejected, the confidence interval is $\alpha_{\beta u} < \phi \leq 1$ and βu is the probability that ϕ lies within that interval. Values of $\alpha_{\beta u}$ when the null hypothesis is rejected are listed in Table B of the Probability Tables Appendix. For example, if the significance level (nge+1)/(NS+1) is 0.05 after 99 shuffles, then with probability 0.95, the actual value of ϕ is less than or equal to 0.089. The 95% confidence interval with a uniform prior is therefore $0.000 \leq \phi \leq 0.089$.

Values of $\alpha_{\beta u}$ when the null hypothesis is not rejected are listed in Table C of the Probability Tables Appendix. For example, if the significance level (nge+1)/(NS+1) is 0.11 after 99 shuffles, then with probability 0.95, the actual value of ϕ is greater than 0.063. The 95% confidence interval with a uniform prior is therefore $0.063 \leq \phi \leq 1.000$.

It is prudent practice to report either the confidence level or a confidence interval whenever the significance of a test statistic is reported. It should be kept in mind, however, that the tables provided in this book are for a uniform prior. That is, the tables assume that in the mind of the researcher, any value of ϕ between 0 and 1 is equally likely.

APPENDIX 3C

THE POWER OF COMPUTER-INTENSIVE HYPOTHESIS TESTS

Computer-intensive hypothesis tests are conducted by sampling from a specified distribution using Monte Carlo simulation. (In the case of approximate randomization tests, the assumed population consists of all permutations of a variable. The original data are a sample of size one from this population.) In Appendix 3A, it was shown that the Monte Carlo significance level, (nge+1)/(NS+1), for a test statistic is valid for any number of Monte Carlo trials NS. That is, the probability of rejecting the null hypothesis that the data are a random sample from the assumed population is no higher than the rejection level of the test. Even though the test is valid for any NS, no matter how small, the power of a test increases as NS is increased.[8] In this appendix, the relationship between NS and the power of a test is explored.

$$\text{Let } \phi = \lim_{NS \to \infty} \frac{nge + 1}{NS + 1}$$

If the null hypothesis is true, then on average ϕ will be 0.5. If the test statistic is selected with care and the alternative hypothesis is true, then ϕ will be very small.

For example, consider the case of sampling from a Normal population with mean μ and variance 1, i.e., $N(\mu, 1)$. A random sample of size M is drawn from this population. Suppose the null hypothesis is that the sample was randomly drawn from a Normal population with mean 0 and variance 1. This hypothesis could be tested by generating NS samples from a N(0, 1) population and comparing the sample means from these NS samples to the sample mean for the original sample. The null hypothesis would be rejected if $(nge+1)/(NS+1) \leq \alpha$.

Suppose the sample mean for the original sample is μ_0. Then ϕ is the probability of drawing a random sample of size M with a mean at least as large as μ_0 from a N(0, 1) population. Thus, in this case $\phi = 1 - \Phi_N(\mu_0\sqrt{M})$, where Φ_N is the cumulative standard Normal probability function. If $\mu = 0$, then ϕ is uniformly distributed on the closed interval [0, 1]. However, as μ increases, the mass of the distribution of ϕ shifts toward zero and eventually all of the probability mass will be concentrated on the point $\phi = 0$. In other words, if μ is large enough, the value of ϕ is, for all practical purposes, always zero.

[8] Hope [1968] demonstrates that the power of a Monte Carlo test for fixed α is an increasing function of NS.

In general, the size of ϕ is an index of the level of departure of the data from the null hypothesis. If the alternative hypothesis is true, ϕ will tend toward zero.

The probability of obtaining a value of the test statistic on any one Monte Carlo trial that is greater than or equal to the value of the test statistic for the original sample is just ϕ. The probability of obtaining nge values of the test statistic in NS Monte Carlo trials that are greater than or equal to the value of the test statistic for the original data is given by the binomial probability:

$$\text{prob}(\text{nge} \mid \text{NS}, \phi) = \binom{\text{NS}}{\text{nge}} \phi^{\text{nge}}(1-\phi)^{\text{NS}-\text{nge}}$$

The null hypothesis is rejected if and only if $(\text{nge}+1)/(\text{NS}+1) \leq \alpha$ or, equivalently, if and only if $\text{nge} \leq \alpha(\text{NS}+1) - 1$. Assuming that NS is chosen so that $\alpha(\text{NS}+1)$ is an integer, then the probability of rejecting the null hypothesis, conditional on α, NS, and ϕ, is given by the cumulative binomial probability function:[9]

$$\text{prob}(\text{reject} \mid \alpha, \text{NS}, \phi) = \sum_{\text{nge}=0}^{\alpha(\text{NS}+1)-1} \binom{\text{NS}}{\text{nge}} \phi^{\text{nge}}(1-\phi)^{\text{NS}-\text{nge}}$$

This expression is the power of the Monte Carlo test conditional on α, NS, and ϕ. Given the rejection level α, the probability of rejecting the null hypothesis is an increasing function of NS and a decreasing function of ϕ. Table 3C.1 shows the probability of rejection for various combinations of α, NS, and ϕ. For example, if NS = 999 and $\phi = 0.08$, then the probability of rejecting the null hypothesis at the $\alpha = 0.10$ level is 0.987.

Unfortunately, this does not tell us what we would really like to know. This power calculation is conditional on ϕ, which depends on the original sample that was drawn. We would really like to know the probability of rejecting the null hypothesis across all samples that could be drawn from the specified population (in this case, N(0, 1)). This probability is a complicated expression for even simple cases and it is difficult to evaluate. The general form of the expression of the power of a Monte Carlo test conditional only on α and NS is as follows.

---

[9] Marriot [1979] constructs similar tables to those that appear in this appendix using the Normal approximation to the cumulative binomial function.

$$\text{prob(reject} \mid \alpha, \text{NS}) = \sum_{\text{nge}=0}^{\alpha(\text{NS}+1)-1} \binom{\text{NS}}{\text{nge}} \int_0^1 \phi^{\text{nge}}(1-\phi)^{\text{NS}-\text{nge}} \text{prob}(\phi) d\phi$$

where prob(ϕ) is a function of the actual population from which the original sample is drawn. Hope [1968] evaluates this expression for the power of a Monte Carlo test for several cases, including the example of samples drawn from a N(μ, 1) population. When the variance of the population is known, the uniformly most powerful parametric test of the hypothesis that the sample is drawn from a N(0, 1) population is a simple test of the sample mean. The power of the Monte Carlo approach cannot exceed the power of this parametric approach, but it can come close.

Using numerical methods, Hope solved for the maximum loss in power for NS = 19 and NS = 39. When the rejection level α is set at 0.05, the maximum loss in power is 0.0855 for NS = 19 and is 0.0428 for NS = 39. This is the maximum loss in power; ordinarily, the loss in power will be less than this for the case of sampling from a N(μ, 1). Assuming that the loss in power is approximately inversely related to NS, the maximum loss in power is of the order 0.02 when NS = 99 and 0.002 when NS = 999.[10] Thus, when NS = 999 and the rejection level is set at 0.05, the loss in power is negligible from using the Monte Carlo approach to assessing the significance of the sample mean rather than the conventional parametric approach, even when the assumptions of the conventional parametric approach are satisfied and it provides the uniformly most powerful approach.

[10]The function "loss in power" = $0.0855(19/\text{NS})^{0.962}$ fits the data. Using this function, the power loss when NS = 99 is 0.0175 and when NS = 999 is 0.0019.

Table 3C.1

The probability of rejecting the null hypothesis in NS Monte Carlo trials as a function of the rejection level α and the asymptotic significance level ϕ

$\alpha = 0.01$

	ϕ									
NS	.010	.009	.008	.007	.006	.005	.004	.003	.002	.001
99	.370	.409	.451	.499	.551	.609	.672	.743	.820	.906
199	.407	.464	.527	.594	.665	.738	.810	.879	.939	.983
499	.441	.533	.630	.727	.817	.892	.948	.982	.996	1.000
999	.459	.589	.718	.832	.917	.969	.992	.999	1.000	1.000
9999	.487	.843	.983	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$\alpha = 0.05$

	ϕ									
NS	.050	.045	.040	.035	.030	.025	.020	.015	.010	.005
99	.445	.539	.637	.734	.823	.897	.951	.983	.997	1.000
199	.461	.593	.724	.838	.921	.971	.993	.999	1.000	1.000
499	.475	.680	.850	.951	.990	.999	1.000	1.000	1.000	1.000
999	.483	.760	.935	.991	1.000	1.000	1.000	1.000	1.000	1.000
9999	.495	.991	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

$\alpha = 0.10$

	ϕ									
NS	.100	.090	.080	.070	.060	.050	.040	.030	.020	.010
99	.464	.600	.732	.845	.926	.973	.994	.999	1.000	1.000
199	.475	.664	.827	.934	.983	.997	1.000	1.000	1.000	1.000
499	.484	.767	.939	.993	1.000	1.000	1.000	1.000	1.000	1.000
999	.489	.855	.987	1.000	1.000	1.000	1.000	1.000	1.000	1.000
9999	.496	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

CHAPTER FOUR

Bootstrap Resampling

Bootstrap resampling is the most recent development in computer-intensive methods. It can be used when the objective of the test is to draw a conclusion about a population based on a random sample. In bootstrap methods, artificial samples are drawn (with replacement) from the sample itself. These bootstrap samples can be used in a variety of ways to estimate the significance level of a test statistic. While several bootstrap methods are very easy to use and are extremely flexible, bootstrap methods appear to be unreliable and should be used with caution.

