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.

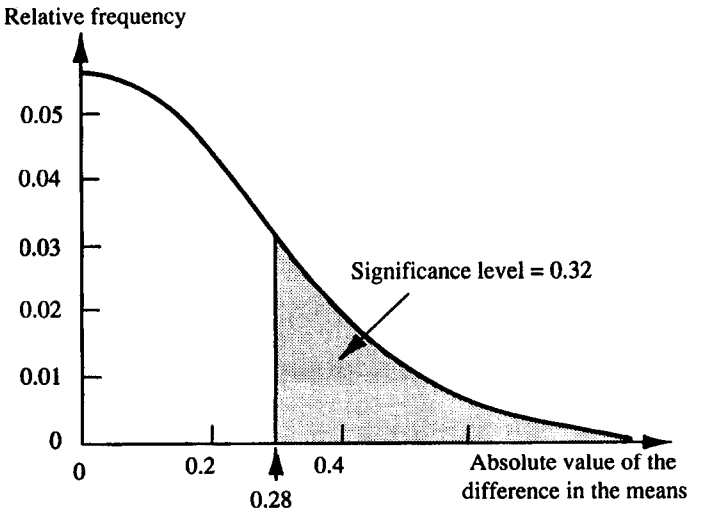

Figure 2.1 Histogram of the absolute value of the difference in mean grades between transfer and nontransfer students

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."

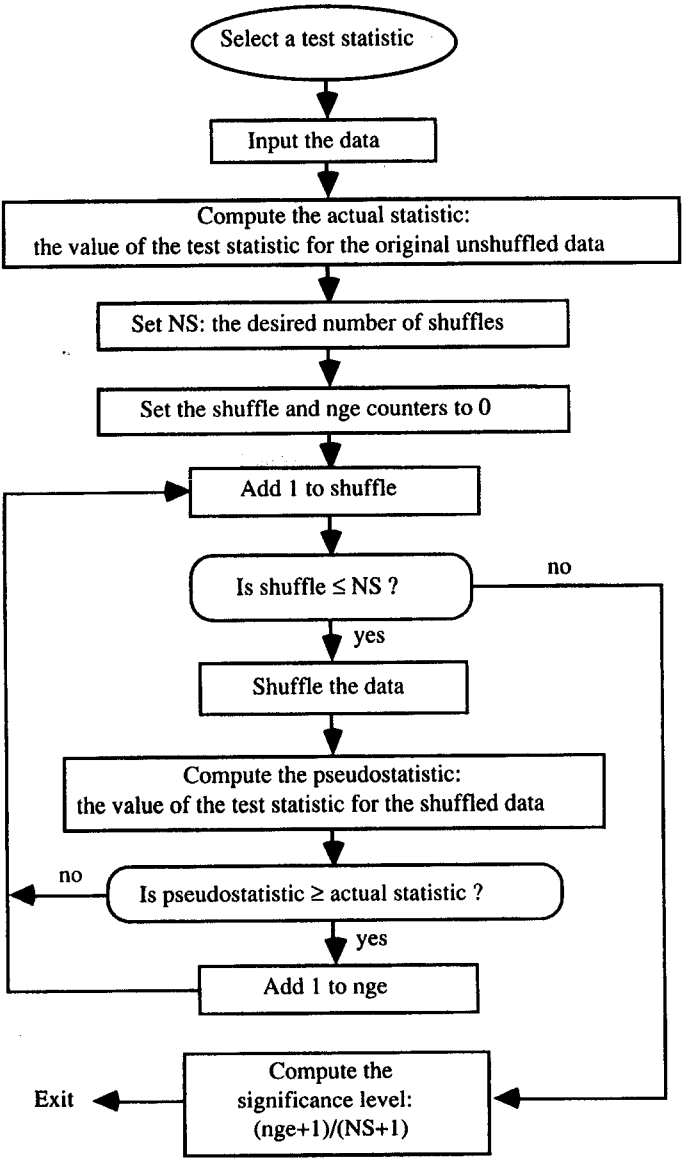

Figure 2.2 Flowchart for an approximate randomization test

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.

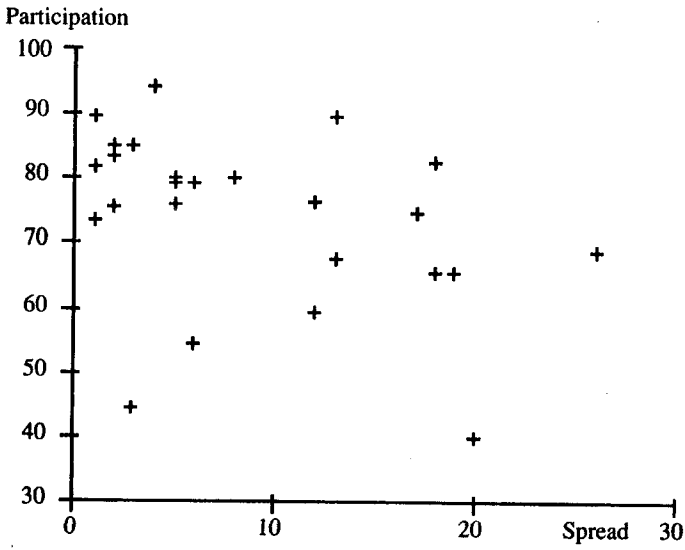

Figure 2.3 Participation rates versus vote spread in the 1844 presidential election

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 slave holdings	High	100 (55%)	81 (45%)	181
	Medium	85 (55%)	68 (45%)	153
	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.