4.1 INTRODUCTION

The last chapter and this chapter are concerned with using a random sample to test a hypothesis about the population from which it was drawn. A test statistic is selected that is sensitive to whether the null or alternative hypothesis is true. To simplify matters, I always define the test statistic so that it will tend to be large for a random sample if the alternative hypothesis about the population is true and to be small if the null hypothesis is true. Since random samples drawn from the same population will differ, there will be natural variability in the test statistic. Thus, relatively large values of the test statistic could occur solely by chance even if the null hypothesis is true. The significance level of a test statistic should be the probability that a hypothetical random sample drawn from the hypothetical null hypothesis population would yield a value of the test statistic at least as large as for the real sample. Bootstrap methods pioneered by Efron for estimating confidence intervals can be modified to estimate significance levels. There are indeed many bootstrap methods that could be used. Two of the methods are discussed in this chapter and two in the appendices, but you should be aware that there are other bootstrap methods that could be used and new methods are being almost continuously developed. The two methods discussed here were selected because they are very easy to use, are relatively easy to understand (and to explain), and seem to work as well as more complicated methods for purposes of estimating significance levels.

To fully explain bootstrap methods, a more formal approach than I have taken in the book to this point is unavoidable. Bear with me and concentrate. Let x be a hypothetical random sample from the null hypothesis population. The function $t(\cdot)$ defines the test statistic; for example, $t(x)$ is the value of the test statistic for the hypothetical random sample x. The hypothetical random sample x is a random variable, so $t(x)$ is also a random variable which has its own probability function. The probability that $t(x)$ would be greater than any particular value, say h, is called the "sampling distribution of $t(x)$" and is formally described by $\text{prob}(t(x) \geq h)$.

Now let x_0 be the real random sample from the real population; $t(x_0)$ is the value of the test statistic for that real sample. A hypothesis test consists of calculating how unusual $t(x_0)$ is relative to the sampling distribution of $t(x)$. That is, the significance of the test statistic ideally is $\text{prob}(t(x) \geq t(x_0))$ and the rule for rejecting the null hypothesis is:

$$\text{Reject if } \text{prob}(t(x) \geq t(x_0)) \leq \alpha$$

The problem in assessing a significance level thus reduces to estimating the sampling distribution of the test statistic under the null hypothesis, i.e., the probability distribution of $t(x)$. If the null hypothesis population can be completely

specified, then the sampling distribution can be estimated using Monte Carlo sampling as described in the previous chapter. The sampling distribution is estimated by drawing simulated random samples from the null hypothesis population. The significance level is essentially the proportion of simulated samples for which the value of the test statistic was at least as large as for the original sample.[1]

A major stumbling block in the application of the Monte Carlo method is the problem of specifying an appropriate null hypothesis population from which to simulate drawing random samples. While the researcher can always fall back on the standard assumptions made in conventional parametric tests (e.g., the population is Normal), sometimes such an assumption may seem courageous or even foolhardy and may not efficiently exploit all of the information in the sample.

In fact, given a sample from a population, the nonparametric maximum likelihood estimate of the population distribution is the sample itself.[2] The startling implication, which Efron [1979] first suggested, is that when the sample contains all of the available information about the population, why not proceed as if the sample is the population for purposes of estimating the sampling distribution of the test statistic? That is, apply Monte Carlo procedures, sampling with replacement from the sample itself.[3] This may seem to be an odd suggestion, but it appears to work surprisingly well in practice.

It is important to be very clear about how this bootstrap resampling is accomplished. A sample can be visualized as a matrix of values, with rows representing cases and columns representing variables. Ordinarily, bootstrap samples are constructed by drawing at random, with replacement, entire rows from the matrix of values. Stated differently, bootstrap samples are constructed by selecting entire cases from the original sample. The values of the various variables identified with a particular case stay together in the bootstrap sample.

Slipping back into formalism, let ${}_{B}x_0$ be a bootstrap sample drawn from the original sample x_0. Then $t({}_{B}x_0)$ is the value of the test statistic for the bootstrap sample ${}_{B}x_0$. Since ${}_{B}x_0$ is a random variable, $t({}_{B}x_0)$ is also a random variable. The probability that $t({}_{B}x_0)$ is at least as large as some value, say h, is called the "bootstrap sampling distribution of ${}_{B}x_0$." This bootstrap sampling distribution can be used in various ways to estimate the sampling distribution of $t(x)$.

---

[1] The idea behind conventional parametric tests is the same, but conventional parametric significance levels are mathematically derived. A conventional parametric significance level can be interpreted as the limiting significance level of a Monte Carlo hypothesis test as the number of Monte Carlo samples is increased without limit.

[2] Efron and Tibshirani [1986]. To be more precise, if the sample consists of the M observations $\{x_1, x_2, ..., x_M\}$, the maximum likelihood nonparametric estimator of the population distribution is the probability function that places probability mass 1/M on each of the observations x_i.

[3] Sampling with replacement is consistent with a population that is essentially infinite. While this may seem unrealistic in many settings, sampling without replacement from the sample would always reproduce the sample itself.

Two methods of estimating significance levels based on the bootstrap sampling distribution are discussed in this chapter. The "shift" method assumes that the bootstrap sampling distribution and the null hypothesis sampling distribution have the same shape but different central locations. The "Normal approximation" method assumes that the null hypothesis sampling distribution is Normal; the bootstrap sampling distribution is used only to estimate the variance of the Normal distribution. Two other bootstrap methods are briefly discussed in the appendices to this chapter. There are other bootstrap methods as well. Unfortunately, at present, all bootstrap methods for testing hypotheses must be regarded as speculative. Little is known about the performance of these methods except for very large samples. A bootstrap method may or may not yield reliable significance levels in a given situation, and it is difficult to tell whether or not it is reliable without knowing more about the population than what is contained in the sample. In general, the results of bootstrap tests should be interpreted with a great deal of caution. However, bootstrap tests have one great advantage over conventional parametric tests – they can be used in situations in which the conventional parametric sampling distribution of the test statistic is not known. For example, bootstrap resampling can be used to test the significance of the median from a small sample. Additionally, the two methods discussed in this chapter can be applied with very little effort to a test of just about any test statistic.

4.2 BOOTSTRAP THEORY

4.2.1 THE SHIFT METHOD

Recall that the null hypothesis sampling distribution is prob(t(x)≥h) and the bootstrap sampling distribution is given by prob(t($_B x_0$)≥h). x represents a hypothetical random sample from the null hypothesis population and $_B x_0$ represents a bootstrap random sample. The real, original sample is represented by x_0 and the value of the test statistic for that sample by t(x_0). Ideally, in a hypothesis test, the rule should be:

$$\text{Reject if } \text{prob}(t(x) \geq t(x_0)) \leq \alpha$$

The problem is how to specify the sampling distribution of t(x). In the shift method, it is assumed that the shapes of the bootstrap sampling distribution and the null hypothesis sampling distribution are the same – but their locations differ. To conduct a hypothesis test, the location of the bootstrap sampling distribution is shifted so that it is centered over the location where the distribution of t(x) should be centered.

Let τ be the expected value of the test statistic for random samples drawn from the null hypothesis population. This is where the sampling distribution of

$t(x)$ should be centered. And let ${}_B\tau$ be the expected value of the test statistic for bootstrap random samples. This is where the bootstrap sampling distribution is centered.

To center the bootstrap sampling distribution over the location where the null hypothesis sampling distribution should be centered, subtract from each value of $t({}_Bx_0)$ its expected value and then add back the expected value of $t(x)$. This transformation bodily shifts the whole bootstrap sampling distribution over so that it is centered over the location where the null hypothesis sampling distribution should be centered. An example of this process is illustrated in Figure 4.1. The bootstrap sampling distribution is depicted in the top panel. (In practice, the expected value of the test statistic for bootstrap samples, ${}_B\tau$, will usually be quite close to the value of the test statistic for the original sample, $t(x_0)$. The difference is exaggerated in the illustration.) The entire bootstrap sampling distribution is shifted to the left in the lower panel so that its mean coincides with the mean of the null hypothesis sampling distribution, τ. The significance level of the test statistic is the shaded portion of the bootstrap sampling distribution to the right of $t(x_0)$.

This hypothesis test can be formally derived as follows. Assume the null hypothesis and bootstrap sampling distributions have the same shapes:

$$\text{prob}(t(x)-\tau \geq h) \approx \text{prob}(t({}_Bx_0)-{}_B\tau) \geq h) \text{ for all } h.^4$$

This particular assumption leads to a test based on the bootstrap sampling distribution that is equivalent to a test based on the sampling distribution of $t(x)$.

$$\begin{aligned}\text{prob}(t(x) \geq t(x_0)) &= \text{prob}(t(x)-\tau \geq t(x_0)-\tau) \\ &\approx \text{prob}(t({}_Bx_0)-{}_B\tau \geq t(x_0)-\tau) \\ &= \text{prob}(t({}_Bx_0) \geq t(x_0)-\tau+{}_B\tau)\end{aligned}$$

Following the above chain of reasoning, the rule

$$\text{Reject if } \text{prob}(t({}_Bx_0) \geq t(x_0)-\tau+{}_B\tau) \leq \alpha$$

is equivalent to the rule

$$\text{Reject if } \text{prob}(t(x) \geq t(x_0)) \leq \alpha$$

And, of course, the null hypothesis should be rejected if $\text{prob}(t(x) \geq t(x_0)) \leq \alpha$.

[4] This is actually a stronger assumption than is required; the sampling distributions need only agree at a single value of h, namely, $h = t(x_0)-\tau$.

Thus, assuming that prob($t(x)-\tau \geq h$) $\approx$ prob($t({}_B x_0)-{}_B\tau) \geq h$), a test based on the bootstrap sampling distribution is equivalent to a test based on the null hypothesis sampling distribution.

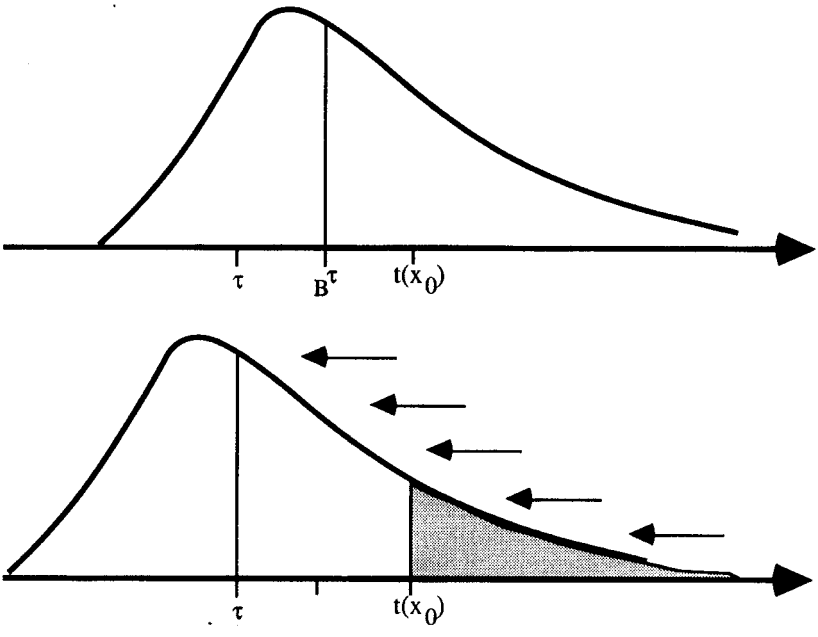

Figure 4.1 The bootstrap sampling distribution shift method

The significance level prob($t({}_B x_0) \geq t(x_0)-\tau+{}_B\tau$) can be estimated using Monte Carlo simulation. The quantity on the right-hand side of the inequality, $t(x_0)-\tau+{}_B\tau$, is a constant. For convenience, this constant will be called the "criterion value." $t(x_0)$ is the value of the test statistic for the original sample. The parameter τ is the expected value of the test statistic for hypothetical random samples drawn from the null hypothesis population.[5] The parameter ${}_B\tau$ is the expected value of the test statistic for bootstrap random samples. The sample mean of the bootstrap samples from the Monte Carlo simulation can be used as the estimate of ${}_B\tau$. The value $t({}_B x_0)$ is a random variable which depends on the bootstrap samples ${}_B x_0$ that are drawn from the original sample.

[5] The expected value of the test statistic for random values drawn from the null hypothesis population may not be known. In that case, a bootstrap test is not feasible. For that matter, a conventional parametric test isn't feasible either. A randomization test, on the other hand, does not require an estimate of τ.

The significance level can be estimated using usual Monte Carlo methods with only minor modifications. The test statistic is computed and saved for each bootstrap sample. After all of the bootstrap samples have been drawn, the sample mean of the bootstrap test statistics is computed, which in turn is used to calculate the criterion value. Then the value of the test statistic for each of the bootstrap samples is compared to the criterion value. The null hypothesis is rejected if $(nge+1)/(NS+1) \leq \alpha$, where nge is the number of bootstrap samples for which the value of the test statistic exceeded the criterion value and NS is the number of bootstrap samples. The method is very easy to implement and requires no thought once a test statistic has been selected.

As stated earlier, this method is equivalent to shifting the bootstrap sampling distribution and using it as if it is the null hypothesis sampling distribution. Note that $\text{prob}(t({}_B x_0) \geq t(x_0) - \tau + {}_B\tau)$ is equivalent to $\text{prob}(t({}_B x_0) - {}_B\tau + \tau \geq t(x_0))$. The bootstrap sampling distribution is shifted by subtracting its mean, ${}_B\tau$, and adding back the mean of the null hypothesis test statistic, τ.

4.2.2 THE NORMAL APPROXIMATION METHOD

Central limit theorems suggest that for many test statistics, if the sample is sufficiently large, the random variable $z = (t(x) - \tau)/\text{stddev}(t(x))$ is approximately standard Normal.[6] The term in the denominator of the random variable z, $\text{stddev}(t(x))$, is the standard deviation of the test statistic when sampling from the null hypothesis population. It may be difficult or impossible to analytically derive the value of this standard deviation without making very strong assumptions about the null hypothesis population. Efron [1979] suggests instead that the standard deviation of the bootstrap sampling distribution be used to estimate the standard deviation of the null hypothesis sampling distribution. Symbolically, we assume $\text{stddev}(t({}_B x_0)) \approx \text{stddev}(t(x))$.

Mechanically, this means bootstrap samples are drawn exactly as before, but they are only used to compute the sample standard deviation. In essence, we assume that the shape of the null hypothesis sampling distribution is Normal and the bootstrap sampling distribution is only used to estimate the variance of the Normal distribution.

The hypothesis test takes the following form:

$$\text{Reject if } \Phi(z \geq z_0) \leq \alpha$$

where $z_0 = (t(x_0) - \tau)/\text{stddev}(t({}_B x_0))$ and $\Phi(z \geq z_0)$ is the probability that a standard Normal random variable would be as large as z_0.

This hypothesis test can be derived as follows:

$$\text{prob}(t(x) \geq t(x_0)) = \text{prob}(t(x) - \tau \geq t(x_0) - \tau)$$

---

[6] The most familiar case is when t(.) is the sample mean of N observations that are drawn from the same population.

$$\text{prob}(t(x)\geq t(x_0)) = \text{prob}\left\{\frac{t(x)-\tau}{\text{stddev}(t(x))} \geq \frac{t(x_0)-\tau}{\text{stddev}(t(x))}\right\}$$

$$\approx \text{prob}\{z \geq (t(x_0)-\tau)/\text{stddev}(t(_B x_0))\}$$

$$\approx \Phi(z \geq z_0)$$

Like the bootstrap sampling distribution shift method, the Normal approximation method is very easy to use, and once a test statistic has been defined the procedure can be carried out automatically.

Template 4.1 in the Program Appendix of your choice can be used to assess the significance of virtually any test statistic using both the bootstrap sampling distribution shift method and the Normal approximation method. The two methods will usually produce very similar significance levels. As long as one of the methods is used, however, there is very little cost to using them both, so you might as well. Most of the template is self-explanatory. However, there is one detail that is quite important and that may be somewhat mysterious. An index variable is used to indicate which observations in the original sample are to be used when the test statistic is computed.[7] This index variable is initialized when the sample observations are read. The initialization sets the first element of the variable equal to 1, the second element equal to 2, and so on. Then the first time the ComputeStatistic subroutine is called, all of the observations in the sample will be included once and only once in calculation of the test statistic. Within the bootstrap resampling loop, new values are assigned to the elements of the index variable. Each of the integers between 1 and M has an equal chance of being assigned to each element of the index variable whenever the DrawBootstrapSample subroutine is called. The index vector identifies which observations in the original sample are included in the bootstrap sample.

Another detail that may be puzzling is that the values of the test statistics for the bootstrap samples are stored and then the program cycles back through all of them to determine the significance level (nge+1)/(NS+1). This is done because the criterion value is a function of the estimated bootstrap sampling distribution mean, which is not known until after all of the bootstrap samples have been drawn.

---

[7] The advantage of using such an index variable is that one subroutine can be used to compute the value of the test statistic for both the original data and the bootstrap samples. There are also some efficiencies to using such an index when the test statistic is a function of more than just one variable.

4.3 AN EXAMPLE

Suppose you have drawn a random sample of 20 values from a population and you are interested in testing whether the mean of the population is greater than zero. These sample values appear in Table 4.1.

Since the mean of the null hypothesis population is zero, the expected mean of any random sample is also zero (i.e., $\tau = 0$). The sample mean for the above data is 0.137 (i.e., $t(x_0) = 0.137$). What is the likelihood that the sample mean would have been this large if the null hypothesis is true?

Table 4.1

A random sample from a population

0.464, 0.060, 1.486, 1.022, 1.394, 0.906, 1.179, –1.501, –0.690, 1.372
–0.048, –1.376, –1.010, –0.005, 1.393, –1.787, –0.105, –1.339, 1.041, 0.279

The significance level of the sample mean is estimated using both the shift and Normal approximation methods in Program 4.1. As will usually be the case, the two methods give very similar answers. Be sure to note how the index variable is used in the computation of the test statistic.

To illustrate the versatility of the bootstrap sampling distribution shift method, suppose you are interested in testing whether the mean value of the observations is zero, after throwing out all observations more than 1.5 standard deviations from the sample mean. (It might, for example, be desirable to delete unusually large observations in a test in order to reduce the influence of outliers.) A program to test the significance of this test statistic is listed as Program 4.2.[8] The important point to note here is that these techniques can be easily applied to testing the significance of quite complicated test statistics with no modification in the template.

[8] It is important to note that the procedure applied to the bootstrap samples exactly mimics the procedure applied to the original sample. That is, for each bootstrap sample, the sample mean and the sample standard deviation are computed and then outliers are deleted based on that mean and that standard deviation. The outliers are not deleted based on the original sample's mean and standard deviation.

4.4 RELIABILITY OF BOOTSTRAP TESTS

Bootstrap methods can be used to estimate a significance level quickly and easily for virtually any test statistic; but, can these significance levels be trusted?

The data in Table 4.2 were actually drawn from a standard Normal population. We can obtain some notion of the reliability of bootstrap methods by drawing many, say 1000, independent samples from this population. For each sample, a test can be conducted on the hypothesis that the population mean is zero using both of the bootstrap procedures (with NS = 99) and, for comparison, a conventional t test. Since the population mean is in fact zero, if the test is reliable the null hypothesis should be rejected about 5% of the time at the 0.05 level and 10% of the time at the 0.10 level. The results are displayed in Table 4.3 for various sample sizes and for two alternative hypotheses: the mean is greater than zero and the mean is less than zero.

Table 4.2
Reliability of a bootstrap resampling test of the mean: 1000 samples of size M drawn from a standard Normal population (NS = 99)

	M=20	M=40	M=80	M=160
α=0.05				
Test that the mean is greater than it actually is				
t test	0.059	0.047	0.053	0.057
shift method bootstrap	0.076	0.064	0.068	0.064
Normal approx bootstrap	0.072	0.056	0.059	0.057
joint bootstrap	0.068	0.056	0.058	0.052
Test that the mean is less than it actually is				
t test	0.050	0.037	0.050	0.048
shift method bootstrap	0.064	0.047	0.048	0.053
Normal approx bootstrap	0.063	0.049	0.057	0.050
joint bootstrap	0.059	0.040	0.046	0.043
α=0.10				
Test that the mean is greater than it actually is				
t test	0.106	0.092	0.115	0.112
shift method bootstrap	0.117	0.095	0.112	0.114
Normal approx bootstrap	0.125	0.101	0.114	0.116
joint bootstrap	0.111	0.092	0.107	0.106
Test that the mean is less than it actually is				
t test	0.101	0.106	0.091	0.098
shift method bootstrap	0.113	0.109	0.093	0.095
Normal approx bootstrap	0.125	0.104	0.095	0.096
joint bootstrap	0.107	0.102	0.086	0.087

For example, when the sample consists of only 20 observations drawn from a standard Normal population, the null hypothesis is rejected 5.9% of the time at the 0.10 level when the parametric t test is used and the alternative hypothesis is that the mean is greater than it actually is. In contrast, the bootstrap shift method rejects the null hypothesis 7.6% of the time and the bootstrap Normal approximation method rejects 7.2% of the time. A slightly more conservative bootstrap test can be obtained by adopting the rule that the null hypothesis is rejected only if both the shift method and the Normal approximation method reject the null hypothesis. Using this "joint bootstrap" rule, the null hypothesis is rejected only 6.8% of the time. The conventional t test is exactly calibrated in this situation, so if the experiment were repeated an infinite number of times, the rejection rate would approach 5% for the conventional t test. This is not the case with the bootstrap tests. It is extremely unlikely that the bootstrap rejection rate would settle down to 5% if more samples were drawn.[9] Even so, the bootstrap rejection levels do not appear to be too far off the mark. While the rejection rates for the parametric t -test are correct for samples from the standard Normal population, the bootstrap rejection rates are reasonably close.

Similar experiments were conducted for two other populations where the assumptions of the t test are not satisfied. Since bootstrap methods do not assume a particular distribution for the population, they may perform better than the conventional t test when its population assumptions are violated.

The first population in this experiment is the standard Lognormal. Values of this population are defined by the relationship $x = \exp(z)$, where z is a standard Normal random variable. The mean of this population is $\sqrt{e}$. The values drawn from the second population, called the Lognormal error population, were generated by a compound process. With probability 0.8 the observation was zero and with probability 0.2, the observation was drawn from the standard Lognormal population.[10] The mean of the Lognormal error population is $0.2\sqrt{e}$.

In all hypothesis tests, the hypothesized mean under the null hypothesis was the actual mean of the population.

The results of those experiments are disclosed in Tables 4.3 and 4.4. The rejection rates for bootstrap methods and the conventional t test are incorrect for samples drawn from these populations.[11] When the alternative hypothesis is that the mean is less than it actually is, the null hypothesis is rejected less

[9] The probability is essentially zero of rejecting the null hypothesis at least 68 times out of 1000, when the probability of rejecting on any one trial is 0.05. Thus the asymptotic probability of rejection at the 0.05 level using the joint bootstrap test is greater than 0.05.

[10] Similar populations have been used to model the errors in recording transactions in accounting records. Ordinarily, there is no error when a transaction is recorded; and when there is an error, the amount of the error is itself a random variable.

[11] The tests reported in the table are one-sided. Had a two-tailed test been conducted, the null hypothesis would have been rejected about the correct number of times. This underlines the importance of conducting one-tailed tests when investigating the performance of procedures for testing hypotheses.

frequently than it should be. When the alternative hypothesis is that the mean is greater than it actually is, the null hypothesis is rejected too frequently.

Table 4.3
Reliability of a bootstrap resampling test of the mean: 1000 samples of size M drawn from a standard logNormal population* (NS = 99)

	M=20	M=40	M=80	M=160
α=0.05				
Test that the mean is greater than it actually is				
t test	0.008	0.010	0.015	0.022
shift method bootstrap	0.014	0.014	0.018	0.027
Normal approx bootstrap	0.011	0.013	0.017	0.024
joint bootstrap	0.009	0.009	0.013	0.021
Test that the mean is less than it actually is				
t test	0.153	0.149	0.124	0.109
shift method bootstrap	0.180	0.167	0.136	0.119
Normal approx bootstrap	0.167	0.161	0.124	0.113
joint bootstrap	0.165	0.158	0.119	0.107
α=0.10				
Test that the mean is greater than it actually is				
t test	0.032	0.038	0.062	0.067
shift method bootstrap	0.041	0.040	0.053	0.060
Normal approx bootstrap	0.049	0.051	0.063	0.068
joint bootstrap	0.037	0.037	0.049	0.055
Test that the mean is less than it actually is				
t test	0.210	0.211	0.178	0.152
shift method bootstrap	0.229	0.216	0.191	0.162
Normal approx bootstrap	0.223	0.217	0.183	0.159
joint bootstrap	0.219	0.211	0.180	0.154

*The values x are defined by $x = e^z$, where z is a random drawing from a standard Normal population. The mean of this population is $\sqrt{e}$.

When the sample size is increased, all of the methods are more reliable. This should be expected. If the sample size is increased without bound, eventually there will be very little to distinguish the sample from the population. If the sample is large enough to be indistinguishable from the population, then the sample will serve quite well as a surrogate for the population and bootstrap methods will usually work quite splendidly.[12]

[12] Bickel and Freedman [1981] and Singh [1981] investigate the asymptotic properties of the bootstrap. Interestingly enough, as Bickel and Freedman point out, the bootstrap does

Table 4.4
Reliability of a bootstrap resampling test of the mean: 1000 samples of size M drawn from a standard lognormal error population* (NS = 99)

	M=20	M=40	M=80	M=160
α=0.05				
Test that the mean is greater than it actually is				
t test	0.003	0.004	0.008	0.008
shift method bootstrap	0.006	0.013	0.012	0.017
Normal approx bootstrap	0.006	0.009	0.010	0.015
joint bootstrap	0.005	0.009	0.009	0.012
Test that the mean is less than it actually is				
t test	0.276	0.222	0.165	0.128
shift method bootstrap	0.328	0.249	0.184	0.151
Normal approx bootstrap	0.301	0.227	0.170	0.134
joint bootstrap	0.299	0.226	0.166	0.131
α=0.10				
Test that the mean is greater than it actually is				
t test	0.012	0.028	0.033	0.051
shift method bootstrap	0.021	0.033	0.037	0.043
Normal approx bootstrap	0.027	0.038	0.039	0.053
joint bootstrap	0.018	0.031	0.031	0.042
Test that the mean is less than it actually is				
t test	0.337	0.278	0.216	0.194
shift method bootstrap	0.363	0.296	0.225	0.200
Normal approx bootstrap	0.352	0.287	0.224	0.193
joint bootstrap	0.347	0.284	0.220	0.192

*The values x are defined by prob(x = 0) = 0.9; prob(x = e^z) = 0.1; where z is a random drawing from a standard Normal population. The mean of this population is $0.1\sqrt{e}$.

The conventional t test and bootstrap methods appear to be roughly equivalent in these particular experiments.[13] This is good news and bad news for the

not always provide asymptotically reliable confidence intervals. For example, the bootstrap is illsuited to assessing the confidence interval for the maximum value of a sample from a uniform distribution.

[13]It may be possible to modify the bootstrap test so that it is more reliable. Efron [1984] suggests elaborate procedures for improving bootstrap confidence intervals that are derived using a particular stochastic model. It appears to me that, using a similar line of reasoning, this particular stochastic model leads to the implication that, unlike the confidence interval, no adjustment of the hypothesis test is called for. Nevertheless, I tried a number of *ad hoc* procedures, including procedures similar to those used by Efron for adjusting confidence intervals, to improve the reliability of the bootstrap test. None worked.

bootstrap. It is good news in the sense that the bootstrap does so well when compared to the conventional parametric test, even though the t test is known to be robust. It is bad news in the sense that there is no apparent advantage to using the bootstrap test when testing the mean for these populations even though the assumptions of the t test are violated.[14] Since the bootstrap avoids making the strong distributional assumptions required for the t test, it might be hoped that it would outperform the conventional parametric test when its assumptions are violated. At least in these experiments, that didn't happen.

There are, however, two great advantages that the bootstrap methods possess. First, they can be applied more or less automatically. All that is required is a definition of the test statistic and the expected value of the test statistic for random samples drawn from the null hypothesis population. Second, bootstrap methods can be used to assess the significance of a test statistic even if its sampling distribution cannot be analytically derived. In a nutshell, bootstrap methods are far simpler to use and are much more flexible than conventional methods.

Nonetheless, there is the problem that bootstrap methods may not be reliable in a given situation. Like all Monte Carlo tests, a bootstrap test is a valid test of a null hypothesis. However, the null hypothesis may not be what the researcher really wants to test. Assuming that the original data are indeed a random sample, the null hypothesis in a bootstrap test is a joint hypothesis with two parts. When the shift method is used, the null hypothesis is that the expected value of the test statistic for random samples drawn from the null hypothesis population is τ and that the shape of the bootstrap sampling distribution approximates the underlying sampling distribution (i.e., $\text{prob}\{t(x)-\tau \geq h\} \approx \text{prob}\{t({}_B x_0)-{}_B\tau) \geq h\}$). The researcher is generally only interested in the first part of this joint hypothesis, but the null may be rejected because either part is false (or, of course, by chance).

The assumption that $\text{prob}\{t(x)-\tau \geq h\} \approx \text{prob}\{t({}_B x_0)-{}_B\tau) \geq h\}$ may or may not be warranted, depending on the test statistic that is used and the population from which the sample is drawn. For example, as discussed above, the bootstrap appears to be unreliable when the mean is tested in a small or moderately sized sample drawn from a Lognormal population or Lognormal error population. The unreliability of the shift method for a test of the mean from these two populations implies that bootstrap sampling distribution does not always satisfactorily approximate underlying sampling distribution.

Figure 4.1 can be used to illustrate one reason why the assumption underlying the shift method may be inappropriate. In the illustration, the bootstrap sampling distribution was shifted to the left. Some values of the test statistic that are under the curve after the shift has been made may be impossible. For

[14] Power comparisons would be misleading. Since the bootstrap test falsely rejects the null hypothesis more frequently than the t test, it is likely that it is technically the more powerful test. However, this power does not come without cost.

example, if the test statistic is the standard deviation of the observations, a negative value of the test statistic is impossible. After shifting, the sampling distribution is clearly misspecified in the left tail if some of the probability mass is over negative values. And if the distribution is misspecified to the left of the origin, it must also be misspecified elsewhere since the total area under the curve must be one.

In the case of the Normal approximation method, the null hypothesis is that (1) the expected value of the test statistic for random samples drawn from the null hypothesis population is τ; (2) $(t(x_0)-\tau)/\text{stddev}(t(x))$ is approximately standard Normal; and (3) the standard deviation of the bootstrap sampling distribution is approximately equal to the standard deviation of the underlying sampling distribution. The only part of this joint hypothesis in which the researcher is really interested is the first part, but the null hypothesis could be rejected because any of the three parts are false. There is some comfort in the observation that the Normal approximation is likely to be adequate for very large samples; but, it may not be satisfactory for small samples. In particular, for small samples, the sampling distribution of the test statistic can be strongly asymmetric; whereas, of course, the Normal distribution is symmetric.

It is possible that general and useful statements can be made concerning situations in which these approximations are exact enough; however, thus far, statisticians have confined their work to the asymptotics (i.e., very large sample properties) of bootstrap methods. Except in pathological cases, if the sample is large enough, bootstrap methods will work satisfactorily simply because the sample becomes indistinguishable from the population if the sample is large enough. On the other hand, the small sample properties of bootstrap methods have been investigated only via simulation for specific cases. It would be risky to extrapolate from those cases to others. I suspect that one would have to know something about the population from which the sample was drawn, apart from the sample itself, in order to be able to conclude with reasonable certitude that the bootstrap is reliable. But, of course, if more were known about the population that just the sample, such information should be taken into account and the bootstrap – which uses just the information in the sample – might be inefficient.

Because of the unknown reliability of the bootstrap method for any given case, it would be not be prudent to base inferences solely on the results of a bootstrap test – unless there is no alternative. There would be no alternative if, for example, a conventional parametric sampling distribution of the test statistic is not known and a randomization test is not feasible.[15]

[15] A necessary condition to conduct a randomization test is that there are at least two variables so that one can be shuffled relative to the other.

4.5 CONCLUSION

Bootstrap methods are used to estimate the sampling distribution of the test statistic by drawing artificial samples from the original sample itself. This procedure is theoretically justifiable when the sample contains all of the available information about the population. Two methods of applying the bootstrap approach to hypothesis testing were discussed – the bootstrap sampling distribution shift method and the Normal approximation method. These methods are very easy to use and very flexible. All that is required is a well-defined test statistic and a specified value of the test statistic under the null hypothesis. However, a bootstrap hypothesis test may not be reliable. And, unfortunately, not enough basic research has been done to characterize when the methods can be expected to be reliable.[16] Until the results of such research are available, use of the bootstrap should be reserved for situations in which other tests are inappropriate.

[16] In specific applications, simulations can be performed to assess the reliability of the method, but such simulations will only be convincing when the population is available for repeated sampling. Biddle, Bruton, and Siegel [1986], for example, suggest that a bootstrap method be used by auditors to assess confidence intervals for errors in accounting records. This would be a dangerous suggestion without some evidence that the bootstrap method they propose actually works in real accounting populations. They take the necessary step of simulating the performance of their bootstrap method using complete accounting records from several firms. To the extent that the accounting records of other firms resemble those of the firms used by Biddle, Bruton, and Siegel, their results can be extrapolated to other firms.

APPENDIX 4A

THE SAMPLE TRANSFORMATION METHOD

The sample transformation method is, at first glance, the most intuitive application of the bootstrap idea to hypothesis testing. In the sample transformation method, the sample values are transformed so that the sample has the characteristic assumed, under the null hypothesis, for the population from which the sample was supposedly drawn. After this transformation, the sample can "stand in" or "proxy" for the population in the usual Monte Carlo simulation of repeated sampling.

Take, for example, the data in Table 4.1. In this case, the null hypothesis might be that the mean of the population is zero, so the sample values are adjusted so that the mean of the sample is zero. This could be done in a variety of ways, but the most obvious is simply to subtract the sample mean from each of the values in the sample. After this simple transformation, the sample has the characteristic that is assumed for the population; namely, a mean of zero. The sampling distribution of the test statistic under the null hypothesis is estimated by drawing samples from this adjusted sample, which proxies for the null hypothesis population.

I have, however, glossed over a difficulty in implementing the sample transformation method. It may not be obvious how to adjust the sample so that it has the characteristic assumed for the population. Suppose that the following seven observations have been drawn from a population: $-2, -1, 2, 2, 4, 5, 8$ and that the test statistic is the median. The null hypothesis might be that the data are a random sample from a population with a zero median. The sample median is 2. Now how should this sample be transformed so that its median is 0? Both of the observations coded 2 would have to be recoded. Should both of them be recoded as 0? Should one of them be recoded as 0 and the other as, say, -1? Should all of the observations be recoded by subtracting 2? The "correct" transformation is not obvious.

Due to this kind of ambiguity, applications of the sample transformation method are limited. In contrast, the sampling distribution shift and Normal approximation methods can be applied almost automatically. When using either of these latter two methods, there is no need to decide how to adjust a sample so that it conforms to assumptions made about the population under the null hypothesis.

APPENDIX 4B

THE BOOTSTRAP RANDOMIZATION METHOD

The bootstrap randomization method is a hybrid of randomization and the bootstrap approaches. It could be used in cases where the hypothesis is that the data are a random sample from a population in which two or more of the variables are stochastically independent. Mechanically, this method is very similar to an approximate randomization test. To simplify matters, suppose there are just two variables. Instead of simply shuffling one of the variables, bootstrap resampling is conducted for one or both of the variables. If one of the variables is considered fixed, then bootstrap samples are drawn for the other variable.[17] If both of the variables are considered to be random variables, then bootstrap samples are drawn for each of the variables independently.

Practically, an approximate randomization test and a bootstrap randomization test will usually agree concerning whether their respective null hypotheses should be rejected. The difference between the two tests is in the interpretation, which has to do with the null hypotheses the tests address. In the case of the approximate randomization test, the null hypothesis is that the variables are independent, which implies that all permutations of one of the variables relative to the other were equally likely. In the case of the bootstrap randomization test, the null hypothesis is that the data are a random sample from a population in which the variables are stochastically independent and that the marginal distributions of the variables in the sample satisfactorily approximate the marginal distributions of the variables in the population. Thus, the bootstrap randomization test might be preferred to the approximate randomization test when the researcher wants to make a direct inference about a characteristic of a population (i.e., that variables are independent in the population) based upon a random sample.[18] Unfortunately, it may be difficult to tell whether the marginal distributions of the variables in the sample satisfactorily approximate the marginal distributions of the variables in the population without knowing more about the population than is contained in the sample. Much work needs to be done by statisticians before this method can be used with confidence without corroborating evidence.

---

[17] Efron and Gong [1983], Freedman [1981], and Freedman and Peters [1984a, 1984b] use bootstrap methods to assess regression standard errors and confidence intervals. Marais [1984] adapts these bootstrap methods to test hypotheses concerning regression statistics. The "percentile" method used by Marais is an example of the bootstrap randomization method and it shows some promise in regression settings where the reliability of conventional parametric tests are suspect.

[18] In the case where the characteristic of the population is known, the researcher may want to test whether the data were drawn at random from the population.

A prudent course of action would be to run an approximate randomization test whenever a bootstrap randomization test is run. As suggested above, the outcomes of the tests (in terms of rejection or nonrejection of the null hypothesis) will only rarely differ. The approximate randomization test has the advantage that it is always a valid test of the simple hypothesis that all permutations were equally likely. The bootstrap randomization test is a valid test of the joint hypothesis that the data are a random sample from a population in which the variables are stochastically independent and that the marginal distributions of the variables in the sample satisfactorily approximate the marginal distributions of the variables in the population. The bootstrap randomization test may not, however, in any given situation be a reliable test of the simpler hypothesis that the data are a random sample from a population in which the variables are stochastically independent.

Conclusion

In this chapter the three major computer-intensive methods for assessing the significance of a test statistic are reviewed and are contrasted with each other and with conventional parametric methods. The method that should be used largely depends on the nature of the null hypothesis. For example, if a researcher is interested in testing the null hypothesis that variables are unrelated, an approximate randomization test should be used in preference to other computer-intensive methods and to conventional parametric methods.

5.1 APPROXIMATE RANDOMIZATION TESTS

Conceptually, approximate randomization tests are the simplest of all statistical methods. The null hypothesis is that one variable (or set of variables) is unrelated to another variable (or set of variables). To estimate the probability distribution of any test statistic under this null hypothesis, you simply shuffle one variable (or set of variables) relative to the others and recompute the value of the test statistic. This procedure ensures that the variables are unrelated.

The characteristics of approximate randomization tests are summarized in Table 5.1.

Table 5.1
Summary of approximate randomization tests

Null Hypothesis

One variable (or set of variables) is unrelated to the other variable(s). Technically, all permutations of one variable (or set of variables) relative to the other variable(s) were equally likely; thus, the data are a sample of size one from the set of all possible permutations.

Procedure

Approximate the distribution of the test statistic under the null hypothesis that the variables are unrelated by repeatedly shuffling one variable (or set of variables) relative to the other variable(s) and recomputing the test statistic.

Advantages

- Can be used to assess the significance of any test statistic.
- The data can be drawn from any population.
- The data need not be a random sample.

Disadvantages

- Cannot be formally used to drawn an inference about a population from which a random sample has been drawn.
- All permutations may not have been equally likely for reasons that are not of interest to the researcher. Therefore, the null hypothesis may be rejected for the wrong reason.

Approximate randomization tests have a number of important advantages over conventional parametric tests. First, the approximate randomization method can be used to assess the significance of any test statistic, even those for which the conventional sampling distribution is not known. Second, the data can be drawn from any population. In particular, the data need not be a random sample (indeed, the data can be the population). On the other hand, if your purpose is to draw an inference about a population from which a random sample

has been drawn, then a conventional parametric test or another computer-intensive method may be more appropriate. Nevertheless, since random samples are, on average, representative of the populations from which they are drawn, if the null hypothesis is rejected using a randomization test on a random sample, it is likely that the null hypothesis would be rejected as well if the test were repeated on the population.

An approximate randomization test is a valid test of the null hypothesis that all permutations of one variable (or set of variables) relative to the other variable(s) were equally likely. A distinction needs to be drawn, however, between the null hypothesis you are really interested in testing and the null hypothesis the test actually addresses. It may be that all permutations of the data were not equally likely – for reasons that do not interest you. For example, suppose you would like to test whether corn prices are related to corn production (economists maintain that prices should increase when production declines and decrease when production goes up). To test this hypothesis, corn prices might be shuffled relative to per capita corn production across a number of years. However, corn prices are likely to drift upward over time due to general price inflation and corn production per capita can drift due to changes in population or technology over time. Thus, there is likely to be a correlation between corn prices and corn production per capita for reasons having little to do with the hypothesis of real interest to the researcher. In general, caution is advised whenever variables are shuffled relative to each other in a way that breaks a temporal connection.[1]

5.2 CONVENTIONAL PARAMETRIC METHODS

The characteristics of conventional parametric methods are summarized in Table 5.2. Conventional parametric methods for testing hypotheses presume the data are a random sample from a particular population. For example, in a conventional t test of a mean, it is assumed that the data are a random sample from a Normal population. Statisticians use mathematical analysis to answer the question, "What is the likelihood that the value of the test statistic would have been as large as x when a random sample is drawn from this population?" Unfortunately, analytical techniques sometimes are incapable of providing an answer to that question. In those cases, Monte Carlo simulation is often used to empirically estimate the sampling distribution of the test statistic.

There are two major drawbacks to relying on conventional parametric methods. First, you may not know the sampling distribution of the test statistic you would like to use. Second, while a conventional parametric test is a valid

[1] This same problem exists with respect to naive application of conventional parametric tests to time series data as well, and the solutions to the problem are the same. When dealing with corn prices, for example, it would be wise to deflate nominal corn prices with a price level index to adjust for general inflation. Diagnostic tests can then be applied to the data to detect problems such as autocorrelation.

test of the null hypothesis that the data are a random sample from a particular population, that may not be the null hypothesis you would really like to test. For example, you may be interested in whether the mean of the population is zero, but a conventional t test of the mean also implicitly tests whether the population is Normal. Therefore, you may reject the null hypothesis for the wrong reason.

Table 5.2
Summary of conventional parametric tests

Null Hypothesis

The data are a random sample from a particular population.

Procedure

Mathematically derive the sampling distribution of the test statistic. If the sampling distribution has already been derived and tables exist, simply look up the significance level of the test statistic in a table.

Advantages

- If the population conforms to the assumptions required to derive the sampling distribution, no other method can do any better.
- Excellent software is available for testing many statistics.
- Requires much less computer time than computer-intensive methods.

Disadvantages

- The assumptions required to derive the sampling distribution may not be valid. Therefore, the null hypothesis may be rejected for the wrong reason.
- The exact sampling distribution may not have been derived for the test statistic the researcher would like to use.

There are some distinct advantages to using conventional parametric methods. The biggest advantage is that when the assumptions of the conventional parametric test are satisfied, no other method can do any better at assessing the significance of a given test statistic. Additionally, conventional parametric tests require much less computer time than computer-intensive methods. As a rule of thumb, computer-intensive methods take on the order of NS times as long to run as conventional parametric tests, where NS is the number of artificial samples that are used to estimate the sampling distribution. When data sets are very large, this can be a very big advantage – particularly since the validity problems associated with using conventional parametric methods tend to diminish as the sample size becomes large.

5.3 MONTE CARLO SAMPLING

The characteristics of the Monte Carlo sampling method are summarized in Table 5.3. The null hypothesis in a Monte Carlo sampling test is the same as in a conventional parametric test; the data are a random sample from a particular population. The sampling distribution of the test statistic is estimated by generating artificial random samples from a computer model of the null hypothesis population.

Table 5.3
Summary of Monte Carlo sampling

Null Hypothesis
The data are a random sample from a particular population.

Procedure
Empirically approximate the sampling distribution of the test statistic by drawing simulated random samples from the null hypothesis population. For each simulated random sample, compute the test statistic.

Advantages
• Can be used even when the exact sampling distribution of the test statistic has not been derived.

Disadvantages
• The assumptions required to fully specify a population may not be valid. Therefore, the null hypothesis may be rejected for the wrong reason.

The only advantage of Monte Carlo sampling over conventional parametric methods is that it can be used even when the sampling distribution of the test statistic has not been analytically derived.

Like conventional parametric tests, significance levels resulting from Monte Carlo tests may be misleading. In order to conduct a Monte Carlo test, the population has to be completely characterized. However, it is usually the case that only one aspect of the population is really of interest (e.g., its mean). The null hypothesis may therefore be rejected because the population from which the random sample was drawn differs from the model of the population used in Monte Carlo sampling in ways that are substantively uninteresting (e.g., the population is assumed to be Normal, but isn't).

5.4 BOOTSTRAP RESAMPLING (DISTRIBUTION SHIFT METHOD)

The characteristics of the bootstrap resampling distribution shift method are summarized in Table 5.4.

Table 5.4
Summary of bootstrap resampling (distribution shift method)

Null Hypothesis

The data are a random sample, the mean of the sampling distribution of the test statistic is a specific value, and the shape of the sampling distribution of the test statistic is well approximated by the bootstrap sampling distribution.

Procedure

Estimate the shape of the sampling distribution by repeatedly sampling (with replacement) from the sample itself.

Advantages

- Can be used when the sampling distribution of the test statistic has not been derived.
- The null hypothesis population does not have to be defined (the sample serves as a proxy for the population).
- The procedure is entirely automatic once a test statistic has been specified.

Disadvantages

- The assumption that the shape of the sampling distribution is well approximated by the bootstrap sampling distribution may not be valid. Therefore, the null hypothesis may be rejected for the wrong reason.

When the distribution shift method is used to assess significance, the null hypothesis is that the data are a random sample, the mean of the sampling distribution of the test statistic is a specific value, and the shape of the sampling distribution of the test statistic is well approximated by the shape of the bootstrap sampling distribution. The test is carried out by sampling with replacement from the sample and recomputing the value of the test statistic for the bootstrapped sample. This procedure is repeated many times, creating an estimate of the bootstrap sampling distribution. It is assumed that the shape of this bootstrap sampling distribution is the same as the shape of the sampling distribution when drawing samples directly from the population; the distributions differ only with respect to their means.

The significance level is estimated by centering the bootstrap sampling distribution over the hypothesized expected value of the underlying sampling distribution and computing the mass under the distribution to the right of the actual value of the test statistic for the real sample.

Despite the somewhat lengthy description, this method is extremely easy to use. All you have to do is define the test statistic and decide what its expected value should be under the null hypothesis – everything else is entirely automatic.

The advantage of this method over Monte Carlo sampling is that it is unnecessary to define the null hypothesis population. Essentially, the sample stands in for the population in the test. The advantage of this method over conventional parametric methods is that a significance level can be estimated for any test statistic – even when the exact sampling distribution of the test statistic has not been derived.

The principal disadvantage of this method is that the null hypothesis may be rejected because the shape of the sampling distribution is not well-approximated by the shape of the bootstrap sampling distribution rather than because the expected value of the test statistic differs from the value that is hypothesized. Unfortunately, it is difficult to tell whether the shape assumption is reasonable or not by just looking at the sample. And if more is known about the population than is contained in the sample, that information should be used and not discarded as it would be in a routine bootstrap test.

On occasion, bootstrap tests have been used as a check on the validity of a conventional parametric test when it is known that assumptions underlying the conventional parametric test are violated. Unfortunately, it appears that bootstrap tests and conventional parametric tests can fail under similar conditions, so this use of bootstrap tests is of questionable utility. Bootstrap tests are most useful in situations where a conventional parametric test does not exist (i.e., the sampling distribution of the test statistic isn't known).

5.5 BOOTSTRAP RESAMPLING (NORMAL APPROXIMATION METHOD)

The characteristics of the bootstrap resampling Normal approximation method are summarized in Table 5.5. The remarks concerning the distribution shift method in the previous section apply here, except that instead of assuming that the shape of the bootstrap sampling distribution approximates the shape of the underlying sampling distribution, it is assumed that the variance of the bootstrap sampling distribution approximates the variance of the underlying sampling distribution which in turn is assumed to be Normal. As before, this critical assumption may or may not be valid, and it is difficult to tell whether it is valid without knowing more about the population than just the information contained in the sample.

Table 5.5
Summary of bootstrap resampling (Normal approximation method)

Null Hypothesis

The data are a random sample from a Normal population, the mean of the sampling distribution of the test statistic is a specific value, and the variance of the sampling distribution is well approximated by the variance of the bootstrap sampling distribution.

Procedure

Estimate the variance of the sampling distribution by sampling (with replacement) from the sample itself.

Advantages

- Can be used when the sampling distribution of the test statistic has not been derived.
- The null hypothesis population does not have to be defined (the sample serves as a proxy for the population).
- The procedure is entirely automatic once a test statistic has been specified.

Disadvantages

- The assumption that the sampling distribution is Normal and that its standard deviation is well approximated by the standard deviation of the bootstrap sampling distribution may not be valid. Therefore, the null hypothesis may be rejected for the wrong reason.

5.6 SELECTING A METHOD FOR ASSESSING SIGNIFICANCE

Once you have selected a test statistic, how should its significance be assessed? Before addressing this question, it should be acknowledged that statisticians might object to separating the selection of a test statistic from the question of how its significance will be assessed. In theory the test statistic and method of assessing its significance should be selected simultaneously to maximize the power of the test. Apart from the difficulty of doing this even under ideal conditions, there is the added complication that power is not the only consideration. The different methods of assessing significance have different likelihoods of falsely rejecting the null hypothesis in which the researcher is really interested. That is, some methods are more reliable than others. I have tried to informally balance issues of reliability and power in the discussion that follows.

Each of the methods discussed above for assessing the significance of a test statistic provides a valid test of a null hypothesis. However, the null hypothesis that is actually tested differs from one method to another. You should be very careful to select from among the methods that could be used, the one that does

the best job of testing the null hypothesis you would really like to see tested. Selection of a method for assessing the significance of a test statistic implicitly involves the selection of a null hypothesis.

Advice concerning selection of a method is presented in condensed form in Figure 5.1. Throughout this section, I assume that a computer-intensive solution is feasible for your problem. If your data set is extraordinarily large or your test statistic is extraordinarily complicated, calculation of the test statistic may require inordinate amounts of computer time. In that case, a computer-intensive approach may not be practical. However, in most fields, this would be quite unusual.

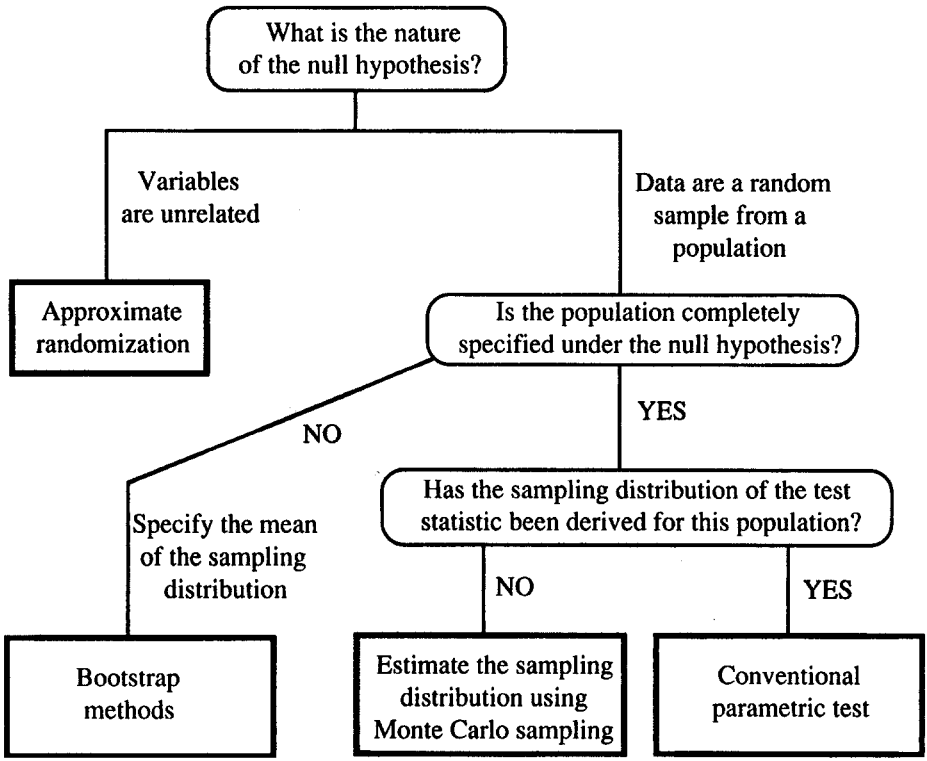

Figure 5.1 Selecting a method for assessing significance

The following discussion should also be tempered by two considerations. First, it is often at present much easier to use conventional methods than computer-intensive methods simply because computer-intensive methods have not yet been incorporated into statistical software packages. Second, it will usually take more effort to communicate the results of computer-intensive test to others, particularly those who have been trained in conventional parametric techniques, since readers are likely to be unfamiliar with the techniques. I don't expect either of these drawbacks to persist for long.

After the test statistic has been selected, the first step is to decide the nature of the null hypothesis you want to test. There are really only two kinds of null hypotheses from which to choose. Either the null hypothesis is that variables are unrelated to each other or the null hypothesis is that the data are a random sample from a population. If you really would like to test whether one variable (or set of variables) is unrelated to another variable (or set of variables), then the approximate randomization method should be used. If you really would like to test whether the data are a random sample from a population, then one of the other methods should be used.

If you want to test whether the data are a random sample from a population, then the next question is whether you want to specify the population in enough detail that the sampling distribution of the test statistic can be estimated by drawing samples from a computer model of the population. If you don't mind invoking the additional assumptions required to build a complete model of a population, then you can proceed with Monte Carlo sampling. Preferably, if the sampling distribution of the test statistic has been derived for that population, you can consult a statistics text for the significance level of the test statistic. Or, you could try to analytically derive the sampling distribution of the test statistic.

If, on the other hand, you would prefer not to assume that the population distribution is Normal or whatever, you could use a bootstrap method to estimate the significance level of the test statistic. To do this, the only additional step required is to specify the mean of the sampling distribution under the null hypothesis. However, since bootstrap methods are rather speculative at this point, I would suggest that you reevaluate the nature of the null hypothesis you would like to test. If, after all, a test of whether variables are unrelated would be appropriate, you would usually be better off using the approximate randomization method to assess significance than to use bootstrap methods.

REFERENCES

Baker, Frank B. and Raymond O. Collier Jr. (1966), "Some Empirical Results on Variance Ratios Under Permutation in the Completely Randomized Design," *Journal of the American Statistical Association*, Sept., 813-820

Barnard, G. A. (1963), *Journal of the Royal Statistical Society B*, 25, 294

Bickel, Peter J. and David A. Freedman (1981), "Some Asymptotic Theory for the Bootstrap," *Annals of Statistics*, 9, 1196-1217

Biddle, Gary, Carol Bruton and Andrew Siegel (1986) "Computer-Intensive Statistics in Auditing," working paper, University of Washington, School of Business, November

Blanchard, Garth, Chee Chow, and Eric Noreen (1986), "Information Asymmetry, Incentive Schemes, and Information Biasing: The Case of Hospital Budgets Under Rate Regulation," *Accounting Review*, Jan., 1-15

Bowen, Robert, Eric Noreen, and John Lacey (1981), "Determinants of the Decision by Firms to Capitalize Interest," *Journal of Accounting and Economics*, 3, August, 151-179

Box, G.E.P. and S.L. Andersen (1955), "Permutation Theory in the Derivation of Robust Criteria and the Study of Departures from Assumption," *Journal of the Royal Statistical Society Series B*, 17, 1-34

Boyett, James M. and J. J. Shuster (1977), "Nonparametric One-Sided Tests in Multivariate Analysis with Medical Applications," *Journal of the American Statistical Association*, September, 665-668

Campbell, Donald T. and Julian C. Stanley (1963), *Experimental and Quasi-Experimental Designs for Research*, Rand McNally College Publishing Co, Chicago

Chung, J.H. and D.A.S. Fraser (1958), "Randomization Tests for a Multi-Variate Two-Sample Problem," *Journal of the American Statistical Association*, 53, 729-735

Cooke, D., A,H. Craven, and G.M. Clarke (1982), *Basic Statistical Computing*, Edward Arnold, Scotland

Daniels, H.E. (1944), "The Relation Between Measures of Correlation in the Universe of Sample Permutations," *Biometrika*, 33, 129-135

Dwass, M. (1957), "Modified Randomization Tests for Nonparametric Hypotheses," *Annals of Mathematical Statistics*, 28, 181-187

Edgington, Eugene S. (1969), "Approximate Randomization Tests," *Journal of Psychology*, 72, 143-149

Edgington, Eugene S. (1970), "Hypothesis Testing Without Fixed Levels of Significance," *Journal of Psychology*, 76, 109-115

Edgington, Eugene S. and Ezinga, G. (1978), "Randomization Tests and Outlier Scores," *Journal of Psychology*, 99, 259-262

Edgington, Eugene S.(1980), *Randomization Tests*, Marcel Dekker, New York

Efron, Bradley (1979), "Bootstrap Methods: Another Look at the Jackknife," *Annals of Statistics*, 7, 1-26

Efron, Bradley and Gail Gong (1983), "A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation," *American Statistician*, 37, Feb., 36-48

Efron, Bradley (1984), "Better Bootstrap Confidence Intervals," LCS Technical Report No. 14, Department of Statistics and Stanford Linear Accelerator

Efron, B. and R. Tibshirani (1984), "Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy," *Statistical Science*, 1, 54-77

Fisher, Ronald (1966), *The Design of Experiments, Eighth Edition*, Hafner Publishing Co, New York

Foutz, Robert V. (1980) "A Method for Constructing Exact Tests from Test Statistics that Have Unknown Null Distributions," *Journal of Statistics and Computer Simulation*, 10, 187-193

Freedman, David (1981), "Bootstrapping Regression Models," *Annals of Statistics*, 9, 1218-1228

Freedman, David and Stephen C. Peters (1984a), "Bootstrapping a Regression Equation: Some Empirical Results," *Journal of the American Statistical Association*, 79, March, 97-105

Freedman, David and Stephen C. Peters (1984b), "Bootstrapping an Econometric Model: Some Empirical Results," *Journal of Business and Economic Statistics*, 2, April, 150-158

Hoeffding, W. (1951), "A Combinatorial Central Limit Theorem," *Annals of Mathematical Statistics*, 22, December, 169-192

Hoeffding, W. (1952), "The Large-Sample Power of Tests Based on Permutations of Observations," *Annals of Mathematical Statistics*, 23, 169-192

Hollander, Myles and Douglas A. Wolfe (1973), *Nonparametric Statistical Methods*, John Wiley and Sons, New York

Hope, A. C. A. (1968), "A Simplified Monte Carlo Significance Test Procedure," *Journal of the Royal Statistical Society*, 30, 582-598

Kempthorne, Oscar (1966), "Some Aspects of Experimental Inference,' *Journal of the American Statistical Association*, March, 11-34

Kempthorne, Oscar and T.E. Doerfler (1969), "The Behavior of Some Significance Tests Under Experimental Randomization," *Biometrika*, 56, 231-248

Klauber, M.R. (1971), "Two-Sample Randomization Tests for Space-Time Clustering," *Biometrics*, 27, March, 129-142

Lehmann, E.L. and C. Stein (1949), "On the Theory of Some Nonparametric Hypotheses," *Annals of Mathematical Statistics*, 20, 28-45

Lipset, Seymour (1960), "The Emergence of the One-Party South – The Election of 1860," in *Political Man*, Doubleday and Company, New York, 344-354

Marais, M. Laurentius (1984), "An Application of the Bootstrap Method to the Analysis of Squared, Standardized Market Model Prediction Errors," *Journal of Accounting Research*, Supplement, 34-54

Marriot, F.H.C. (1979), "Barnard's Monte Carlo Tests: How Many Simulations?," *Applied Statistics*, 28, 75-77

Noether, Gottfreid E. (1949), "On a Theorem by Wald and Wolfowitz," *Annals of Mathematical Statistics*, 20, September, 455-458

Noreen, Eric and James Sepe (1981), "Market Reactions to Accounting Policy Deliberations: The Inflation Accounting Case," *Accounting Review*, April, 253-269

Noreen, Eric (1986), "On the Validity of Randomization Tests," working paper, School of Business, University of Washington

Pearson, E.S. (1937), "Some Aspects of the Problem of Randomization, *Biometrika*, 29, 53-64

Pitman, E.J.G. (1937), "Significance Tests Which May Be Applied to Samples from Any Population I and II," *Journal of the Royal Statistical Society*, Supplement 4, 119-130, 225-232

Press, William H., Brian P. Flannery, Saul A. Teukolsky, and William T. Vetterling (1986), *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge

Rubinstein, Reuven Y. (1981), *Simulation and the Monte Carlo Method*, John Wiley and Sons, New York

Scheffe, Henry (1943), "Statistical Inference in the Non-Parametric Case," *Annals of Mathematical Statistics*, December, 305-332

Siegel, Sidney (1956), *Nonparametric Statistics for the Behavioral Sciences*, McGraw-Hill, New York

Singh, Kesar (1981), "On the Asymptotic Accuracy of Efron's Bootstrap," *Annals of Statistics*, 9, 1187-1195

Tsutakawa, R.K. and B.L. Yang (1974), "Permutation Tests Applied to Antibiotic Drug Resistance," *Journal of the American Statistical Association*, 69, 87-92

Wald, A. and J. Wolfowitz (1944), "Statistical Tests Based on Permutations of the Observations," *Annals of Mathematical Statistics*, December, 358-372